COMPUTER VISION BASED AUTOMATED MONITORING AND ANALYSIS OF EXCAVATION PRODUCTIVITY ON CONSTRUCTION SITES

Chen Chen

A Thesis

in the Department

of

Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Building Engineering) at

Concordia University

Montreal, Quebec, Canada

February 2021

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Chen Chen

Entitled: **Computer Vision Based Automated Monitoring and Analysis of Excavation Productivity on Construction Sites**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Building Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examing Committee:

Chair
Dr. Youmin Zhang

External Examiner
Dr. Saiedeh Razavi

External to Program
Dr. Nizar Bouguila

Examiner
Dr. Fuzhan Nasiri

Examiner
Dr. Osama Moselhi

Thesis Supervisor (s)
Dr. Amin Hammad

Dr. Zhenhua Zhu

Approved by

Dr. Michelle Nokken, Chair of Department or Graduate Program Director

49/4/2021
Date of Defence

Dr. Mourad Debbabi Dean, Gina Cody School of Engineering and Computer Science

ABSTRACT

Computer Vision Based Automated Monitoring and Analysis of Excavation Productivity on Construction Sites

Chen Chen, Ph.D.

Concordia University, 2021

Construction equipment is the main component of construction production, and the equipment costs are usually one of the constructor's biggest expenditures. The productivity of construction equipment plays an important role in completing construction projects within schedule and under budget. Therefore, effectively monitoring equipment productivity is critical to improve construction productivity and control the cost. In order to monitor the productivity of the equipment, numerous research works are focusing on monitoring the operation process of the equipment. Through the continuous monitoring of equipment operations, detailed information about the productivity and performance indicators, such as activities, idling time, and work cycles can be estimated. In recent years, different information technologies, such as machine learning and Real-time Location Systems (RTLS), have been used for equipment productivity monitoring. Based on the type of the data collected, technologies can be categorized into computer vision (CV)-based methods and sensor-based methods. The CV-based methods collect equipment operation data from site surveillance cameras as videos or images. Sensor-based methods install sensors and/or tags (such as Radio Frequency Identification (RFID), Global Positioning System (GPS), Ultra-wideband (UWB), Inertial Measurement Unit (IMU), etc.) on the equipment and the construction site to collect the position and pose information of the equipment. Accordingly, the work states and activities of the equipment are identified by analyzing the data collected from cameras or sensors. The location and trajectory data collected from the sensors can be used to directly estimate the activity of the equipment. Finally, based on the activity information, the productivity of the equipment can be estimated in the form of equipment operation time or soil quantity. However, the previous studies did not consider the activity recognition and productivity analysis of multiple excavators. Moreover, the reasons that cause equipment low productivity, such as idling

reasons, have not been considered in previous CV methods. Furthermore, the sensor-based methods can be expensive and impractical to use for the large fleet of construction equipment. Compared with the sensor-based methods, CV-based methods are growing and becoming more efficient.

This research aims to monitor the productivity of the excavator using a CV-based method and has the following objectives: (1) Comparing the CV-based and sensor-based methods of earthmoving equipment productivity monitoring to identify the advantages and limitations of each approach, and proposing a roadmap for the future work directions of automated equipment productivity monitoring; (2) Developing an end-to-end CV-based method for automatically recognizing activities of multiple excavators; (3) Developing a framework to analyze the productivity of the excavator based on the activity recognition results; and (4) Improving the earthmoving productivity by identifying the idling reasons of excavators and trucks.

A systematic literature review is conducted to cover the recent studies in the area of activity recognition and to compare the CV-based methods with the sensor-based methods for earthmoving equipment productivity monitoring. Then, a roadmap is proposed to suggest future research directions for automatic equipment productivity monitoring. Accordingly, a framework that integrates three Convolutional Neural Networks (CNNs) is proposed for automatic detection, tracking, activity recognition and productivity analysis of excavators. The proposed framework has been tested with the videos recorded from real construction sites. The overall activity recognition has achieved 87.6% accuracy. The productivity calculation has achieved 83% accuracy, which indicates the feasibility of the proposed framework for automating the monitoring of excavator's productivity. Furthermore, knowing that idling is one of the main reasons for low productivity, a CV-based method was proposed to identify the idling reasons in earthmoving operations by analyzing the workgroup and interactive work states of the excavators and trucks. In this method, the activities of the excavators and trucks were identified using CNNs. Then, the onsite camera was calibrated to estimate the proximity and identify workgroups of excavators and trucks. By calculating the relationships between each excavator and the surrounding truck(s), the potential reason for idling can be identified. The proposed method has been validated with videos from several construction sites showing promising results.

The contributions of the research are: (1) Proposing a roadmap to show the future research directions of automatic equipment productivity monitoring; (2) Proposing a CV-based method to recognize the activities of multiple excavators with the state-of-the-art three dimensional (3D) CNN; (3) Proposing a framework that integrates the detection, tracking, and activity recognition techniques to monitor the operation process and analyze the productivity of the excavator; and (4) Developing a method to automatically identify the idling reasons of excavators and trucks based on their interactive work states, which contributes to a better understanding of earthmoving productivity under the dynamic and complex construction site conditions.

ACKNOWLEDGMENTS

I would like to thank all the people who have helped and inspired me during my four years of study at Concordia. First and foremost, I would like to express my sincere gratitude to my supervisors Dr. Amin Hammad and Dr. Zhenhua Zhu for their continuous support of my study and research. I appreciate all their contributions of time, patience, knowledge and ideas through the last several years. Their perpetual enthusiasms and conscientious attitude towards research have motivated me. I hope that I could be as enthusiastic, serious, and energetic as they are in my future work.

I would also like to thank the members of my committee, Dr. Saiedeh Razavi, Dr. Fuzhan Nasiri, Dr. Nizar Bouguila and Dr. Osama Moselhi for spending their precious time reading this thesis and for providing me with constructive feedback and suggestions. I gratefully acknowledge them for their valuable guidance and comments.

I feel very fortunate for studying with my following research group members: Ms. Negar Salimzadeh, Ms. Neshat Bolourian, Ms. Wenjing Chu, Dr. Xiaoning Ren, Mr. Amr Amr, Mr. Bo Xiao, and Mr. Bingfei Zhang. They make the lab a convivial place to work. I am also extremely thankful for my research colleagues Mr. Yusheng Huang and Mr. Mohammod Akbarzadeh, for their kind help in my research work. This research has been supported by Indus.ai and Hydro Quebec, and I would also like to acknowledge the support of my colleagues Dr. Walid Ahmad and Mr. Konan Donald from these two companies. I would also like to thank Dr. Jinwoo Kim and Dr. Seokho Chi for kindly providing the construction surveillance video for one of the case studies.

Lastly, I would like to express my deep appreciation to my parents for their love and support. They raised me with love of knowledge and supported me in all my pursuits. Without their encouragement, I would not have made it this far.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
3D	Three Dimensional
CV	Computer Vision
CSK	Circulant Structure of tracking-by-detection with Kernels
CNN	Convolutional Neural Network
CWT	Continuous Wavelet Transform
DCF	Dual Correlation Filter
DES	Discrete Event Simulation
DT	Dense Trajectory
DTW	Dynamic Time Warping
GDP	Gross Domestic Product
GMM	Gaussian Mixture Model
GPS	Global Positioning System
GPU	Graphical Processing Unit
HOF	Histogram of Optical Flow
HOG	Histogram of Oriented Gradients
KLT	Kanade and Lucas Tracker
IMU	Inertial Measurement Unit
ICT	Information and Communication Technology
LCY	Loose Cubic Yard
LSTM	Long Short-Term Memory
MBH	Motion Boundary Histogram

R-CNN	Region-based Convolutional Neural Network
R-FCN	Region-based Fully Convolutional Network
ResNet	Residual Neural Network
RFID	Radio Frequency Identification
RGB	Red, Green, and Blue
RNN	Recurrent Neural Network
RTLS	Real-time Location Systems
SIFT	Scale-Invariant Feature Transform
SORT	Simple Online and Real-Time
SSD	Single Shot Detector
SVM	Support Vector Machine
TLD	Tracking-Learning-Detection
TSN	Temporal Segment Network
UWB	Ultra-wideband
YOLO	You Only Look Once

CHAPTER 1: INTRODUCTION

1.1 General Background

Construction is one of the major industries in Canada, which has more than 1.2 million employees and contributes about 6% of Canada's Gross Domestic Product (GDP) (Statistic Canada 2011). However, it is considered as a low productivity industry compared with other industries. As shown in Figure 1-1, Quebec's overall industrial productivity (solid line) has been growing from 1982 to 2017 with a growth of more than 55.9% between 1989 and 2017 (Quebec Wood Export Bureau 2019). However, the productivity of the construction industry (dotted line) is decreasing. Moreover, few companies are investing in innovation, which could promote productivity in the construction industry. For example, in 2011, the companies in Quebec that have invested in innovation are 4% and 50% in construction and manufacturing, respectively (Statistic Canada 2011). Therefore, it is necessary and pressing to increase productivity in the construction industry.

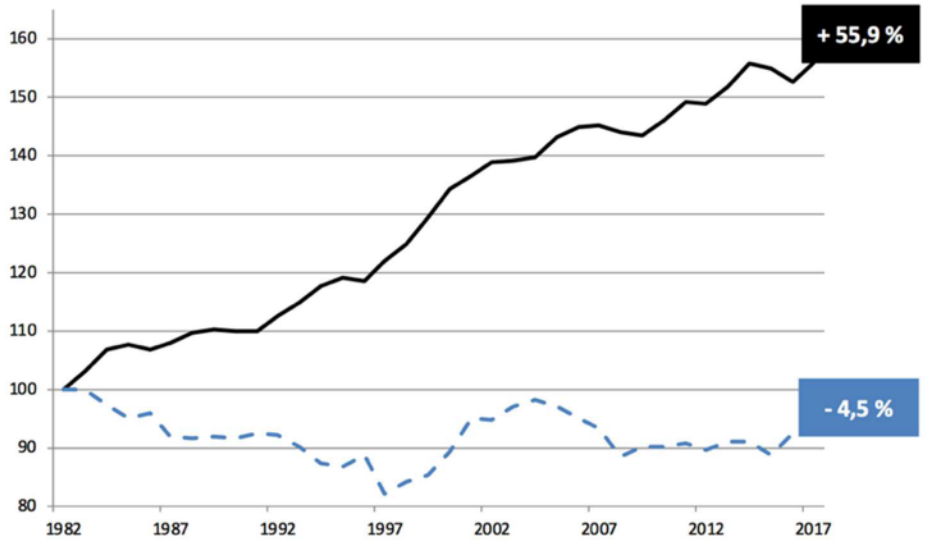


Figure 1-1 Construction productivity in Quebec from 1982-2017 (Quebec Wood Export Bureau 2019)

The construction equipment is one of the most important components in a construction project, which contributes much to the construction productivity (Kim et al. 2018a). Moreover, the equipment cost is usually one of the constructor's biggest expenditures (Peurifoy et al. 2009). Therefore, effectively managing the equipment operations is critical to improve construction

productivity and control the construction cost. However, existing practices for measuring and analyzing the equipment operations are considered as time-consuming, expensive, and error-prone work (Golparvar-Fard et al. 2013) as site superintendents are required to manually observe and record the entire operation process for each construction equipment on the construction site (Kim et al. 2018a). Not only the observation task, but also the data analysis and calculation are labor-intensive. Therefore, numerous researchers created automatic methods to accurately and efficiently monitor equipment operations and analyze productivity. This is especially true for the excavators, since earthmoving operations are highly dependent on equipment such as excavators and trucks.

1.2 Problem Statements and Research Gaps

Existing methods for automatically recognizing the activities of the construction equipment could be classified into sensor-based methods and computer vision (CV)-based methods. The sensor-based methods recognize equipment activities by analyzing its location, acceleration, velocity and orientation information from sensors, such as the Global Positioning System (GPS) (Oloufa et al. 2002; Pradhananga et al. 2013) and Inertial Measurement Unit (IMU) (Akhavian and Behzadan et al. 2015). However, they require attaching sensors to monitor the equipment, which may not be feasible for the rented or old equipment. Additionally, the sensor-based methods have difficulty in categorizing the activities in detail. For example, the GPS-based methods identify equipment activities based on its position changes; and therefore, they cannot precisely identify the activity when the equipment is located at the same position while performing different activities (Kim et al. 2017).

Compared with the sensor-based methods, the CV-based methods do not have the aforementioned drawbacks. Furthermore, they can visualize the equipment's state directly from images and videos, so that it becomes easy to identify false recognitions and analyze the reasons behind low productivity. For example, the increase in the excavator's swinging time may be due to avoiding passing workers or the unsuitable truck positions. Moreover, the productivity analysis could be made based on the automatic monitoring and analyzing the surrounding information provided by videos.

So far, the CV-based methods recognize the activities by analyzing the motion trajectories or the equipment location information in video frames (Yang et al. 2016b). They cannot classify the

activities into detailed categories (e.g. digging, hauling, dumping, etc.) in long video sequences. In addition, most of the existing methods are not able to recognize and analyze the activities when multiple pieces of equipment are captured simultaneously. Moreover, existing methods just focus on activity recognition. Their practical values are not well discussed, such as the productivity analysis and low productivity reasons identification. As a result, it is not clear how these methods can be used for productivity monitoring in construction projects.

1.3 Research Objectives

The lack of a CV-based method with the ability to monitor the operation process of excavator to improve the productivity of earthmoving work, brings the attention of this research. This research aims to achieve the following objectives: (1) Comparing the CV-based and sensor-based methods of earthmoving equipment productivity monitoring to identify the advantages and limitations of each approach, and proposing a roadmap for the future work directions of automated equipment productivity monitoring; (2) Developing an end-to-end CV-based method for automatically recognizing activities of multiple excavators; (3) Developing a framework to analyze the productivity of the excavator based on the activity recognition results; and (4) Improving the earthmoving productivity by identifying the idling reasons of excavators and trucks.

1.4 Thesis Organization

The structure of this thesis is presented as follows:

Chapter 2 Literature Review: This chapter reviews the current state of statistics, concepts, technologies and methods related with automatic construction equipment monitoring and productivity analysis. Furthermore, a roadmap is proposed to introduce the advanced research methods of construction equipment productivity monitoring and indicate future directions for research and the industry to take up based on current technologies.

Chapter 3 Overview of The Proposed Research Framework: This chapter describes the overall proposed framework of this research work. There are three main methods adopted in this work, and each of them was explained briefly in this chapter.

Chapter 4 Vision-based Activity Recognition and Productivity analysis: This chapter goes through the details of the proposed method for CV-based excavators' activity recognition and productivity analysis and demonstrates the feasibility of the method with real site surveillance videos.

Chapter 5 Automatic Identification of Idling Reasons in Earthwotk Operations Based on Excavator-truck Relationships: The method for idling reasons identification and interative opertions analysis between excavators and trucks are introduced and validated with three case studies.

Chapter 6 Summary, Contributions, and Future Work: In this chapter, the summary of the work done in this research is provided followed by the contributions. The further possibilities for extending and improving the current work are explained in the future work section.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In this chapter, the status of current practices in automatic construction equipment operation monitoring and productivity analysis are reviewed. The application of CV-based and sensor-based equipment detection, tracking, activity recognition, and equipment productivity calculation methods are also reviewed. All these technologies are the ones that this research plans to build on and augment. At last, the limitations and the research gaps in the existing works are highlighted, and a roadmap for future research is proposed.

2.2 Productivity Monitoring in Construction

The construction industry, with the annual expenditures reaching over $1,231 billion in the United States (Statistic 2018) and $625 billion in Canada (Statistics Canada 2018), plays a significant role in the economy of North America. Therefore, productivity improvement in the construction industry has a significant impact on improving the GDP in North America (Arditi and Mochtar 2000). Since construction equipment is the main component of the construction production (Kim et al. 2018a), an accurate analysis of the productivity of the equipment is critical for construction productivity control, project control, and construction planning (Ok and Sinha 2006). Recognizing the activity of the equipment is important to manage its productivity. Knowing the activity of the equipment could help measure its actual work hours, and compare the input and output of the equipment over time (Goodrum et al. 2002). These methods of construction equipment activity recognition and operation monitoring are explained in Section 2.4 and Section 2.5.

2.3 Earthmoving Equipment Monitoring

Earthwork operations are one of the most important works in construction projects, which are highly mechanized since they involve moving large quantities of soil within a limited period of time for subsequent works (Tam et al. 2002). In earthwork operations, a group of construction equipment usually works together to accomplish tasks, such as excavating, hauling, and compacting soil (Peurifoy et al. 2009; Montaser and Moselhi 2014). Excavators and trucks are the

most widely used resources for earthwork. During earthwork operations, an excavator and a fleet of trucks usually work as a team to dig and transport the soil to the dumping site (Hola and Schabowicz 2010). In this process, the excavator works under the cycle of digging and swinging for loading a truck, and swinging back for digging (Hola and Schabowicz 2010; Kim and Chi 2019; Chen et al. 2020a). Meanwhile, trucks are loaded before hauling the soil under three main states, which are moving, stopping to be loaded, and idling. Therefore, the productivity of earthwork operations is generally calculated based on the number of work cycles of the excavator and trucks (Kim and Chi 2020; Montaser et al. 2012; Zhang 2008). Idling is another important work state of the excavator, which has a great influence on productivity. Idling may be caused by different reasons. For example, an excavator may be idling because the next truck to be loaded did not arrive (Kim and Chi 2019). To avoid this idling, a certain number of trucks should be arranged to work with the excavator and to keep the excavator working at capacity (Peurifoy et al. 2009). In other conditions, the idling may be caused by machine malfunction, congested site or safety problems. Therefore, to decrease the idling time and reduce cost, it is essential to identify the reasons that cause idling and provide solutions that fit the situation.

In recent years, different information technologies, such as machine learning (Azar et al. 2013; Kim and Chi 2017) and Real-time Location Systems (RTLS) (Montaser et al. 2012; Louis and Dunston 2018) have been used for equipment productivity monitoring. Based on the types of the data collected, they can be categorized into CV-based methods and sensor-based methods. The conceptual workflow of these methods is shown in Figure 2-1. The CV-based methods collect equipment operation data from site surveillance cameras as videos or images. Sensor-based methods install sensors and/or tags (e.g., Radio Frequency Identification (RFID), Global GPS, Ultra-wideband (UWB), IMU, etc.) on the equipment and the construction site to collect the position and pose information of the equipment. Accordingly, the work states and activities of the equipment are identified by analyzing the data collected from cameras or sensors. For example, the location and trajectory data collected from the sensors can be used to estimate the activity of the equipment directly. The visual data are usually processed with CV-based methods (e.g. machine learning and deep learning) to identify the equipment activity. Finally, based on the activity information, the productivity of the equipment can be estimated in the forms of equipment operation time or soil quantity.

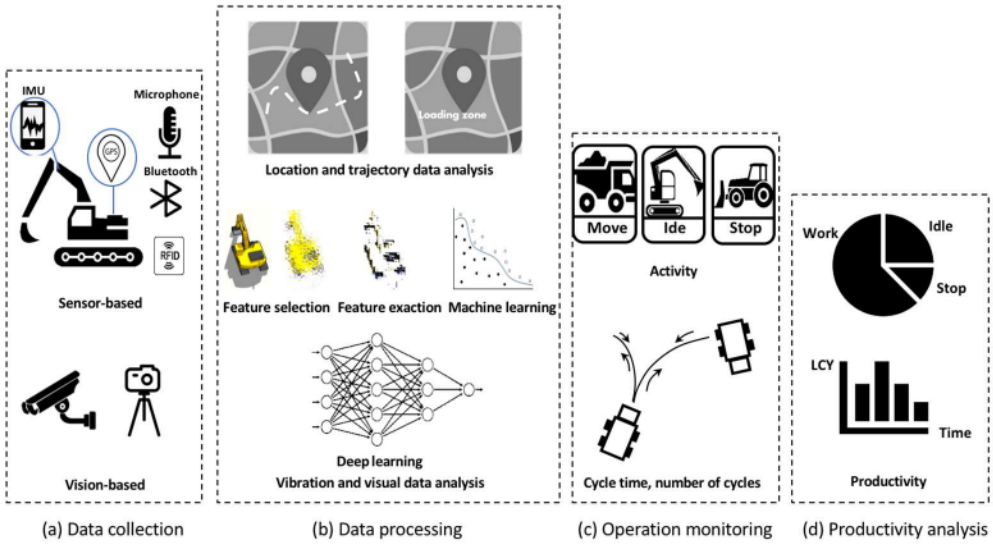


Figure 2-1 Conceptual workflow of equipment productivity monitoring

2.4. CV-based Equipment Productivity Monitoring

The application of CV technology for equipment monitoring has been popular in recent years for several reasons, such as the development of promising object detection and tracking algorithms in computer science, and the widespread of low-cost cameras installed in construction sites to collect videos and images. A general CV-based equipment monitoring framework consists of several steps. First, the equipment detection is used to recognize the particular type of equipment in the image or video frames. Then, different pieces of equipment are continuously tracked in all video frames. The detection and tracking methods provide the spatial position and movement information of the equipment. Accordingly, activity recognition and pose estimation are conducted to evaluate the work states of the equipment, which are necessary for productivity analysis. Figure 2-2 shows the distribution of research topics of CV-based equipment monitoring articles. 20 (43%) papers focused on equipment detection and tracking; 11 (23%) papers focused on equipment activities recognition; 11 (23%) papers focused on equipment productivity monitoring; and 5 (11%) papers focused on equipment pose estimation. The number of productivity monitoring papers started increasing in 2019 (as shown in Figure 2-3), since the technologies of productivity monitoring are based on detection, tracking and activity recognition.

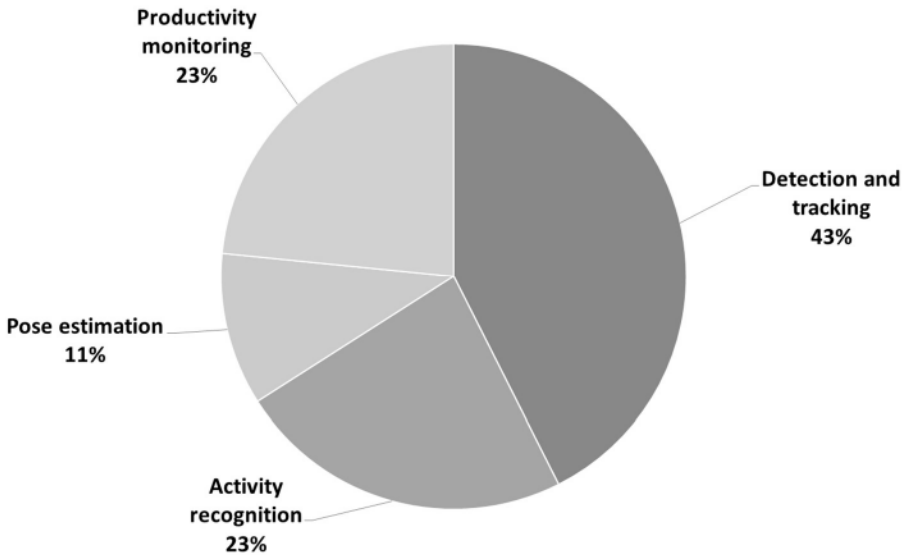


Figure 2-2 Distribution of research topics of CV-based papers (for the period 2007-2020)

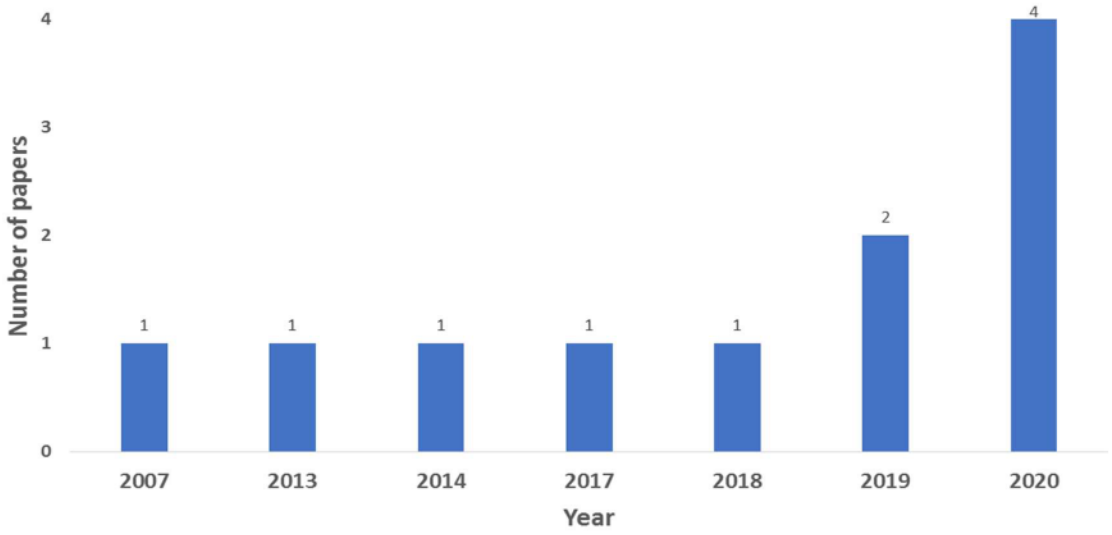


Figure 2-3 Distribution of CV-based productivity monitoring papers

2.4.1 CV-based equipment detection methods

The application of CV in equipment monitoring starts with equipment detection. The summary of existing CV-based detection methods is shown in Table 2-1. Most of the earlier works used feature-based methods to detect equipment, such as excavators and trucks, in images and videos. Those works first extract features to represent the appearance characteristics of the equipment. Then, classifiers are trained to identify the equipment by classifying the vectors created from the features.

The Histogram of Oriented Gradients (HOG) is the most widely used feature descriptor. Azar and McCabe (Azar et al. 2013; Azar and McCabe 2012b) used HOG feature and Support Vector Machine (SVM) classifier to detect dump trucks in videos. The method could detect eight different orientations of the dump truck with an accuracy of 86.8%. Memarzadeh et al. (2013) combined HOG with color features and Hue-Saturation values, and used SVM classifier to recognize workers, excavators and trucks in video frames with an accuracy of 98.83%, 82.10%, and 84.88% respectively. Tajeen and Zhu (2014) used HOG features and latent SVM classifier, which recognized five classes of equipment (backhoe, dozer, excavator, loader, and roller) from images, with the accuracy of 70% (backhoe), 82% (dozer), 64% (excavator), 66% (loader) and 78% (roller). Rezazadeh and McCabe (2012a) indicated that compared with other objects, the accuracy of the excavator detector is relatively low because excavators have countless deformable forms. In order to train a HOG-based detector with limited images, they created the part-based model, which recognized excavators by detecting the boom. The method achieved lower misclassifications than general HOG-based methods. Kim et al. (2017) also used HOG and SVM classifier to recognize the trucks and workers in the construction images with the accuracy of 95.7%.

Some methods recognized moving equipment by subtracting it from the background. Zou and Kim (2007) used color space values to isolate the excavator in images with snow and soil backgrounds. Chi and Caldas (2011) segmented the regions of moving objects from the image with Gaussian Mixture Model (GMM) algorithm (Reddy et al. 2009), then, they used two classifiers, Bayes classifier (Fukunaga 2013) and 4-layers neural network, to classify the segmented parts into workers, backhoes and loaders. The error of Bayes and neural network are 3.35% and 3.9%, respectively. Kim et al. (2016) and Bügler et al. (2014; 2017) also used GMM and Bayes networks to identify excavators and trucks in video frames. However, the background subtraction method is limited to detecting moving and idling objects, which may not be sufficient in the application for identifying other activities of the equipment.

Despite these efforts, previous methods were not able to differentiate specific equipment within the fleet. To fulfill this gap, Azar (2016) proposed a marker-based recognition method to recognize individual excavators and trucks in videos. The method first created and attached markers on the equipment; then, the equipment can be detected by recognizing the markers.

The development of deep learning methods using Convolutional Neural Network (CNN) also affected equipment detection. A major difference between feature-based methods and CNN is their way of learning features of the objects. CNN can automatically learn representative features from images in the dataset; whereas feature-based methods just extract human-designed features, which is not easy for the complex construction environment (Kim et al. 2018c). Numerous CNNs have been used for various equipment detection tasks with better performances than feature-based methods (Roberts and Golparvar-Fard 2019; Chen et al. 2020a). Fang et al. (Fang et al. 2018b) used Faster Region-based Convolutional Neural Network (Faster R-CNN) to detect excavators with the accuracy of 95%. Kim et al. (2018c) trained Resnet-50 with 2920 images to recognize four classes of equipment (loader, excavator, dump truck and concrete mixer truck) with the average precision of 96.33%. Kim et al. (2020) trained Faster R-CNN (Fang et al. 2018b), Single Shot Detector (SSD) (Liu et al. 2016) and You Only Look Once (YOLO) (Redmon et al. 2016) detectors with the same dataset to detect excavators, trucks, forklifts and loaders and got the accuracy of 93%, 92.1% and 89.7%, respectively for the three methods.

Without training a classifier to recognize any specific type of equipment, Kim and Chi (2017) and Kim et al. (2018d) applied a tracking-based method called Tracking-Learning-Detection (TLD) (Kalal et al. 2012) to identify the target equipment in video frames. The method first selects the target equipment that needs to be identified with a bounding box. Then, a tracker and a detector will learn online to search the target in the next video frame based on the trajectory, as well as the spatial, gray-value variance, and pixel variance information.

Table 2-1 Summary of equipment detection methods used in the reviewed papers

Detection methods	Type	References
Feature-based methods	HOG + SVM	Azar and McCabe (2012a) Azar et al. (2013) Azar and McCabe (2012b) Tajeen and Zhu (2014) Kim et al. (2017)
	HOG, color, Hue-Saturation + SVM	Memarzadeh et al. (2013)
	Color	Zou and Kim (2007)
GMM-based methods	GMM, Bayes network, Two-layers CNN	Chi and Caldas (2011)
	GMM	Bügler et al. (2014; 2017)
	GMM+Bayes network	Kim et al. (2016)
Marker-based methods	Barcode marker+HOG+SVM	Azar (2016)
TLD methods	Trajectory, spatial and gray-value variance, pixel variance	Kim and Chi (2017) Kim et al. (2018d)
CNNs-based methods	Resnet-50	Kim et al. (2018c)
	Faster R-CNN	Chen et al. (2020a) Fang et al. (2018b)
	CNN+LSTM	Kim and Chi (2019)
	ResNeXt-101	Roberts and Golparvar-Fard (2019)
	Faster R-CNN + SSD + YOLO	Kim et al. (2020)

2.4.2 CV-based equipment tracking methods

The tracking methods aim to associate and get the trajectory of each piece of equipment across all video frames. Various methods were used to track equipment such as mean-shift tracking (Gong and Caldas 2011), Kanade and Lucas Tracker (KLT) (Azar et al. 2013), counter-based and point-based algorithms (Park et al. 2011), kernel covariance (Teizer 2015), and Kalman filter (Kim et al. 2016; Kim et al. 2017). Park et al. (2011) compared and evaluated three widely used tacking methods, contour-based method, kernel-based method and point-based method to track workers, equipment and materials in construction sites. They indicated that the kernel-based method is the most appropriate method for tracking construction-related resources considering the occlusion, illumination, and scale variation conditions. Xiao and Zhu (2018) evaluated the accuracy and robustness of 15 visual tracking methods on 20 different construction scenarios. They found the self-occlusion of the excavators made them difficult to be tracked, and only Dual Correlation Filter (DCF) (Henriques et al. 2015) and Circulant Structure of Tracking-by-detection with Kernels (CSK) (Henriques et al. 2012) methods could complete the tracking under heavy occlusions.

Zhu et al. (2016) proposed a particle-based tracker, which could track workers and equipment continuously (i.e., roller, truck, and dozer) even when they have a long period of collusions. Their method represented target objects by a set of particles and required no offline training. Yuan et al. (2017) improved the point tracking method (Lucas and Kanade 1981) by providing a failure checking mechanism to track excavators. They first generated the optical flow images of the excavator. Then, they tracked the key points based on the premise that the target between two consecutive frames continues with constant image brightness. Kim and Chi (2017) integrated hybrid methods such as median-flow algorithm (Kalal et al. 2010) and a pyramidal Lucas-Kanade algorithm (Bouguet 2000) to estimate and represent object motions among consecutive frames and track excavators in long videos. Hybrid tracking methods were also applied to avoid long time occlusion and interclass variations. Chen et al. (2020a) used a deep Simple Online and Real-Time (SORT) tracker (Wojke et al. 2017), which combined CNN and Kalman filter to track excavators and trucks.

2.4.3 CV-based equipment activity recognition methods

Based on equipment detection and tracking methods, numerous research works have been focused on developing more practical methods for monitoring the operation and productivity of the equipment. One important topic for equipment monitoring is activity recognition, since the activity information has a direct relationship with the productivity analysis. So far, there are mainly three approaches of CV-based activity recognition, which are feature-based methods, rule-based analysis, and spatial-temporal CNN methods (Figure 2-4). In the following subsections, these three approaches are discussed in detail.

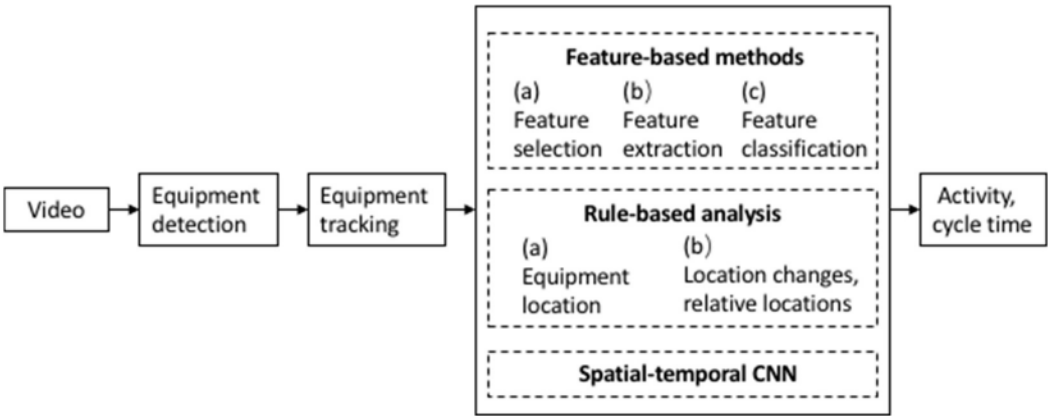


Figure 2-4 CV-based equipment activity recognition workflow

(1) Activity recognition with features

The feature-based methods are similar to feature-based equipment detection methods. After collecting the spatial features in each video frame, the features in consecutive video frames are encoded to a vector or matrix for the activity classification. These feature-based methods are used in early works for human activity recognition in CV as shown in Section 2.6.1. Gong et al. (2011) used a three dimensional (3D)-Harris corner detector to detect the interest points, and used HOG and Histogram of Optical Flow (HOF) descriptors to describe interest points in every two consecutive frames. Then, they trained the Bayesian neural network (Li et al. 2003) to classify three kinds of backhoe activities (i.e., relocating, excavating and swinging). The results showed that the HOG feature has a better performance than HOF with an overall accuracy of 86.33%. Similarly, Golparvar-Fard et al. (2013) used 3D HOG as spatiotemporal features (Figure 2-5) and SVM as a classifier to recognize four activities of excavators (i.e., digging, hauling, dumping and swinging) and three activities of trucks (i.e., filling, dumping and moving). Roberts and Golparvar-Fard (2019) used neural networks to get the bounding boxes of the excavators and trucks in video frames. Then, they extracted the HOG, HOF and Motion Boundary Histograms (MBH) features of excavators and trucks in the bounding boxes every 20 frames. At last, an SVM classifier was trained to classify the activities of the equipment into eight categories as shown in Table 2-2. One major limitation of the feature-based methods is that they can only recognize single activity of each piece of equipment in short video clip, since the motion trajectories of the features drift a lot with the progress of the time (Chen et al. 2020a). The gap in recognizing consecutive activities in long video sequences has limited the application value of these methods in safety and productivity monitoring.

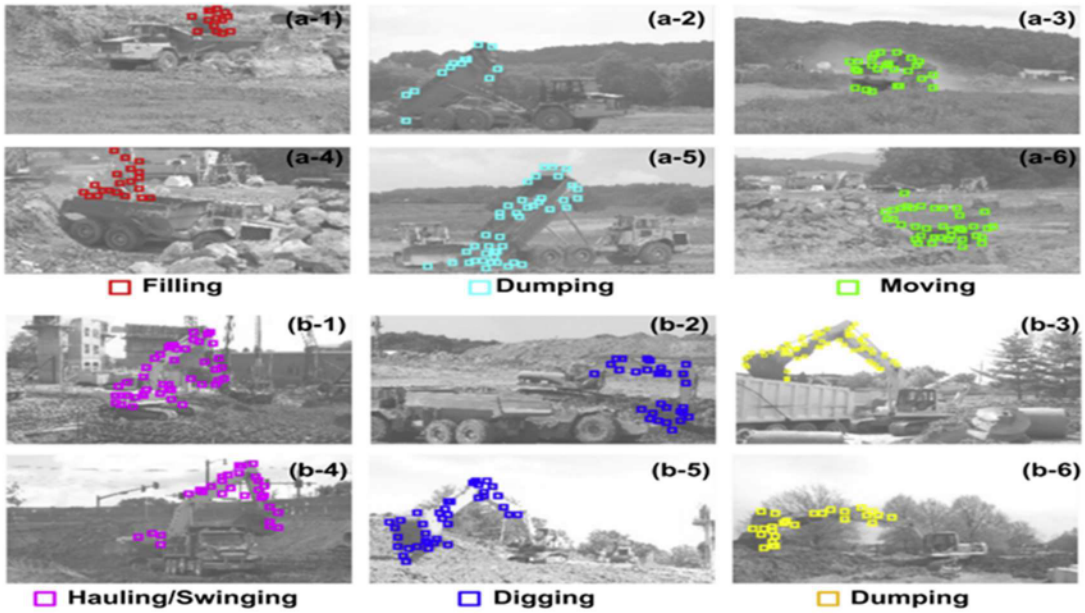
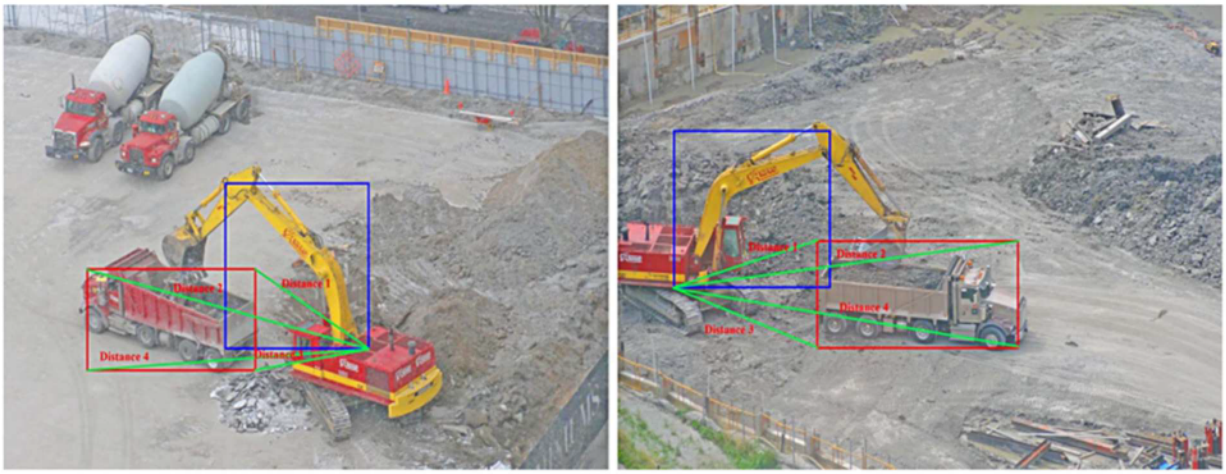


Figure 2-5 Example features from both truck and excavator (Golparvar-Fardetal et al. 2013)

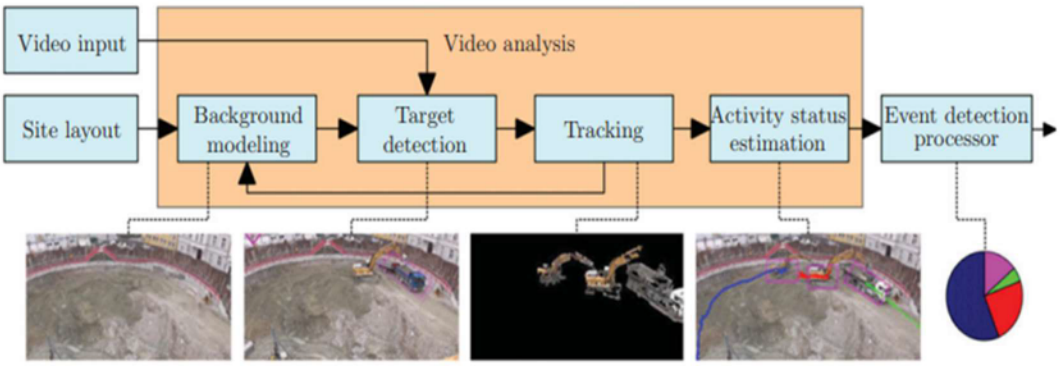
(2) Activity recognition with rule-based methods

The rule-based methods are widely used for productivity analysis. These methods recognized activities based on the detection and tracking results. They first extract pixel coordinates of equipment in video frames from the detection and tracking results. Then, activities are identified by analyzing the changing coordinates in video frames or the relative distances between different pieces of equipment. Zou and Kim (2007) used color features detector and centroid tracker to get the coordinates of the excavators in video frames. Accordingly, the idling and stopping activities of the excavator were differentiated by comparing the changing of centroid coordinates in consecutive video frames. In the video of the resolution of 640 × 480 pixels, if the distance of centroid coordinates in consecutive frames changed less than 1.6 pixels, the excavator was identified as idling. Azar et al. (2013) used HOG-based detectors to get the coordinates of excavators and trucks. Then, an SVM classifier was trained to identify loading activities using the vectors computed from the distance between the base point of the excavator and four corners of the dump truck (Figure 2-6(a)). The cycle time was managed as the duration between two loading activities. Bügler et al. (2014; 2017) divided the construction site into several interest regions; then, they used feature-based method proposed by Golparvar-Fard et al. (2013) to identify the static and moving activities of excavators and trucks (Figure 2-6(b)). If both the excavator and truck were

detected in the earthmoving region, and their distance is less than the threshold, the activity is identified as filling. The method has been tested with 2 days' videos. The average accuracy of filling activity was 81.8%. With the given information of the excavator's bucket volume and hourly bucket numbers, they calculated the production as the volume of soil that is excavated, which is 17.44% different from the ground truth. Azar (2017) recognized five kinds of equipment (bulldozer, excavator, truck, grader, and roller) with HOG-based classifier, and used the Bayesian network to estimate the probability of activities based on the other contents detected in the same frame. Considering that the excavator and truck are within a certain distance at the working state of loading, Kim et al. (2018b) identified loading activity by comparing the distance changes between the two pieces of equipment with the thresholds. Kim et al. (2018d) also recognized activities with the changing of centroid coordinates in consecutive video frames. In order to obtain more accurate results, they adopted an interaction analysis between excavators and trucks. For instance, if the excavator is working while the nearest truck is stopping, the activity of that truck should be identified as working. This interactive analysis helps to distinguish between working trucks interacting with the excavator, and idling trucks waiting to be loaded. The method achieved an average accuracy of 91.27%, which is 15.59% higher than non-interactive analysis.



(a) Distances between the corners of trucks and the base point in both left and right configurations (Azar et al. 2013)



(b) Workflow of the automatic surveillance system (Bügler et al. 2017)

Figure 2-6 Examples of the detection-tracking-based methods

The methods based on the detection and tracking results can recognize consecutive activities in long video sequences and calculate the duration of each activity in order to estimate the productivity of the equipment. However, they still have the following limitations. Instead of using the 3D real distance, these methods applied 2D pixel distance for proximity estimation. Considering that most cameras are installed with an inclined angle in the construction site, the 2D pixel distance cannot reflect the real spatial relationships between different pieces of equipment in the real construction scenario. In addition, these methods need to pre-define thresholds and adjust them with the change of camera positions, which is inconvenient to use. The limitations of the rule-based methods can be overcome in the spatial-temporal CNN methods.

(3) Activity recognition with spatial-temporal CNNs

In most recent works, the activity of the equipment was directly recognized with spatial-temporal neural networks. For example, Kim and Chi (2019) combined a CNN and a LSTM network to recognize the excavator's activities of digging, hauling, dumping and swinging. In their work, they assumed that during the operation, the excavators usually work in the sequence of digging, hauling, dumping and swinging. Therefore, they created a hybrid neural network, which has 10-layers CNN and 2-layers LSTM to learn the visual and sequential features, respectively. Because of the gradient descent problem of LSTM, this method has limited recognition accuracy in long videos. Chen et al. (2020a) used 101-layers 3D Residual Neural Network (ResNet) to learn the spatial-temporal features of the activity in every consecutive 16 video frames, and recognized the digging, swinging

and loading activities of the excavator. The method achieved 87.6% accuracy on a 1 h video test. One limitation of this method is that a large video dataset is needed to train the 3D ResNet.

2.4.4 CV-based equipment pose estimation methods

Pose estimation is also used to identify the work states of excavators. Lundeen et al. (2016) and Feng et al. (2015) attached several markers on excavators to estimate the poses. Yuan et al. (2017) used 2D tracking and stereo cameras to get the 3D coordinates of the key nodes of the excavator. Soltani et al. (2017) estimated the skeleton of the excavator by detecting its parts. Soltani et al. (2018) applied stereo cameras to estimate the 3D pose of the excavator from 2D detections. Since the 3D pose estimation retrieved the 3D position of the equipment in global coordinates, it was applied for safety management based on proximity analysis.

2.4.5 Equipment idling state recognition with thermal images

Some works identified equipment idling states to estimate the resulting air pollution using thermal images, which are collected with infrared cameras. Infrared cameras absorb the heat energy from the surface of the equipment and give an illustration in the forms of thermal images. Therefore, it is used to recognize idling states of vehicles and equipment based on the temperature of engines (Bastan et al. 2018). Lustbader et al. (2012) applied thermal images to identify the idling state of the truck and compared the diesel fuel use in idling and working states. Jain et al. (2018) used thermal images to differentiate between the truck engine on and engine off idling conditions, and monitor gas emissions. Thermal images are useful in high occlusion and poor lighting conditions to identify the equipment (Pazhoohesh and Zhang 2015). However, most of existing works applied the thermal images to monitor the emissions and not for equipment productivity monitoring.

2.4.6 CV-based productivity analysis methods

The summary of CV-based methods for productivity analysis is shown in Table 2-2. Most of the existing works calculated productivity by recognizing the activities of the equipment. Some works calculated productivity by estimating the proportion of working or idling times in the whole operation time. For example, Zou and Kim (2007) calculated that during the 10,800 s investigation time, the excavator was idling 3,800 s with a working rate of 64.8%. As another example, during 2 h and 27 min captured excavation video, Azar et al. (2013) calculated that there is 1h 35 min of excavator's loading activities. Some works identified the number of work cycles of the excavators

or trucks. With the given information of the bucket volume, they could calculate the volume of soil that has been excavated. Bügler et al. (2014; 2017) recognized loading time as described in Section 2.4.3; then, using the excavator's bucket volume and hourly bucket numbers, they further calculated the productivity in the volume of soil that has been excavated, which is 17.44% less than the ground truth. Kim et al. (2018b) recognized activities of excavators and trucks, then, based on the activity's information, a method for simulating equipment processes was created to estimate the cycle numbers of trucks and calculate the earthmoving productivity. Chen et al. (2020a) identified the number of work cycles of the excavator from the sequential relationship of the recognized activities. With the known bucket volume, the productivity of the 60.2 min excavation was calculated as 131.16 LCY/h, which is 17% lower than the ground truth.

Instead of productivity analysis with activity recognition, some research works calculated productivity with other information. Kim et al. (2019) trained a Region-based Fully Convolutional Network (R-FCN) (Dai et al. 2016) to recognize the license plate numbers of dump trucks when they pass through the gate of the construction site. Accordingly, by measuring the time interval of each dump truck coming and leaving the site, the number of trucks' cycles and the total volume of earthmoving was calculated. Kim and Chi (2020) proposed a multi-camera-based productivity monitoring method. They installed two cameras at entry and loading zones, separately. Then, by matching the dump trucks in two cameras with the queueing discipline analysis, the cycle time and number of cycles of the dump truck could be estimated.

Table 2-2 Comparison of CV-based papers of activity recognition and productivity analysis

Method		References	Activities		Dataset	Result	Test video	Calculate activity duration	Calculate productivity
			Excavator	Truck					
Activity recognition	Feature-based methods	Gong et al. (2011)	Relocating, excavating, swinging		150 videos, 7 fps, 10s per video	Accuracy: 86.33%	6-10 fps 720×480 pixel		
		Golparvar-Fard (2013)	Digging, hauling, dumping, swinging	Filling, dumping, moving	895 videos, 250×250 pixels, 4~16 s	Accuracy: 86.33% (excavator) 98.33% (truck)			
		Roberts and Golparvar-Fard (2019)	Idling, swinging, loading, moving, dumping	Moving, filling, hauling	480 × 272 pixels, 25 fps, 514 videos	Accuracy: 86.8% (excavator) 88.5% (truck)			
	Rule-based methods	Zou and Kim (2007)	Idling, stopping			Accuracy: 96%	10800 s	√	
		Azar et al. (2013)	Loading	Loading	7 videos with 51 min, 640×480	Accuracy: 95%	2 h, 27 min	√	
		Bügler et al. (2014; 2017)	Filling			Accuracy: 82.66%	20 h	√	√
		Azar (2017)	Bulldozer: excavation, spreading, excavator: excavation, trenching, loading, truck: backfilling, loading, hauling, compaction, grader: spreading, ditch cutting, roller: compaction		49 videos	Precision: 65%-78%			
		Kim et al. (2018b)	Dumping, idling	Loading, hauling		Accuracy: 98.4%	8 h 720×480 pixel	√	√
		Kim et al. (2018d)	Idling, traveling, working	Idling, working	150 min videos with 720 × 1280 or 1080×1920 pixels)	Accuracy: 89.36% (excavator), 93.80% (truck)	100 min	√	
	Spatial-temporal CNNs methods	Kim and Chi (2019)	Digging, hauling, dumping, swinging		121 min video with 720 × 1280 pixels and 10 fps	Precision: 90.09%	32,365 images	√	
		Chen et al. (2020a)	Digging, swinging, loading		381 videos clips with 1280 × 720 pixels and 25fps	Accuracy: 92%	1h	√	√
License plate recognition		Kim et al. (2019)	Truck work cycles			Accuracy: 96.76%	2,867,484 s	√	√
Cameras matching		Kim and Chi (2020)	Truck work cycles			Accuracy: 97.6%	88 min 720 × 1280 pixels	√	√

2.5. Sensor-based Equipment Productivity Monitoring

Sensor-based methods attach different sensors on the equipment to localize and track its equipment movement. These methods use telematics and logging devices to transmit and record the data. By processing the position and pose data retrieved from the sensors with classification or simulation methods, the productivity of the equipment can be estimated. In this section, sensor-based papers are reviewed in four categories based on the type of sensors used in each paper: (1) RTLS sensors, which can detect the locations and trajectories of the equipment; (2) vibration sensors, which can get the movement and pose information of the equipment; (3) audio sensors, which collect sounds during equipment operations; and (4) hybrid sensors, which use more than one kind of sensors for equipment productivity monitoring. The distribution of each kind of sensors used in the reviewed papers is shown in Figure 2-7. RTLS sensors are most frequently used in 14 papers (41%). Vibration sensors are used in 11 papers (32%), and they can extract more detailed work state of the equipment. Audio sensors appeared in three papers (9%) related to equipment activity recognition since 2017. Six papers (18%) applied multiple sensors for equipment monitoring. There are also ambient weather sensors, which are used to monitor the influence of weather conditions on the productivity, such as temperature, humidity, and wind (Ibrahim and Moselhi 2014; Song et al. 2017; Salem et al. 2018).

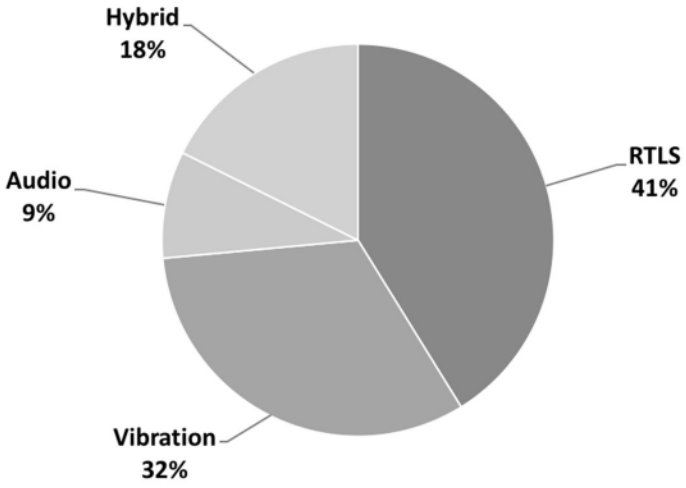
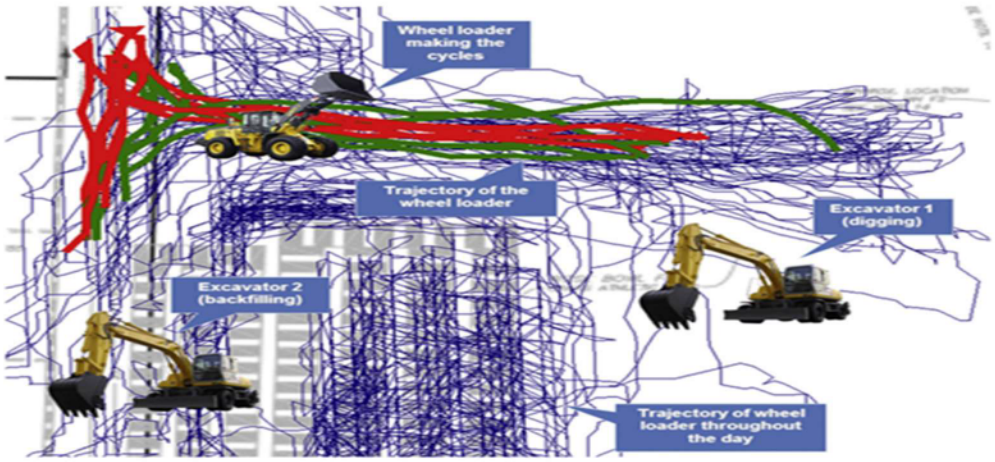


Figure 2-7 Distribution of sensors used for equipment monitoring in reviewed papers

2.5.1 RTLS sensors

GPS, RFID and UWB are widely used RTLS sensors, which can provide location and trajectory data of the equipment. GPS is a satellite-based navigation sensor system that can obtain the longitude, latitude, and altitude data of the equipment (Alshibani and Moselhi 2016). In the implementation of GPS, some methods estimated activities by observing the trajectories of the equipment on the map. Montaser et al. (2012), Montaser and Moselhi (2014), and Alshibani and Moselhi (2016) used GPS sensors to collect trajectories of the trucks, and estimated the durations of the activities (e.g. load, travel, return, etc.) by analyzing trucks' trajectories on the map. Ahn et al. (2020) also used GPS to collect path trajectories of trucks on the map to estimate the transportation cost. Some GPS-based methods separated construction site into different kinds of work zones, such as excavation and loading zones. By observing the location of the equipment in specific work zones, the activities or cycle time can be estimated, as shown in Figure 2-8. For example, Han et al. (Han et al. 2008) attached GPS sensors on the trucks to estimate the loading and traveling time based on the locations of the trucks on the construction sites. Song and Eldin (2012) separated the site into work zone and non-work zone to estimate the work time of the trucks. Pradhananga and Teizer (2013) installed GPS on excavators and skid steer loaders, and separated the site into gravel and excavation zones. By observing the locations of the excavators and loaders, the durations of excavation and loading were estimated. Song et al. (2017) and Louis and Dunston (2018) divided the construction site into excavation and dump zones to estimate the trucks activities such as enter, exist, load, etc.

(a) GPS sensor mounted on the equipment



(b) Trajectories of the equipment

Figure 2-8 GPS sensor-based method (Pradhananga and Teizer 2013; Oloufa et al. 2002)

UWB is a radio frequency positioning system that uses a triangulation method to estimate the location of the equipment based on the propagation durations of the signals from the tag to the receivers (Li et al. 2017). Teizer et al. (2008) attached UWB tags on the hook of a crane to recognize the lifting activity of the crane. Cheng et al. (2011) tested the feasibility of commercially available UWB systems in tracking equipment, materials, and workers in a large construction site. Vahdatikhaki and Hammad (2014) divided the construction site into dumping, hauling, loading, and excavation zones to estimate the activities of a truck and an excavator based on the locations calculated form UWB sensors. Instead of tracking the positions of the equipment, Vahdatikhaki et al. (2015) attached 6 tags on different components of the excavator to identify its 3D pose with UWB sensors.

RFID uses electromagnetic transmission to identify and track tags attached on the objects (Ni et al. 2003). The RFID system consists of readers and tags. The tags are attached to the equipment to transfer the digital data to the readers with radio waves (Lu et al. 2011). The RFID can be used for applications of distance estimation, scene analysis, and proximity (Bouet and Santos 2008). Using the triangulation algorithm with the propagation time of the signal, the distance of the tag to the readers can be estimated. In the scene analysis, first, a finger-print map is collected. Then, the location of the tag is estimated by matching the fingerprint with the online measurement. The proximity estimation relies on dense deployment of antennas. In the equipment monitoring, the readers can have a pre-determined power level to define a certain range in which it can detect

RFID tags. Therefore, by placing tags in different work zones, the equipment locations can be estimated. Montaster and Moselhi (2012) installed RFID readers at the entrance gates of loading and dumping regions; by recording the entrance and existing times, the loading and dumping durations were calculated (Figure 2-9).

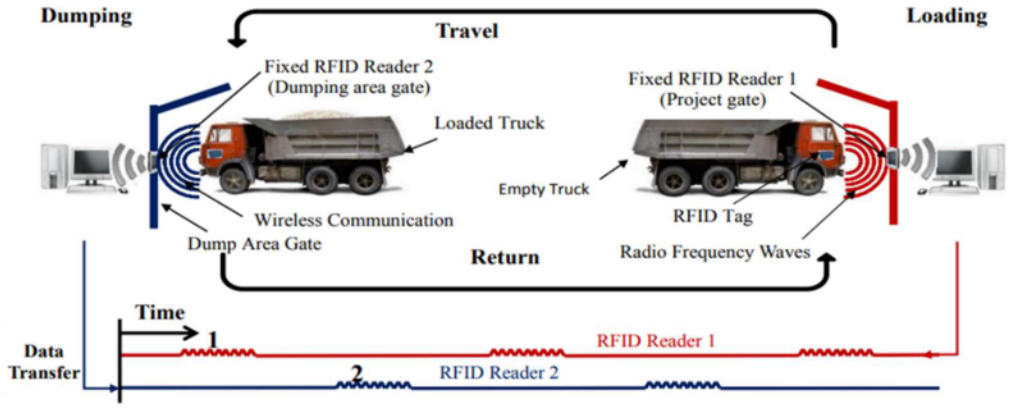


Figure 2-9 Overview of the RFID method (Montasser and Moselhi 2012)

One limitation of these positioning sensors is the acquired data are limited to location and time information, which makes it hard to distinguish between productive and idling states of the equipment. In addition, these records do not provide enough information to estimate the quantities of the excavated soil or confirm that the trucks are fully loaded.

2.5.2 Vibration and orientation sensors

In contrast with the RTLS, the accelerometers detect the vibration signals of the equipment and the gyroscopes detect the orientations of the equipment. By processing the vibration and orientation data, the pose and the state of the equipment can be estimated. The IMU sensor consists of an accelerometer, a gyroscope, and a magnetometer, which can get the acceleration and orientation data of the equipment. The movement data collected by IMUs are usually classified with machine learning algorithms to identify the activity and the work cycle of the equipment. Ahn et al. (2012) mounted two accelerometers inside the cabin of the excavator; then, by observing the overall shape of the vibration signals (e.g. increasing and decreasing trends), three activities of working, idling, and engine-off were identified. Akhavian and Behzadan (2012) attached IMU sensors on the bed of a truck and the boom of a loader to estimate the activities based on the orientation and acceleration data (Figure 2-10). For example, an increasing boom angle to the

horizontal line and a constant bed angle close to zero indicated that the loader is raising its boom while the truck is waiting to be loaded. Akhavian and Behzadan (2015) used two smartphones to collect the accelerometer and gyroscope data of a loader. The raw data were represented by 12 features such as the mean, variance, peak, root mean values, etc. At last, SVM and CNN were used to classify the features and recognize the engine off, idling, moving, scooping and dumping activities of the loader. Similarly, Ahn et al. (2015) investigated the feasibility of measuring operation efficiency using accelerometer data for classifying engine-off, idling and working of the excavator operations. Mathur et al. (2015) attached a smart phone inside the cabin of the excavator to capture the engine vibration signatures in form of 3D acceleration. Then, a classifier was trained to measure the cycle time of the excavator. Kim et al. (Kim et al. 2018a) attached a smart phone with an inbuilt IMU sensor to the front window of the excavator, and used the Dynamitic Time Warping (DTW) algorithm to recognize its work cycle. Bae et al. (2019) used DTW to classify joystick signals of the excavator and identify activities. Hernandez et al. (2019) attached two IMU sensors on the body of a roller and recognized six work states with LSTM. Rashid and Louis (Rashid and Louis 2019) used IMU sensor and LSTM to recognize the activities of a loader. Rashid and Louis (2020) attached three IMU sensors to the bucket, stick and boom of an excavator to identify the movements, and used sliding window and artificial neural network to recognize activities. The results showed that the bucket is the best place to collect motion data in order to identify the activities of the excavator. Slaton et al. (2020) classified accelerometer data with CNN to distinguish six activities of the excavator.

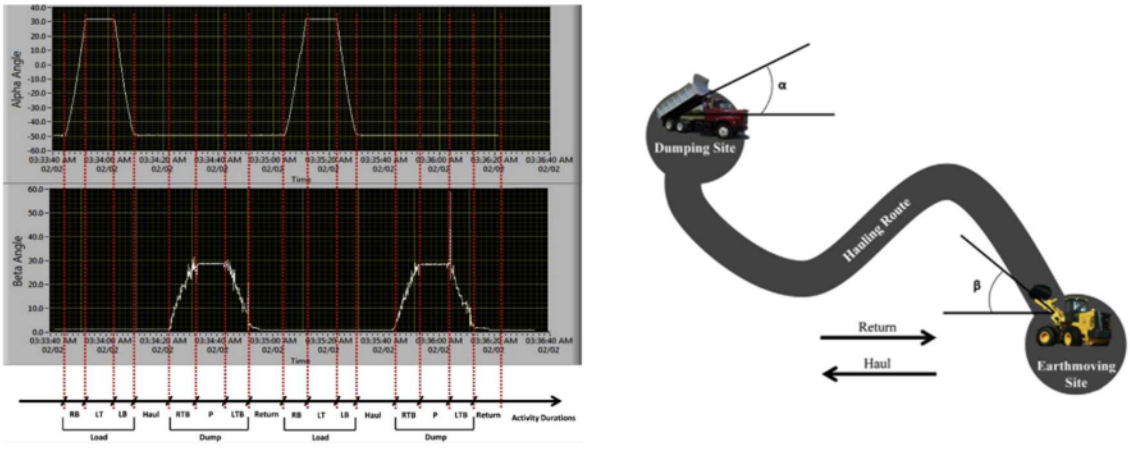
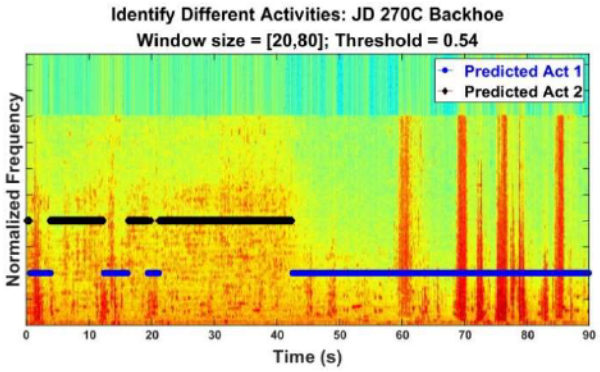


Figure 2-10 IMU sensor-based method (Akhavian and Behzadan 2012)

2.5.3 Audio sensors

In recent years, there have been developments in audio-based methods, which use the sounds generated from heavy equipment to recognize their activities, as shown in Figure 2-11. Similar to vibration signals, audio signals also contain different acoustic patterns of the equipment operation process. To identify the activity of the equipment from the signals, researchers generally applied four main processes: (1) collecting equipment sounds data from a microphone; (2) signal filtering or augmentation; (3) extracting features; and (4) training classification models (Santosh et al. 2010). Cheng et al. (2017; 2019) used audio singles to differentiate major and minor activities of excavators, loaders and dozers with Short-time Fourier Transform (SIFT) features and SVM classifier. Sabillon et al. (2020) used SIFT and Continuous Wavelet Transform (CWT) features combined with SVM classifier to recognize activities of the excavator and dozer. Then, a Markov chain filter was applied to analyze the cycle time and productivity of the equipment. The audio-based methods have limitations in crowded and noisy jobsites as it is difficult to recognize the equipment activities in detail.

(a) Microphone placing at the jobsite for collecting audio files



(b) Predicted activity label for the equipment

Figure 2-11 Application of audio-based activity recognition (Cheng et al. 2017)

2.5.4 Hybrid sensors

Hybrid sensors are used to get more accurate equipment operation information. Kim et al. (Kim et al. 2013) installed displacement and pressure sensors on each of the cylinders at the boom, arm and bucket to measure the movement, pose, and real bucket load of the excavator. Then, the simulation methods are used to optimize the bucket-tip path within the data retrieved from the sensors. Ibrahim and Moselhi (2014) used GPS to track the location of the equipment, strain gauges to measure the load weight of the truck, accelerometer to monitor loading and dumping activities of the loader and trucks, and barometric pressure to measure the weather condition. Based on the data collected with the sensors, they created an automated data processing algorithm to estimate the earthmoving productivity in near-real time. Sherafat et al. (2019) used IMU and microphones to collect vibration and audio data during excavators' operations. Then, they manually synchronized two kinds of data based on their similarity of signal spikes. In the test, SVM was applied to classify the fused data to recognize the activities of an excavator (stop, shove, move, turn), which achieved 20% higher accuracy than using only IMU or audio data.

Hybrid sensors were also used to monitor different productivity-related factors. Akhavian and Behzadan (2013) used UWB and attitude and heading reference system (AHRS) to capture the positions and boom angles of the truck, and Zigbee-enabled weight sensors to measure the weight of material transported by trucks. Vasenev et al. (2014) used GPS to track the locations of trucks, pavers, and rollers, and temperature sensors on the pavers to monitor the temperature of asphalt mat. Salem et al. (2018) collected data with GPS, IMU, soil water content sensors, and load cells to estimate the influence factors on the earthmoving productivity based on the soil condition, hauling and road condition, equipment operational condition and weather condition. The sensor-based equipment monitoring papers are categorized based on the type of sensors as shown in Table 2-3.

Table 2-3 Overview of sensor-based methods

Method	References	Sensors	Monitoring	Test	Result	Objectives		
						Activity duration	Productivity	Cycle time/number
RTLS	Han et al. (2008)	GPS	Truck: dump, load, haul, return	15 h, real project	96 % accuracy for productivity calculation	√	√	√
	Teizer et al. (2008)	UWB	Mobile crane: location	2 days, real project	The standard deviation for the position measurement was 0.09 m and 0.15 m in the X and Y direction, respectively			
	Cheng et al. (2011)	UWB	Equipment trajectory: mobile crane, tractor, material hauling trailer	43 min 22 s, real project	75% data within 2 m error, 99.9% data within 4 m error	√		
	Montaser et al. (2012)	GPS	Truck: load, dump, travel	8 h per day, 11 days, real project		√	√	√
	Montaser and Moselhi (2012)	RFID	Truck: load, travel, dump, return	51 min 37 s, real project		√		√
	Song and Eldin (2012)	GPS	Truck: load, haul, in work zone, wait, out work zone, return	Real project	Decrease cycle-time prediction error by 6%			√
	Pradhananga and Teizer (2013)	GPS	Steer loader: haul, return, unload, cycle, speed	12 h, real project	1.15 m – 0.36 m precision	√		√
	Montaser and Moselhi (2014)	GPS	Truck: load, travel, dump, return	12 days, real project		√	√	√
	Vahdatikhaki and Hammad (2014)	UWB	Excavator: relocation, swing to load, loading, swing to truck, dumping Truck: maneuvering for dumping, dumping, returning, maneuvering for loading, hauling	4 min 45 s, indoor library test	87% accuracy for activity duration estimation	√		√
	Vahdatikhaki et al. (2015)	UWB	Excavator pose estimation	91 s, indoor lab test	95% accuracy, 45-48 cm			
	Alshibani and Moselhi (2016)	GPS	Truck: load, haul, dump, return	131.47 min, real project		√	√	√
	Song et al. (2017)	GPS	Truck: enter, exit, load, haul, idle time	12 days, real project		√		√
	Louis and Dunston (2018)	GPS	Truck: haul, dump, return, wait, cycle time	800 min, real project		√	√	√
	Ahn et al. (2020)	GPS	Truck: pick-up loaded trailer, entering, inside, leaving	Real project	86% - 88% accuracy for activity duration estimation	√		
Vibration and orientation	Ahn et al. (2012)	Accelerators	Excavator: work, idle, engine-off	30 min, real project	79% accuracy for activity duration estimation	√		
	Akhavian and Behzadan (2012)	3D orientation, accelerometer	Loader and truck: load, dump, haul, return, idle	3 min, indoor library test		√		√
	Akhavian and Behzadan (2015)	Smart phone (IMU)	Loader: engine off, idle, moving, scooping, dumping	400 s, real project	90% accuracy for activity recognition	√		
	Ahn et al. (2015)	Smartphone (IMU)	Excavator: engine off, idle, work	100 s, real project	95% accuracy for activity recognition	√		

Table 2-3 Overview of sensor-based methods (continued)

Method	References	Sensors	Monitoring	Test	Result	Objectives		
						Activity duration	Productivity	Cycle time/number
Vibration and orientation	Mathur (2015)	Smartphone (IMU)	Excavator: idle, wheel-base motion, cabin rotation, bucket/arm movement	62.53 min, real project	74% accuracy for activity recognition	√		√
	Kim et al. (2018a)	Smartphone (IMU)	Excavator: rotating clockwise, rotating anti-clockwise, not rotating, work cycle	116 excavator's work cycle, real project	91.83% accuracy for work cycle recognition	√		√
	Bae et al. (Bae et al. 2019)	Joystick signals	Excavator: digging, leveling, trenching	70 min, real project	100 % accuracy for activity recognition	√		
	Hernandez et al. (2019)	IMU	Roller: forward high/low, backward high/low, forward off, backward off	20 min, real project	77.1% accuracy for activity recognition			
	Rashid and Louis (2019)	IMU	Excavator: engine off, idle, scoop, dump, swing loaded, swing empty, move forward, move backward, and level ground Loader: engine off, idle, scoop, raise, dump, lower, move forward loaded, move backward loaded, move forward empty, move backward empty	555 activities, real project	96.2% accuracy for activity recognition			
	Rashid and Louis (2020)	IMU	Excavator: engine off, idle, scooping, dumping, swing loaded, swing empty, moving forward, moving backward, ground levelling	4000 s, real project	92.1% accuracy for activity recognition			
	Slaton et al. (2020)	IMU	Excavator: idle, travel, scoop, drop, rotate (right, lift), various	30 min, real project	77.6% accuracy for activity recognition	√		
Audio	Cheng et al. (2017)	Microphone	Major activity, minor activity (excavator, backhoe, loader, dozer, hydraulic hammer, dumper, compactor)	90 s each machine, real project	85% - 90% accuracy for activity recognition	√		
	Cheng et al. (2019)	Microphone	Major activity, minor activity (excavator, loader, backhoe, dozer, mixer)	250 s, each machine	87% accuracy for productivity calculation	√		
	Sabillon et al. (2020)	Microphone	Excavator, loader, dozer, concrete mixer: major activity, minor activity	30 min, real project	Average 84% accuracy for activity recognition, Productivity predict 9.5-12.81% error	√	√	√
Hybrid	Akhavian and Behzadan (2013)	UWB, ZigBee, AHRS	Truck: load, haul, dump, return	60 s, indoor library test	95% confidence for activity recognition, 95% accuracy for productivity calculation	√	√	√

Table 2-3 Overview of sensor-based methods (continue)

Method	References	Sensors	Monitoring	Test	Result	Objectives		
						Activity duration	Productivity	Cycle time/number
Hybrid	Kim et al. (2013)	Displacement, pressure	Excavator torque trajectory	140 s, real project				
	Ibrahim and Moselhi (2014)	GPS, accelerometer, strain gauge, barometric pressure	Loader: hauling volume Truck: load, haul, dump, return	10.83 h, real project	98% accuracy for productivity calculation	√	√	√
	Vasenev et al. (2014)	GPS, IMU, temperature	Roller location (track equipment in the field)	16 h, real project	Mass of paved mixture: 1650 tons per day	√	√	
	Salem et al. (2018)	Pressure, moisture, humidity, luminosity, IMU, GPS, weather station	Influence factors of the earthmoving productivity	Data analysis				
	Sherafat et al. (2019)	Mobile, microphone	Test 1 Excavator: stop arm/shovel movement moving forward/backward turning right/left Test 2 (1) Excavator: stop, drilling, rotating/moving arm, moving forward/backward (2) Loader: stop, arm/shovel movement, moving forward/backward, turning right/left (3) Lift: maneuvering forward/backward, raising/lowering (4) Lift: stop, moving forward/backward, moving arm (5) Excavator: scraping and moving/rotating arm (6) Excavator: extending arm, rotating cabin (7) Dozer: stop, moving forward, moving backward (8) Concrete truck: pouring concrete, moving (9) Loader: stop, moving forward/pushing soil, moving backward (10) Vibrator: stop, vibrating	87 s, real project	Test 1 92% accuracy for activity recognition Test 2 Activity recognition accuracy: (1) 87.51%, (2) 92%, (3) 93.85%, (4) 87.2%, (5) 86.45%, (6) 86.32%, (7) 95.6%, (8) 93.61%, (9) 87.14%, (10)91.66%			

2.5.5 Sensor-based productivity analysis methods

The simulation and operation analysis are two widely used methods for productivity monitoring of sensor-based methods. The productivity is generally estimated in forms of activity durations, equipment cycle time, and soil volumes. The overview of productivity analysis of sensor-based methods is shown in Table 2-4. Simulation methods are widely used in sensor-based methods to capture the complex interactive activities between earthwork equipment and increase the accuracy of productivity analysis (AbouRizk 2010). These methods use the location or vibration data retrieved from sensors as input. Then, the productivity and cycle times are predicted by simulating the logic of the earthmoving equipment work cycle. Akhavian and Behzadan (2012) estimated the activities of a truck and a loader with IMU sensors, and simulated the work cycle of the two pieces of equipment with a discrete event simulation (DES) model based on the activity information. Some works (Louis and Dunston 2018; Song and Eldin 2012; Akhavian and Behzadan 2013; Montaser and Moselhi 2014; Alshibani and Moselhi 2016) used RTLS sensors to get the location, speed and direction of the trucks and further identified the hauling and returning times. With the predefined trucks' operation logic, they simulated and predicted the cycle time of the truck with the real-time data from sensors. Kim et al. (2013) used pressure and displacement sensors to track the moving trajectory of the excavator's boom and bucket. Then, they created a motion planning system to simulate and optimize the bucket-tip path of the excavator. Ibrahim and Moselhi (2014) developed a hardware to collect the location, vibration, soil type and weather data from GPS, accelerometer, strain gauge, and barometric pressure, respectively. Then they created activity recognition and volume calculation algorithms to calculate the earthmoving productivity based on the data collected from the sensors. Vahdatikhaki and Hammad (2014) proposed a rule-based simulation method, which can capture, process, analyze, filter and visualize the equipment operation processes with various RTLS data. Vasenev et al. (2014) proposed a framework to use data retrieved from GPS, IMU and temperature sensors monitoring for simulating the productivity of asphalt paving operations. The framework analyzed the number and work time of trucks, compactors working with each paver, and amount of asphalt mixture paved by the paver. Then, the position received from the sensors; and the productivity information are transferred into a server database to support future decision making. These simulation-based methods conducted real-time data-driven productivity analysis, which can concurrently monitor several pieces of equipment. Alshibani and Moselhi (2016) combined GPS with Geographic Information System (GIS) data to

get the trajectories of the trucks in earthmoving work. Song et al. (2017) collected data with different sensors to analyze the influence factors (e.g. humidity, wind speed, temperature, idle time average speed, etc.) on the earthmoving productivity using the linear regression method.

Other sensor-based methods analyzed productivity with the operation information of the equipment. For example, some research works (Akhavian and Behzadan 2015; Mathur et al. 2015) calculated the duration of all recognized activities and analyzed the productivity of the equipment with the portion of idling states in the total work time. Pradhananga and Teizer (2013) estimated the speed of the loader from its trajectory and stationary times in different working zones. Kim et al. (Kim et al. 2018a) estimated excavators' activities by classifying vibration data retrieved from IMU sensors, and determined work cycle of the excavator from the order of the activities. Bae et al. (Bae et al. 2019) recognized activities of the excavators and calculated the duration of digging and leveling in 20 min's work. Sabillon et al. (2020) analyzed cycle times and the number of cycles per hour from the activities. With the observed average fill factor, they calculated the productivity of the backhoe as the soil volume excavated per hour.

These sensor-based methods may have limited use in real construction projects because the sensor system could be difficult to install on the rented and old equipment. In addition, the installation and disassembly of sensor systems also requires labor and cost. This is especially true for the RFID sensor systems, which need comprehensive infrastructure to be installed on both equipment and jobsite. Moreover, construction sites have complex environments with noise, which may disturb or obstruct the wireless signal transmission and decrease the accuracy of the activity recognition results.

Table 2-4 Overview of productivity analysis of sensor-based methods

References	Type of study		Productivity monitoring		
	Simulation	Operation analysis	Activity duration	Cycle time/numbers	Soil volume
Akhavian and Behzadan (2012)	√		√	√	
Song and Eldin (2012)	√		√	√	
Akhavian and Behzadan (2013)	√		√	√	√
Pradhananga and Teizer (2013)		√	√	√	
Kim et al. (2013)	√				
Ibrahim and Moselhi (2014)	√		√	√	√
Montaser and Moselhi (2014)	√		√	√	√
Vahdatikhaki and Hammad (2014)	√				
Vasenev et al. (2014)	√		√		√
Ahn et al. (2015)		√	√		
Akhavian and Behzadan (2015)		√	√		
Mathur et al. (2015)		√	√	√	
Alshibani and Moselhi (2016)	√		√	√	√
Song et al. (2017)		√	√	√	
Louis and Dunston (2018)	√		√	√	√
Kim et al. (Kim et al. 2018a)		√	√	√	
Bae et al. (Bae et al. 2019)		√	√		
Sabillon et al. (2020)		√	√	√	√

2.6 Activity Recognition in Computer Vision

So far, numerous research works have been done in video activity recognition in the computer vision field. The existing methods can be generally divided into two categories. The first category uses hand-crafted local features to extract and represent the object's movement information in the videos (Wang et al. 2015a). This kind of feature is calculated by capturing motion and appearance features from the video. Effective features are considered to be discriminative for the target activates in the video and robust to rotation, occlusion, and background noise (Zhang et al. 2017). The second category relies on deep learning which trains deep neural networks to classify the activities in the videos.

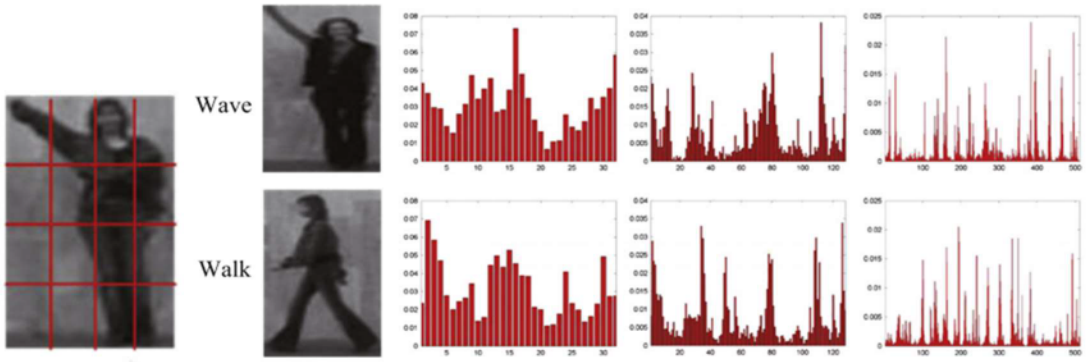
2.6.1 Feature-based methods

Most of the earlier works have used hand-crafted features in activity recognition. These works first extract the features which represent the appearance and/or motion character of the activity from the video frames, and then, encode the features into a fixed-length vector. At last a classifier is

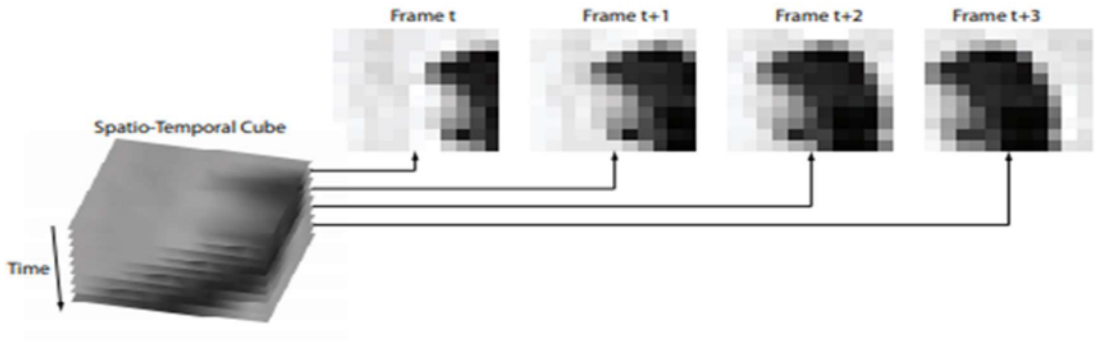
trained to identify the activities. General hand-crafted features consist of two types: regular spatial-based features and motion-based features.

(a) Spatial features

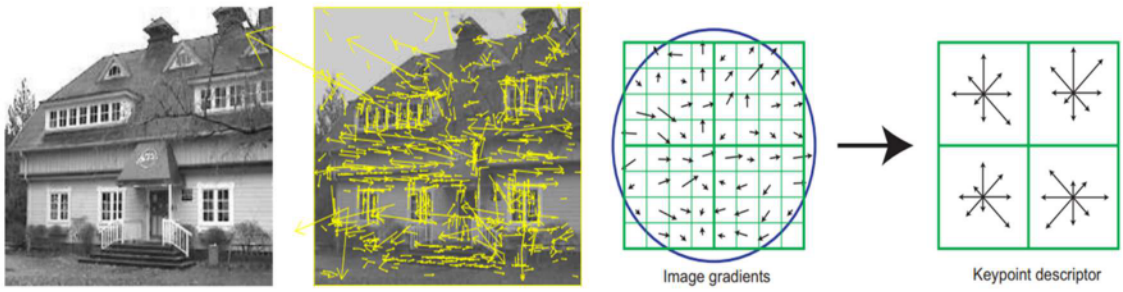
The spatial features are the ones that represent the spatial appearance, orientation and gradient of the activity in the video. A rich set of spatial-based features have made frequent appearance in the activity recognition field. For instance, the HOG (Dalal and Triggs 2004), 3D-HOG (Klaser et al. 2010) and SIFT feature, which are also widely used in equipment activity recognition as introduced in Section 2.4.3. HOG features capture the appearance information of the activity by calculating the gradient direction histogram of the local area of the image. Wang et al. (2013) (Figure 2-12(a)) used the HOG descriptor to encode the human figure in each frame of the video, and accurately recognize human activities in video sequences. 3D HOG is an extension of HOG feature to represent the activity in video sequences. It views the video as spatiotemporal volumes (Figure 2-12(b)) and calculates the spatiotemporal gradients to extract the shape and motion information of the activity in the video. Weinland et al. (2010) used the 3D HOG descriptor to recognize human activities. The 3D HOG features represent the information of the activities from the sequences of images that have been concatenated into data volume, which could achieve robustness to occlusion and viewpoint changes. SIFT (Loew 2004) descriptor (Figure 2-12(c)) uses polar coordinates to obtain the gradient magnitudes and orientations of the key points in the image. The 3D SIFT descriptor is an extension of the SIFT, which uses additional angle information to represent the direction of the gradient. Scovanner et al. (2007) used 3D SIFT to describe the spatiotemporal region of the frames in the video to classify the 10 types of human activities, which achieved improved performance over the previous methods.



(a) Example of ‘‘wave’’ and ‘‘walk’’ human figures and the corresponding HOG descriptors at different spatial levels (Dalal and Triggs 2004)



(b) Spatiotemporal volume of the video (Lowe 2004)



(c) Example of the SIFT feature descriptor (Lowe 2004)

Figure 2-12 Examples of the spatial-feature-base methods

(b) Motion features

Motion features are extracted from the optical flow, which can represent the motion of the objects in the video. The motion features mainly contain HOF, MBH, Dense Trajectory (DT) and improved Dense Trajectory.

Optical flow is a method to calculate the motion information of the object between adjacent frames. It uses the change of pixels in the time domain and the correlation between adjacent frames in the image sequence to find the corresponding relationship between the previous frame and the current frame. By separating the optical flow into vertical and horizontal components and blurring them with Gaussian, an artificial set of motion channels are created (Kang and Wildes 2016).

HOF captures the local motion of the pattern by quantifying the direction of the optical flow vectors (Kang and Wildes 2016). Laptev and Pérez (2007) recognized the human activities in movies with HOF feature combined with the SIFT feature to improve the performance. The MBH (Dalal et al. 2006) is calculated from local orientation and gradient of the optical flow. It can suppress the camera motion while preserving local relative motion of pixels (Kang and Wildes 2016).

Another form of the descriptor is the DT descriptor, which is a kind of descriptor that tracks the path of sampled feature points in the video frames over time. The DT descriptor is shown in Figure 2-13. DT first requires dense sampling the feature points of each frame in the video clips, and then, the sampled points are tracked using optical flow to get the trajectory of the motion (Kang and Wildes 2016). At last, the activity is recognized with the classification of the trajectories. In the DT-based methods, feature points sampled from the frames using HOG, HOF and MBH features have shown to be successful on a number of challenging datasets (Wang et al. 2015a). Yang et al. (2016a) used HOG, HOF and MBH features with DT to recognize the worker's activity on the construction site, and the results showed that the MBH feature got the best recognition performance. Numerous approaches have been done to improve the performance of DT-based method. Eleonora and David (2012) used saliency-mapping algorithms with DT features to represent the salient region in more detail. The improved method conducted a more compact video representation, and improved action recognition accuracy. Jiang et al. (2012) integrated trajectory descriptors with the pairwise trajectory locations to build a simple method to separate foreground from background, and eventually, got a robust activity recognition performance. The improved dense trajectory (Wang and Schmid 2013) is another improvement of the DT feature. It removed the trajectories

consistent with the camera motion, and significantly improved the recognition performance. Current studies show that improved dense trajectory descriptors and Fisher Vector (Willems et al. 2008) representations yield superior performance on various datasets (Zhang et al. 2016). Similarly, Wang et al. (2016) improved the DT by removing the background trajectories and using warping optical flow, which achieved better recognition results.

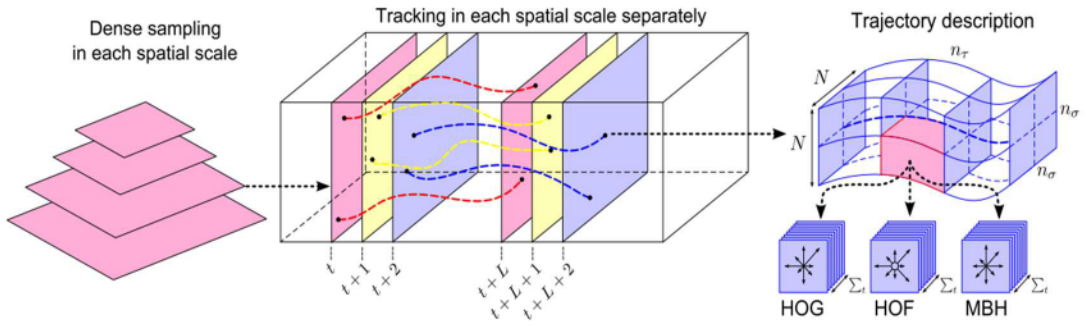


Figure 2-13 Illustration of dense trajectory description (Wang et al. 2013)

2.6.2 CNN-based methods

In recent years, with the development and application of deep learning methods in image classification tasks, there have been a lot of attempts to implement deep learning methods for video activity recognition. The deep learning methods could achieve high accuracy in activity recognition. Furthermore, they are much simpler to use than the traditional feature-based methods since the representation of the spatial and temporal features can be automatically learned by the CNN. In deep learning, a deep neural network with multiple layers is built up for automated feature extraction. Specifically, each layer in the neural network performs a non-linear transformation on the outputs of the previous layer, so that through the deep learning models the data are represented by a hierarchy of features from low-level to high-level. The well-known deep learning models include CNN, RNN, LSTM, etc. There are three mainstream methods in deep convolutional neural network applications in video activity recognition: two-stream methods (Simonyan and Zisserman 2014), deep 3D CNN (Tran et al. 2015) methods and the combination of multiple features methods.

(a) Two-stream methods

Simonyan and Zisserman (2014) first proposed the two-stream method. In their two-stream method, two CNNs were used to extract the special and temporal information from the frames and optical flows of a video separately. Figure 2-14 presents the architecture of the two-stream neural network.

The spatial stream CNN took single RGB frames as input to extract the appearance representation of the activity in videos; and the temporal stream CNN, with the same structure as the spatial one, took several optical flows as input to describe the temporal features between consecutive video frames. At last, the two CNNs were fused by averaging the softmax scores to provide the final activity recognition results. The result showed the method of Simonyan and Zisserman (2014) achieved the accuracy of 88% on testing human activities.

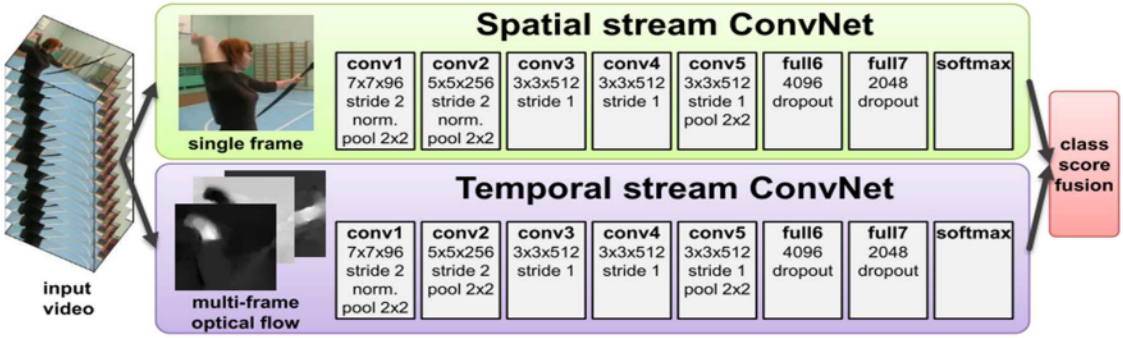
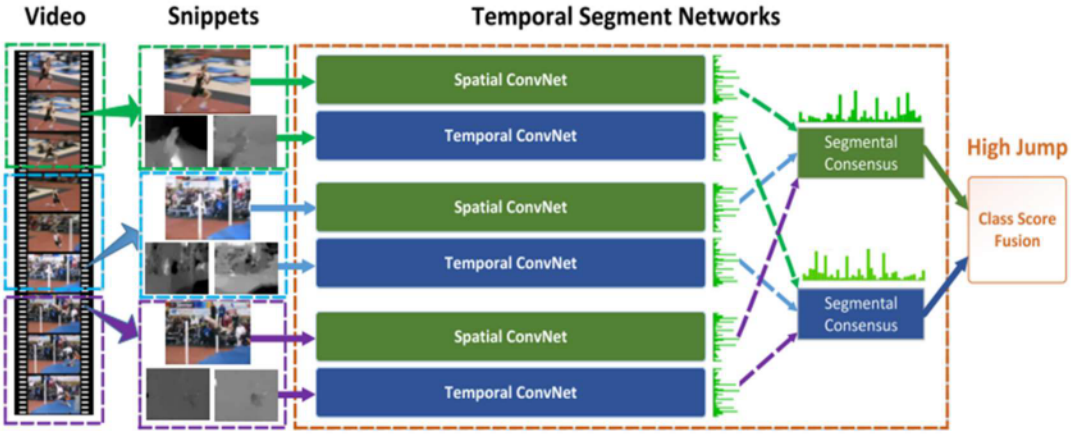


Figure 2-14 Illustration of the two-stream architecture (Simonyan and Zisserman 2014)

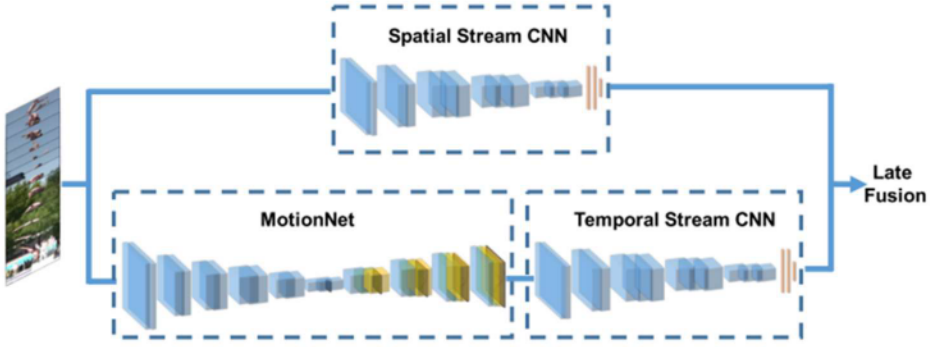
Based on the original regular two-stream CNN method, numerous variations and improvements have been produced. Some of the methods focused on adjusting the structure of the network to improve activity recognition accuracy. For instance, Wang et al. (Wang et al. 2015b) applied the data augmentation, multi-GPU, pre-training and smaller learning rates techniques to train a very deep two-stream CNN, which improved the performance of the original two-stream method to 91.4% accuracy. Donahue et al. (2017) employed the RNN that used LSTM cells instead of the widely used CNN in the two-stream method to operate the special and temporal flow. The recurrent network with LSTM cells has an advantage in learning the long-range temporal information from the video sequences.

Some methods focused on adjusting the input source of the network to improve the activity recognition performance. Wang et al. (2016) presented the novel Temporal Segment Network (TSN) (Figure 2-15(a)), which could improve the efficiency of the original two-stream network. Instead of operating all the frames in the video, TSN worked on a sequence of short snippets sparsely sampled from the input video clip, such that it could efficiently capture the temporal feature of the video with the lower computational resource. Since the most computation-intensive part in the two-stream approach comes from optical flow calculation, Zhang et al. (2016) replaced

the optical flow with the motion vector in the original two-stream method. The test result showed that this real-time CNN method is 27 times faster than the original two-stream method. Zhu et al. (2017) presented a hidden two-stream method (Figure 2-15(b)) It used a fully convolutional neural network implicitly captured the motion information between consecutive frames instead of explicitly computing the optical flow. The experimental result showed that the hidden two-stream method was 10 times faster than the traditional method. Although the two-stream methods could achieve high accuracy in activity recognition, they are not efficient to use because they take long time and huge computation resources to extract the optical flows in the videos.



(a) Temporal segment network (Wang et al. 2016)



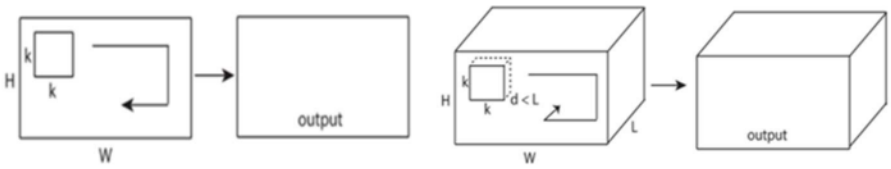
(b) Hidden two-stream network (Zhu et al. 2017)

Figure 2-15 Examples of improved two-stream network

(b) Deep 3D CNN methods

3D CNN (Tran et al. 2015) is a deep neural network, which can learn the spatial-temporal features of the video from merely RGB images. In this method, a 3D convolutional neural network is trained

for activity recognition. Tran et al. (2015) first proposed the deep 3D CNN architecture, which used the 3D convolutional and pooling kennels to capture the appearance and motion information of the activity from the video sequences, as shown in Figure 2-16. Then, the model was trained on a large-scale video dataset to recognize the activities. In terms of performance, this original 3D CNN-based method is inferior to the original two-stream method (Wang et al. 2016), but it is much more efficient than the two-stream method. Specifically, the 3D CNN method could process 672 frames per second, which is about 30 times faster than the two-stream method.



(a) 2D convolution (b) 3D convolution

Figure 2-16 2D and 3D convolutional operations

Accordingly, numerous works have been done to improve the 3D CNN method. Hara et al. (2018) created a 101 layers' 3D convolutional network based on the residual network architecture (Wu et al. 2017), which verified that the deep 3D network could achieve state-of-the-art performance in video activity recognition. Varol et al. (2018) created a 3D neural network with the long-term temporal convolutions, which could capture the long-temporal features of the activity. Diba et al. (2017) proposed a novel deep 3D network with variable 3D convolution kernel depths. This architecture was designed to concatenate temporal features-maps extracted at different temporal depth ranges, rather than fixed 3D homogeneous kernel depths, and it achieved good performance in both long and short video datasets.

(c) Other deep learning methods

There are also well-performed deep learning methods for activity recognition that couldn't be categorized into the previous two kinds of methods. Feichtenhofer et al. (2016) (Figure 2-17) combined the two-stream method with the 3D CNN method, which could keep both the special and temporal features extracted from these two methods. Zolfaghari et al. (2017) (Figure 2-18) created a three-stream 3D CNN architecture, which has three types of input images, such as Red, Green, and Blue (RGB) image, optical flow and sematic segmented pose images. Each type of the input

image was operated with a 3D CNN, and fused with a Markov chain model at last. The method improved the activity recognition performance over the baseline methods, and illustrated the contribution of the pose feature.

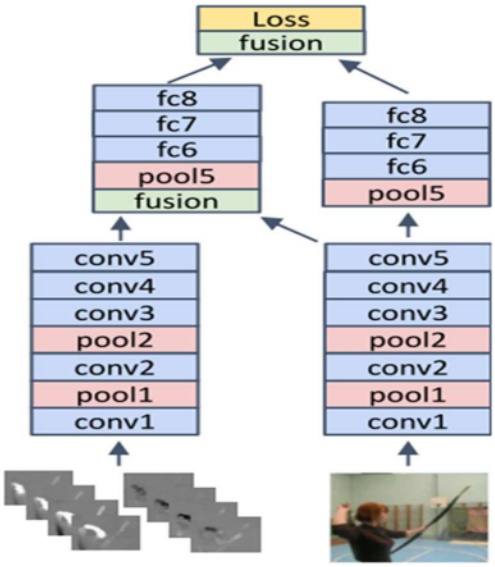


Figure 2-17 Two-stream and 3D fusion method (Feichtenhofer et al. 2016)

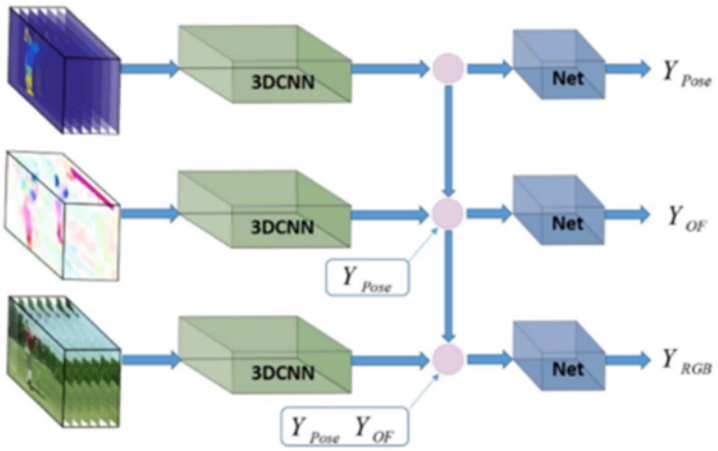


Figure 2-18 Three-stream 3D-CNN architecture (Zolfaghari et al. 2017)

Girdhar and Ramanan (2017) introduced the attention model to the activity recognition task, which replaced the pooling layer in a standard CNN with a weighted polling layer. The attention model could learn the specific region of the input image that is most relevant to the task, and it has produced competitive results on widely benchmarked datasets. Zang et al. (2018) (Figure 2-19) proposed an attention-based CNN, which embeds the attention model into a multi-stream CNN to recognize human activity. This novel network takes RGB, optical flow and wrapped flow images

as input, and an attention layer in the network is used to effectively reducing the negative effect of redundant information/noises to get a better recognition performance. The proposed method has achieved the accuracy of 94.6% and 70.5% for human activity recognition on two challenging video datasets of UCF-101 (Soomro et al. 2012) and HMDB-51 (Kuehne et al. 2013).

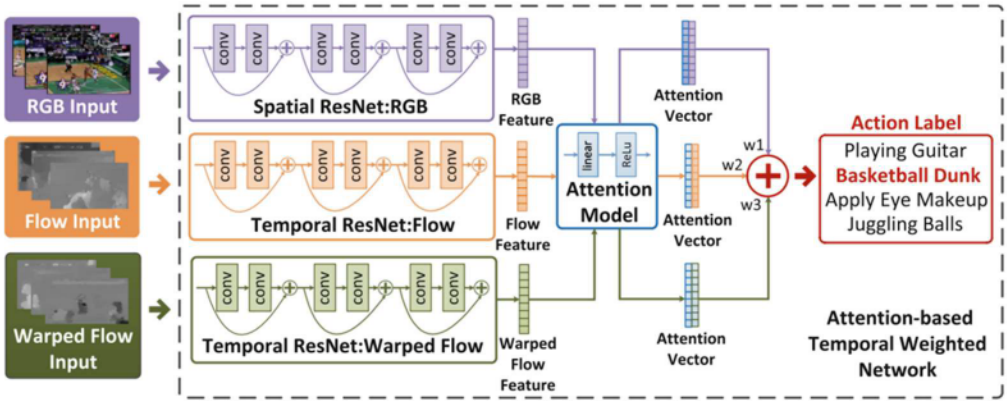


Figure 2-19 Network architecture of attention-based method (Zang et al. 2018)

Although deep learning methods can automatically learn the representation features of the activity in the video, they ignore the intrinsic difference between the spatial and temporal domains (Wang et al. 2015a). Therefore, Wang et al. (2015a) presented a novel trajectory-pooled deep-convolutional descriptor method, which incorporated the hand-crafted feature with the deep-learned feature by using the strategy of trajectory-constrained sampling and pooling. Shi et al. (2017) implemented the DT features with the CNN to descript the long-term video actions, which achieved good performance. The current research shows that the combination of DT or improved DT with the deep learning-based method could always get improved performance in video activity recognition.

2.7. Productivity-related Factors Monitoring

The previous literature review (Section 2.4.4 and Section 2.5.5) indicated that various factors affect the equipment's productivity. CV-based and sensor-based methods have their advantages in monitoring different productivity influence factors which are compared in this section.

2.7.1 Impact factors of equipment productivity

The construction equipment usually operates in different and complex construction environments. The operation conditions have a strong influence on the productivity of earthmoving equipment.

In order to measure and predict productivity more accurately, numerous research works have been conducted on exploring the key factors that have effects on equipment productivity. Smith (Smith 1999) created a linear regression model to estimate the productivity of earthmoving equipment and indicated that the bucket volume, truck travel time, number of trucks and haul length are the main factors that influence earthmoving productivity. Edwards and Griffiths (2000) used a two-layer CNN to predict the excavator's output based on the cycle time. They found that the machine weight, digging depth and machine swing angle (i.e. the angle between digging and loading points) are the three main productivity influence factors. Similarly, swing angle, machine weight and dig depth were also identified as the main factors of excavator's productivity by Edwards and Holt (2000) and Yang et al. (2003). Tam et al. (2002) predicted the productivity of excavators with a feedforward neural network based on the factors of relative positions between excavators and materials, site obstructions, operator's skill and type of soil. Hola and Schabowicz (2010) combined the feed-forward propagation network with a conjugate gradient algorithm, and predicted excavation productivity based on the number of excavators and trucks, excavator bucket volume, truck loading capacity and type of load material. There are also research works that analyzed the influence of productivity based on one specific factor. Moselhi et al. (2011) created a decision support system called WEATHER to estimate the impact of weather conditions on equipment productivity. Yoon et al. (2014) observed and compared the loading time of the excavators with respect to the relative position with trucks, and indicated that the horizontal and vertical distances, and swing angles between the excavator and truck have influence on the loading time. Holt and Edwards (2015) indicated that the excavator operator's ability is an important excavator's productivity factor, which could impact other factors. They analyzed the relationship between the operator's skill levels with excavator's productivity using Caterpillar hydraulic excavator productivity model. Then, they increased the productivity estimation accuracy of the model by adding the operator ability factor. Manyele (2017) used a computer dispatch system to explore the main factors that influence excavators' and trucks' productivity in mine projects. The test results have shown that the matching between the excavator and truck (i.e. loading position, truck-shovel combinations) is the major factor affecting the loading efficiency, and hence productivity. Salem et al. (2018) used experts' investigation and fuzzy model methods to identify and evaluate the factors affecting the productivity of earthmoving works. The water content of soil, operators' skills, snowy road condition, foggy weather, and waiting durations were identified as the most important

factors of productivity. Seresht and Faye (2019) explored the most influential factors using interview surveys with construction experts. They identified the positive and negative factors related to labor and equipment productivities, among which personal protective equipment, past experience of crew with project configurations, and equipment operator experience are the top three positive effect factors; oil price and its fluctuations, weather conditions and global economic outlook are the top three negative factors. Based on the review, the factors that most influence equipment productivity are listed in Table 2-5.

Table 2-5 Influence factors of equipment productivity

Main factors	Influence factors
Excavator	Operator ability
	Operator fatigue
	Bucket volume
	Relative positions between excavator and material
	Dig depth
	Mechanical problems
Truck	Bed volume
	Travel time (e.g. traffic conditions)
	Haul length, slops, road quality
	Driver fatigue
	Mechanical problems
Fleet	Number of trucks working with an excavator
Relative location of truck with respect to excavator	Swing angle
	Relative height
Site conditions	Type of soil
	Site congestion
Weather	Weather

2.7.2 Productivity impact factors analysis with CV-based and sensor-based methods

Some of the reviewed papers have focused on analyzing the influence of different factors on equipment productivity using the data collected by sensors and cameras from real construction projects. Both the sensor and CV-based methods have their advantages and disadvantages in monitoring different kinds of influential factors. The comparison of the factors that can be monitored with the sensor-based and CV-based methods are shown in Table 2-6.

Table 2-6 Impact factors analysis between CV and sensor-based methods

Methods	References	Excavator				Truck					Site condition		Weather
		Operator fatigue	Bucket volume	Idling time	Cycle time	Cycle time (load, haul, idle, dump)	Bed volume	Travel time (e.g. traffic condition)	Driver skill	Number of trucks	Material type	Site congestion	temperature, wind, humidity
Sensor-based methods	Han et al. (2008)					√				√	Δ		
	Montaser et al. (2012)					√	Δ	√		√	Δ		
	Song and Eldin (2012)					√							
	Kim et al. (2013)		√										
	Ahn et al. (2015)			√									
	Ibrahim and Moselhi (2014)						√		√	√			√
	Montaser and Moselhi (2014)					√		√					
	Vasenev et al. (2014)						√						
	Alshibani and Moselhi (2016)					√				√			
	Song et al. (2017)					√		√	√		Δ	Δ	√
	Salem et al. (2018)			√		√	√	√			√		√
	Louis and Dunston (2018)		Δ		Δ	√				√			
	Sabillon et al. (2020)		Δ		√								
Vision-based methods	Zou and Kim (2007)			√									
	Azar et al. (2013)				√	√							
	Bügler et al. (2014; 2017)		Δ		√	√							
	Kim et al. (2018b)			√	√	√				√			
	Kim et al. (Kim et al. 2018c)			√	√								
	Li et al. (2019; 2020)	√											
	Kim et al. (2019)					√		√		√			
	Kim et al. (2020)					√				√			
	Chen et al. (2020a)			√	√								
	Chen et al. (2020b)			√									

√ The factor is automatically monitored

Δ The factor is manually monitored

In the sensor-based methods, GPS and IMU sensors were widely used to estimate the travel condition (Montaser and Moselhi 2014), idling time (Salem et al. 2018; Ahn et al. 2015), number of trucks (Alshibani and Moselhi 2016), the cycle time of the equipment, and the speed (MontaserAli et al. 2012; Sabillon et al. 2020) of the equipment. Typically, the speed of trucks was also used for monitoring the skill of truck drivers (Ibrahim and Moselhi 2014; Song et al. 2017).

Pressure measurement sensors can be used to measure the soil volume in the bed of trucks (Ibrahim and Moselhi 2014; Vasenev et al. 2014) and in the bucket of excavators (Kim et al. 2013). In addition, weather conditions such as the temperature, wind speed, and humidity were also monitored by sensors to estimate their influence on equipment productivity (Ibrahim and Moselhi 2014; Song et al. 2017). Some sensor-based methods also take the material type (MontaserAli et al. 2012; Han et al. 2008) and site congestion (Song et al. 2017) into consideration. The related information was observed by the researchers but not collected by sensors.

The CV-based methods monitor the number of excavators and trucks with detection and tracking methods (Kim et al. 2018c; Kim and Chi 2020). The cycle time, travel time, and idling time of the equipment can be monitored with activity recognition methods (Azar et al. 2013; Zou and Kim 2007; Kim et al. 2018c; Chen et al. 2020a). Specifically, the travel time and cycle time of the truck was also monitored with plate number recognition (Kim et al. 2019). In addition, Chen et al. (Chen et al. 2020b) estimated idling reasons of excavators by analyzing the work states between excavators and trucks. Li et al. (2019; 2020) monitored the mental fatigue of the operator by tracking the eye-movement with eye tracking cameras.

Figure 2-20 shows the factors that were monitored in the reviewed papers. Truck cycle time (22%), number of trucks (13%), excavator cycle time (12%), excavator idling time (11%), and truck travel time (8%) are the most frequently studied factors. Other factors, such as the relative position of the excavator with respect to the truck and soil, swing angle, relative height between excavator and truck, and site congestion are difficult for both the sensors and cameras to capture and have not been studied with any automatic method.

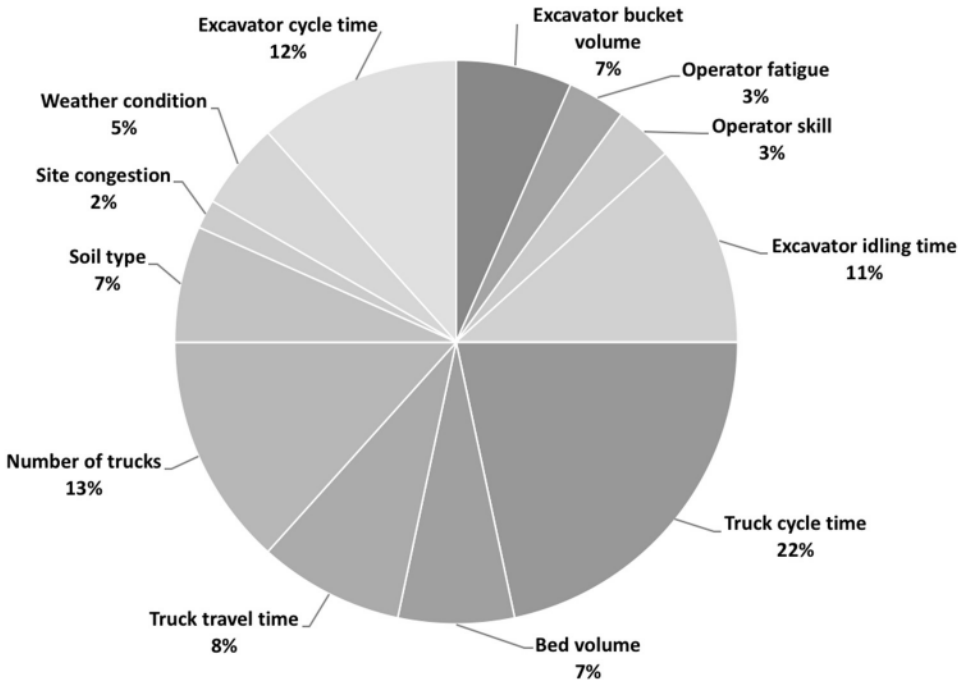


Figure 2-20 Factors monitoring in reviewed papers

2.8 Roadmap

Hannus et al. (2003) proposed "Construction Information and Communication Technology (ICT) Roadmap", which shows the developments of new and emerging technologies towards information and communication technology of construction. Following Hannus's concept, a roadmap about the automatic equipment productivity monitoring is proposed in this section. The proposed roadmap focuses on novel research topics of construction equipment productivity management and indicates future works for research and the industry to take up based on current technologies. As shown in Figure 2-21, there are mainly three steps toward equipment productivity management at different time spans: (1) Take-up: leveraging existing technologies; (2) Research and Development (R&D): developing new technologies that are clearly defined and under explorations; and (3) Emerging: future works that need potential solutions. The following paragraphs will explain the proposed roadmap starting from the take-up technologies to the emerging ones.

The take-up technologies have been thoroughly reviewed in this research including: field data capturing (i.e. CV-based and sensor-based), equipment operation monitoring and equipment activity recognition (Sections 2.4 and 2.5), and productivity influence factor analysis (Section 2.7).

However, these take-up technologies still have limitations, and further R&D is needed to overcome these limitations as explained in the following.

- **Equipment detection and tracking.** As reviewed in Sections 2.4.1 and 2.4.2, CNN-based methods have been widely used for equipment detection and tracking. However, the performance of these methods is highly influenced by the volume of training dataset, light condition and objects occlusions. Therefore, to overcome these limitations, existing detection and tracking methods still need to be improved. For example, Kim et al. (2020) proposed database-free method, which aims to minimize the volume of training data and cost of human labeling, and maximize the detection performance. In their work, a deep active learning approach was applied to automatically evaluate and label the unlabeled data based on a small amount (10%) of manually labelled data.
- **Equipment ID identification with CV-based methods**. Existing CV-based methods can only detect equipment type, without a specific ID for each equipment, and then use tracking methods to identify specific equipment. However, tracking methods have ID switch and jump issues when objects are occluded. Some works (Azar 2016; Feng et al. 2015; Kim et al. 2019) used marker or license number to identify equipment ID, which are limited in far field conditions. CV-based ID identification is still in the development stage.
- **End-to-end activity recognition.** In order to recognize activities of multiple pieces of equipment, existing methods have to use detection and tracking methods to identify and localize the equipment. In CV, single-stage methods have been proposed to localize and recognize human activities simultaneously. Köpüklü et al. (2020) created You Only Watch Once (YOWO) network, which integrated 2D and 3D CNNs, channel fusion and attention module, and bounding box regression module to localize and recognize 24 classes of human activities (UCF101-24 dataset) with the accuracy of 80.4%. Liu et al. (2020) improved YOWO by adding the knowledge distillation method, which can learn long range motion features provided by the optical flow. The method has been tested on UCF101-24 dataset and achieved the current best accuracy of 83.5%.
- **Productivity analysis.** Most of the existing papers monitor productivity by calculating the earthmoving productivity or estimating the idling time of the equipment. However, in order to improve the productivity of the equipment, the reasons of low productivity should be analyzed

in detail based on the equipment operation processes. For example, in order to reduce the idling time, this research (Chapter 5) analyzed the reasons that caused excavators idling by observing the interactive work states between excavators and trucks.

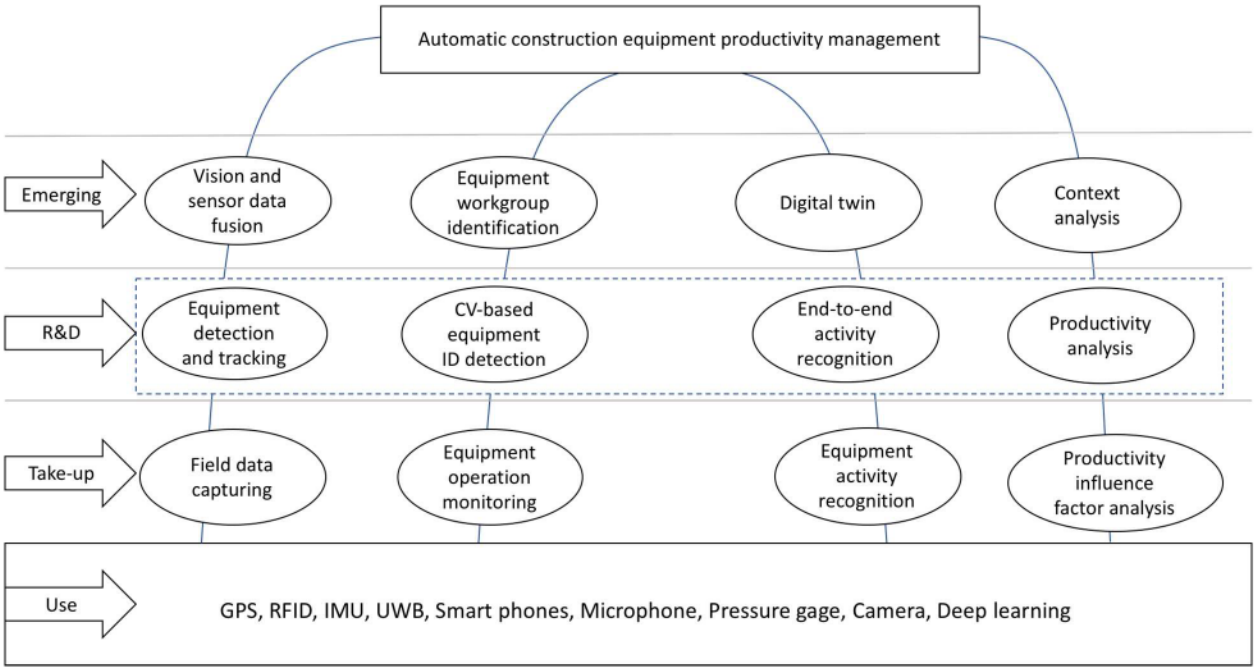


Figure 2-21 Roadmap for automatic equipment productivity monitoring

Furthermore, the integration of the R&D technologies can be further developed to fill the needs of more advanced emerging technologies:

- **Data fusion technology**. Data fusion is a multidisciplinary research area that can combine advantages from multiple resources, enhance the reliability of the measurements and improve the results (Shahandashti et al. 2011). Most of the existing works use either CV-based or sensor-based methods, which have limitations in the application. Some researchers have attempted to investigate the opportunities of fusing both methods. For example, Sonltani et al. (2018) indicated that the CV-based method for equipment monitoring cannot get the accurate 3D pose information, which is essential for safety control. They calibrated the camera to the 3D real coordinates of the excavator and fused the GPS data with camera data by aligning their coordinates. As mentioned in the R&D section, the equipment detection and tracking technologies still have limitations and are under development. Data fusion technology may help overcome such limitations of objects occlusion and imitated training dataset. To fill the gap of ID switch and fragmentation errors when objects are occluded in CV-based tracking methods,

Cai and Cai (2020) fused the tracking results of CV-based and sensor-based methods by matching the 3D locations. They first calibrated the stereo cameras and converted the 2D coordinates of the workers detected in the video frame into 3D coordinates. Accordingly, the vision-based locations were matched with the sensor-based locations in 3D coordinates system. Chen et al. (2019) fused the CV-based workers' posture data and UWB-based position data with a risk matrix. Then, they evaluated the risk of the work based on the positions of the workers. All aspects of data fusion should be studied in the future to benefit the complementary data sources (i.e. CV and sensor).

- **Equipment workgroup identification**. Many researchers have indicated the importance of interactive operation analysis between several pieces of equipment in earthmoving work (Azar et al. 2013; Bügler et al. 2017). The interactive work analysis can provide comprehensive explanation about the equipment productivity in complex interaction scenarios (Kim et al. 2018d). Existing works identified loading activity and work cycle of the equipment based on analyzing the interactive work between excavators and trucks (Bügler et al. 2017). However, they focused on the specific number of equipment that can be viewed in video frames, which is not sufficient to be used in real construction scenarios for analyzing the complete fleet of equipment. With the development of the equipment ID identification, the workgroup identification can be performed on specific equipment, which will add more practical value to equipment productivity monitoring.
- **Digital twin**. The digital twin is new concept of smart manufacturing emerging with the rise of Industry 4.0, which focuses on reflecting the lifecycle of physical products into virtual data by using information technologies such as sensors, simulation, and databases (Haag and Anderl 2018). As discussed in Section 2.7, the productivity of equipment is influenced by various factors. The digital twin can be used to simulate the actual behavior of the equipment in a virtual environment based on the real-time activity recognition results. Based on the simulation of the entire operation processes, the factors that may cause low productivity can be found and avoided in the real project. BIM is a virtual platform which contains digital information of buildings for enhancing planning, construction, and maintenance over the life cycle. Khajavi et al. (2019) installed sensors on building façade to collect temperature, humidity and light data and reflected them in the BIM model with the digital twin technology. They indicated that, in

the future, the digital twin may be used to monitor more construction-related issues, such as equipment productivity, using BIM.

- **Context analysis.** From the previous review in Sections 2.4 and 2.5, both CV-based and sensor-based methods could extract limited information of the equipment (e.g. location, trajectory, weather, human factors, and activity) directly from the videos or sensor data. The summary of factors that can be automatically monitored is shown in Table 2-7. In order to have a comprehensive understanding of the equipment operation conditions, context analysis is necessary. For example, to improve equipment productivity, the reasons for low productivity have to be identified beyond the reasons related to the interaction between the equipment, which requires analyzing complex context conditions, such as the work states, operator's performance, site condition, weather condition, etc. From the review Section 2.7, some factors have been identified to have important impacts on equipment productivity, but have not been analyzed using automated methods (e.g. the relative position between the excavator and digging point, swing angle, relative height between excavator and truck, and site congestion). With the development of the data fusion methods and context analysis, these factors can be monitored and analyzed automatically in future works.

Table 2-7 Factors which can be monitored with CV-based and sensor-based methods

Main factors	Influence factors	CV-based methods	Sensor-based methods
Excavator	Operator ability	√	√
	Operator fatigue	√	√
	Bucket volume		√
	Relative positions between excavator and material	√	√
	Dig depth		√
	Mechanical problems		√
Truck	Bed volume		√
	Travel time (e.g. traffic conditions)	√	√
	Haul length, slops, road quality		√
	Driver fatigue	√	√
	Mechanical problems		√
Fleet	Number of trucks working with an excavator	√	√
Relative location of truck with respect to excavator	Swing angle		√
	Relative height		√
Site conditions	Type of soil		√
	Site congestion	√	√
Weather	Weather		√

2.9 Summary

The literature was reviewed in fields pertinent to CV-based methods for equipment operation monitoring, sensor-based methods for equipment operation monitoring, influence factors of earthmoving work, and activity recognition methods in CV.

The main findings of this review are: (1) The recent works of automatic equipment monitoring are grouped in two categories of CV-based methods and sensor-based methods considering the data collection methods. The advantages and limitations of each method have been discussed in detail; and (2) The existing methods of equipment productivity monitoring are summarized and extensively discussed from three aspects: (a) estimating the time of a specific activity (e.g. idling, loading, digging etc.), (b) calculating the productivity of earthmoving in soil volume, and (c)

analyzing the factor that effect to the productivity. In order to get the productivity information, the CV-based methods focused on recognizing the activity and number of cycles of the equipment. Most of the sensor-based methods calculated the locations and poses of the equipment to estimate their work states. The audio methods analyzed the work state based on the sound of the engine.

After the literature review, the limitations and gaps in current research works are identified accordingly:

- The existing CV-based methods are not efficient to use in the real construction scenario. The feature-based methods are computationally intensive and cannot be used to recognize consecutive activities in long videos. Also, the detection-tracking-based methods need to pre-define the threshold and adjust it with the change of the camera, which is not efficient to be used.
- The precision of the detection-tracking-based methods is not enough for further data analysis. These methods are unable to identify the detailed activity of the equipment, which is essential for productivity calculation and analysis.
- Most of the existing studies are focusing on monitoring or recognizing the activity of one single piece of equipment. These methods are not sufficient to be applied in real construction projects, which have numerous equipment working simultaneously.
- Previous studies usually focused on equipment activity recognition, however, limited research work has been conducted to analyze the productivity of the equipment.
- There is no related research work focusing on exploring the reasons that cause equipment's low productivity with the CV-based method, which is important for analyzing the operation performance and improving the productivity of the equipment.

Futhermore, a roadmap is proposed to show the advances in each automatic equipment productivity monitoring method, and future research directions of this domain are proposed. The roadmap shows the research paths towards more comprehensive equipment productivity management by integrating different research areas and leveraging new advancements in computer science.

CHAPTER 3: OVERVIEW OF THE PROPOSED RESEARCH FRAMEWORK

3.1 Introduction

In Chapter 2, automatic methods for construction equipment monitoring and productivity analysis were reviewed, and a roadmap is proposed to explore the future research directions for automatic equipment productivity monitoring. This chapter provides an overview of the proposed research framework. For excavators' operation monitoring, a state-of-the-art 3D CNN and sliding window method are used to recognize the activities of the excavators. Then, the activity information is used to analyze productivity. Moreover, in order to improve the earthmoving productivity, the workgroup of excavators and trucks are identified to analyze the idling reasons of the excavator.

3.2 Overview of The Research Framework

Figure 3-1 shows an overview of the proposed framework, which has three components. First, a comparison between sensor-based and CV-based methods for equipment productivity monitoring is conducted, and a roadmap is created to explore the future work directions of automated equipment productivity monitoring. Then, the activity of multiples excavators in the site surveillance videos is automatically recognized. Finally, the time of each activity of the excavator is analyzed to calculate its working cycle and productivity. To improve the productivity of the excavator by reducing its idling time, the idling reasons of the excavator are further identified based on the interactive operations between excavators and trucks.

3.2.1 Roadmap of automatic equipment productivity monitoring

The objective of this component is to propose a roadmap to illustrate the technology path forward for automatic equipment productivity monitoring and provide future directions that will support the development of full automation in monitoring construction equipment. The main tasks of this component include:

(1) Comparing the automatic equipment productivity monitoring methods.

(2) Proposing a roadmap of equipment productivity monitoring.

The details of this component are given in Chapter 2.

3.2.2 Excavators activity recognition and productivity analysis

The objective of this component is to propose a method, which can recognize the activities and analyze the productivity of excavators in the long video sequences. The proposed method first integrates the detection and tracking technologies to localize all excavators in the video frames. Second, by analyzing the changes of centroid coordinates and areas of bounding boxes of the excavators in consecutive frames, idling states are identified. Then, a neural network is trained to recognize the activities of each excavator into detailed categories (i.e. digging, loading, and swinging). Finally, the productivity related indicators, such as the time of each activity, the cycle numbers of the excavator, the duration of each working cycle, and the productivity of the excavator, are calculated based on the activity recognition results. The main tasks of this component include:

(1) Detecting and tracking the excavators and trucks in video frames;

(2) Recognizing the idling state of the excavator with the sliding window method;

(3) Applying the 3D CNN method to recognize three types of activities of the excavator in the video (i.e. digging, loading and swinging);

(4) Calculating the duration of each activity, the number of work cycles, and the productivity of the excavator;

(5) Testing the proposed method in the videos recorded from real construction sites, and evaluating its performance.

The details of this component are given in Chapter 4.

3.2.3 Idling reasons identification in earthwork operations

This component aims to create a method that can identify the idling reasons of earthmoving equipment such as excavators and trucks. First, the types of equipment and their positions in video frames are retrieved from the detection and tracking results. Then, the activities of the excavators are derived from the activity recognition results. In addition, the camera is calibrated to calculate the 3D global positions of the excavators and the trucks. By analyzing the relative positions of the excavators and the trucks in consecutive frames, the workgroups are identified. At last, the idling

reasons can be identified by analyzing the activities of the excavators and trucks in the same workgroup. The main tasks in this component include the following:

(1) Identification of excavators and trucks locations and activities based on the detection and tracking results.

(2) Calibrating the camera and calculating the real distances between excavators and trucks.

(3) Comparing the distances between each excavator and the trucks and identifying their work groups.

(4) Identifying the reasons of idling by analyzing the interactive operation states between the excavators and trucks based on their activities and positions.

(5) Testing the proposed method in the videos recorded from real construction sites, and evaluating its performance.

The details of this component are given in Chapter 5.

3.3 Summary

This chapter presents an overview of the proposed research methodology. The proposed methodology includes three main components. Specifically, the first component is to propose a roadmap based on the literature review to show the technology path and future research directions of full automation in monitoring construction equipment. The second component is to monitor the operation process and analyze the productivity of the excavators in the video frames. The third component is to identify the idling reasons of excavators and trucks and improve the earthmoving productivity.

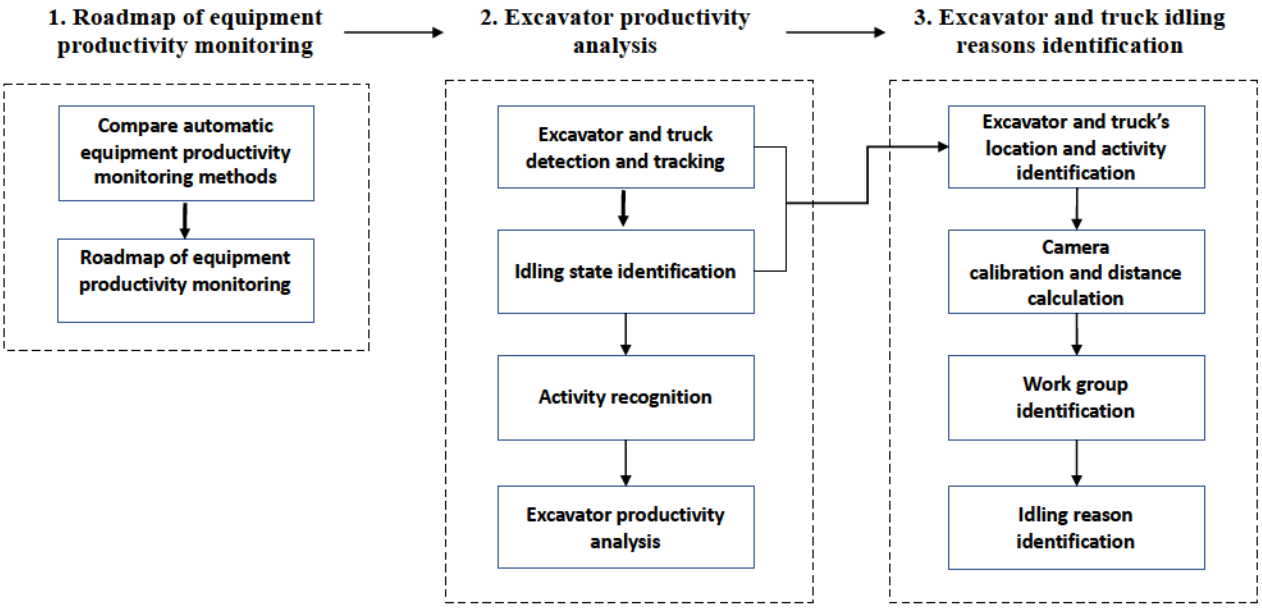


Figure 3-1 Overview of the proposed framework

CHAPTER 4: VISION-BASED ACTIVITY RECOGNITION AND PRODUCTIVITY ANALYSIS

4.1 Introduction

As explained in Section 2.4.3, existing CV-based methods still have limitations for automatically recognizing the activities of construction equipment. For instance, they cannot classify the activities into detailed categories (e.g. digging, hauling, dumping, etc.) in long video sequences. In addition, most of existing methods are not able to recognize and analyze the activities when multiple pieces of equipment are captured simultaneously. Moreover, existing methods just focus on activity recognition. Their practical values are not well discussed. As a result, it is not clear how these methods can be used for productivity control in construction projects.

The main objective of this chapter is to propose a framework to automatically recognize activities and analyze the productivity of multiple excavators. Excavators are selected because they are the most common earthmoving equipment, and they have various types of activities during their working process. In this framework, an excavator detector and a tracker are first integrated to identify and locate the excavators in video frames. Based on the changes of the centroid coordinates and areas of the bounding boxes of the excavators in consecutive frames, idling states are identified and the corresponding frames are cut from the video. Then, 3D CNN is used to precisely recognize the excavators' activities and classify them into detailed types (e.g. digging, loading and swinging). Finally, an algorithm is developed to analyze the activity recognition results and calculate the productivity of the excavators. The feasibility of the proposed framework has been tested and validated on the construction surveillance videos collected from real construction sites.

4.2 Framework of Excavator Activity Recognition and Productivity Analysis

The framework for multiple excavators' activity recognition and productivity analysis is shown in Figure 4-1. The framework contains five main modules: excavator detection, excavator tracking, idling state identification, activity recognition, and productivity analysis. First, a detector is used to identify all the excavators in video frames. The detection results provide two kinds of data, i.e. equipment type and region. Second, each individual excavator and its trajectory are identified

through the detection-based tracking. The tracking results return the identification (ID) number and bounding boxes for each excavator. By analyzing the changes of centroid coordinates and areas of bounding boxes of the excavators in consecutive frames, idling states are identified. Then, the activities of the tracked excavators are recognized with a spatio-temporal feature-based 3D CNN model. Based on the activity recognition and idling state identification results, each video frame is labeled accordingly. Finally, the productivity of each excavator is calculated by compiling the activity recognition results. The details about each module in the framework are provided in the following sub-sections.

4.2.1 Excavator detection

The purpose of this module is to extract the excavators in video sequences. An accurate detection method is used here to detect multiple excavators in long construction site surveillance videos. Construction sites have complex working environments, which include various movements of workers and equipment. In order to isolate each excavator, the excavator detector is used to precisely get the regions of the excavators in video frames for the activity recognition use.

The Faster R-CNN (Girshick 2015) model was applied to detect the excavators in video frames. Faster R-CNN is selected as it has a high accuracy and low processing time. Additionally, experimental results of numerous research works have verified the high performance of Faster R-CNN in various construction objects detection (Luo et al. 2018a; Luo et al. 2018c; Fang et al. 2018a). The Faster R-CNN method includes three main steps. First, a CNN is applied to extract features of the whole image and produce a feature map. Then, a region proposal network uses the feature map to select 300 regions that are distinguished from background regions, and may include target equipment. Finally, the proposed regions are converted to 7*7*512 dimension feature vectors and passed to the fully connected layer to classify whether the region is an excavator or not. When the video is taken as the input of the Faster R-CNN model, it will return the type of the equipment and rectangular bounding boxes, which indicate the regions of the objects in video frames. The coordinates of bounding boxes from the detection results will be used as the input of the next tracking module.

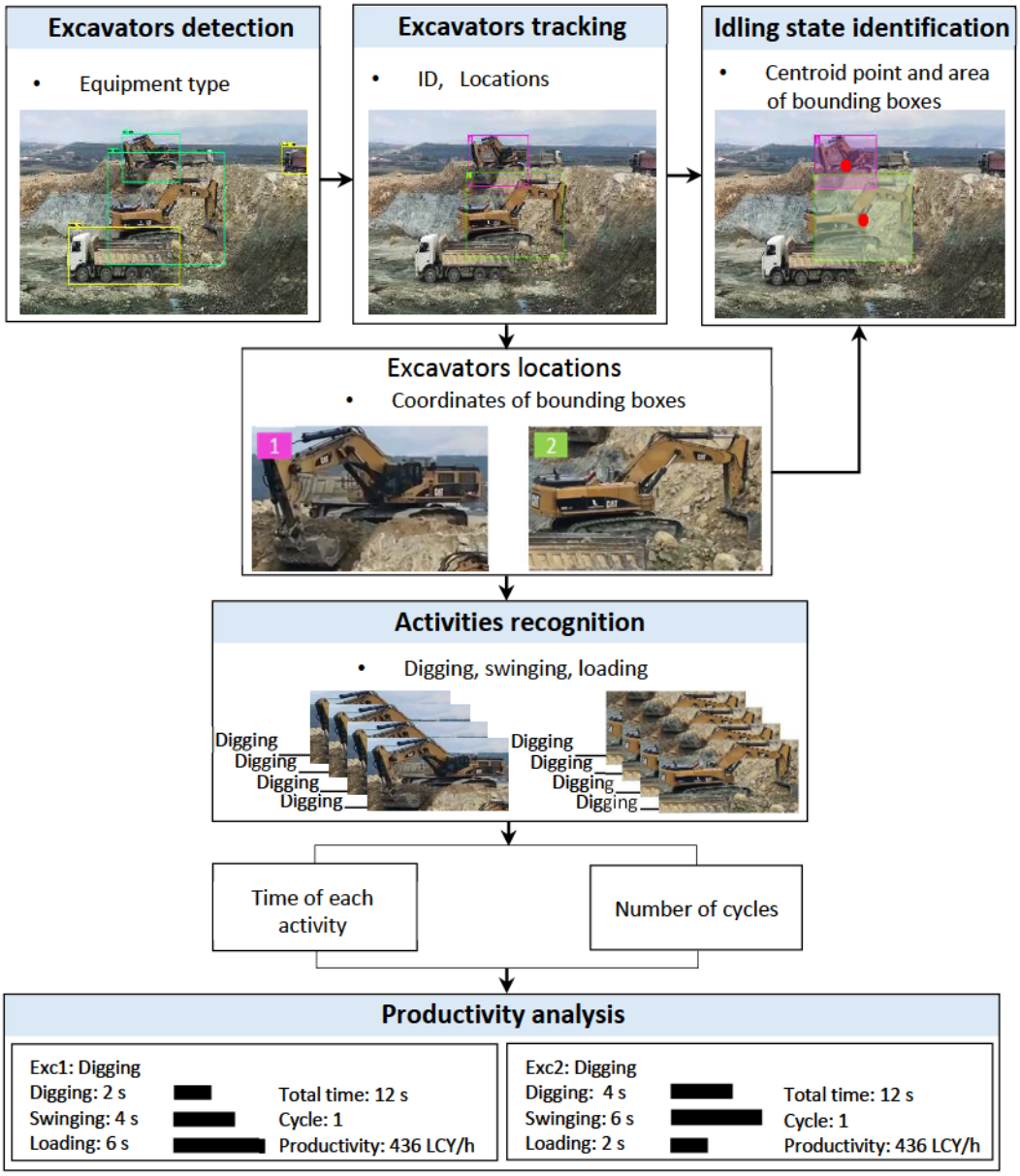


Figure 4-1Workflow of the proposed framework

4.2.2 Excavator tracking

In this module, the most state-of-the-art multi-object deep SORT tracker is applied to associate the same excavators detected in the previous step across all the frames in the video (Wojke et al. 2017). The deep SORT tracker is selected since it achieved good performance on MOT 16 dataset, and it could track the objects through long periods of occlusions in the video. The deep SORT tracker uses both appearance and trajectory information provided by the detection results to track the excavators in video frames. In the trajectory aspect, the tracker uses a constant velocity model

(Bewley et al. 2016) to predict the bounding box of the target excavator in the current frame based on the previous ones. In the appearance aspect, a ten-layer CNN model is used to extract the appearance features of the region inside the bounding box, and represent it with 128-dimensional vectors. Finally, the detected excavator in the current frame is related to the predicted bounding boxes with the Hungarian algorithm (Kuhn 1955). As a result, the same excavator is detected and matched in different frames. The region for the detected and matched excavator is further extracted and resized to 112 × 112 pixel and input to the activity recognition module. The tracking results could be considered as several activity tubes, as shown in Figure 4-2. Each tube represents the exact regions of a certain excavator across the video frames, which is the basis of the next step.

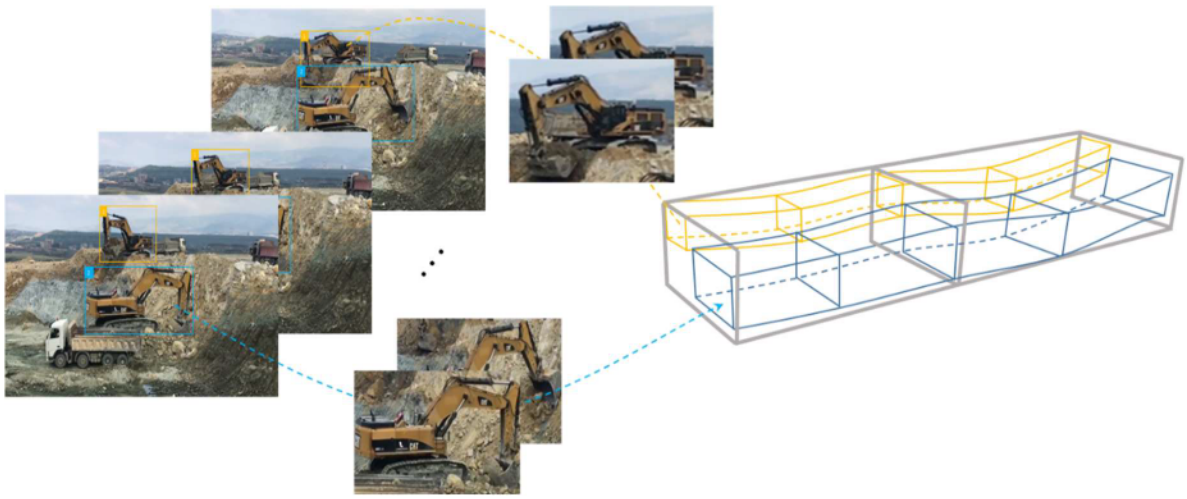


Figure 4-2 Tracking results schematic plot

4.2.3 Idling state identification

The purpose of this module is to identify the idling state of the excavators. The workflow of idling state identification is shown in Figure 4-3. First, the pixel coordinates of all detection/tracking bounding boxes in *K* frames of the video are extracted. The centroids (x_i, y_i) and areas (a_i) of the bounding boxes are calculated. Also, the average area $\bar{a}$ for all the bounding boxes is estimated. For each video frame, the centroid distance changes (Δd_i) and area changes (Δa_i) between the current frame and the *nth* frame after the current one are calculated using Equations 4-1 and 4-2;

$$Centroid\ coordinates\ change\ (\Delta d_i) = \sqrt[2]{(y_{i+n} - y_i)^2 + (x_{i+n} - x_i)^2} \quad \text{Equation 4-1}$$

$$Area\ change\ (\Delta a_i) = |a_{i+n} - a_i| \quad \text{Equation 4-2}$$

where (x_i, y_i) and a_i are the centroid coordinates and area of the bounding box in frame i. Moreover, a sliding window with L consecutive frames and the step of one frame is applied to calculate the standard deviation of distance changes $STD(\Delta d_i)$ and the standard deviation of area changes $STD(\Delta a_i)$. Figure 4-4(a) and Figure 4-4(b) show an example of changes of distance (Δd) and the concept of sliding window for calculating $STD(\Delta d)$, respectively. If $STD(\Delta d_i)$ is less than d', and $STD(\Delta a_i)$ is less than $\alpha\bar{a}$, the excavator's state in that particular frame will be identified as idling. The threshold d' represents a negligible change of the distance between the centroids, and α is a percentage of the average bounding box area so that $\alpha\bar{a}$ represents a negligible change of the area. Both d' and α are determined based on the resolution of the video and the distance of the excavator from the camera.

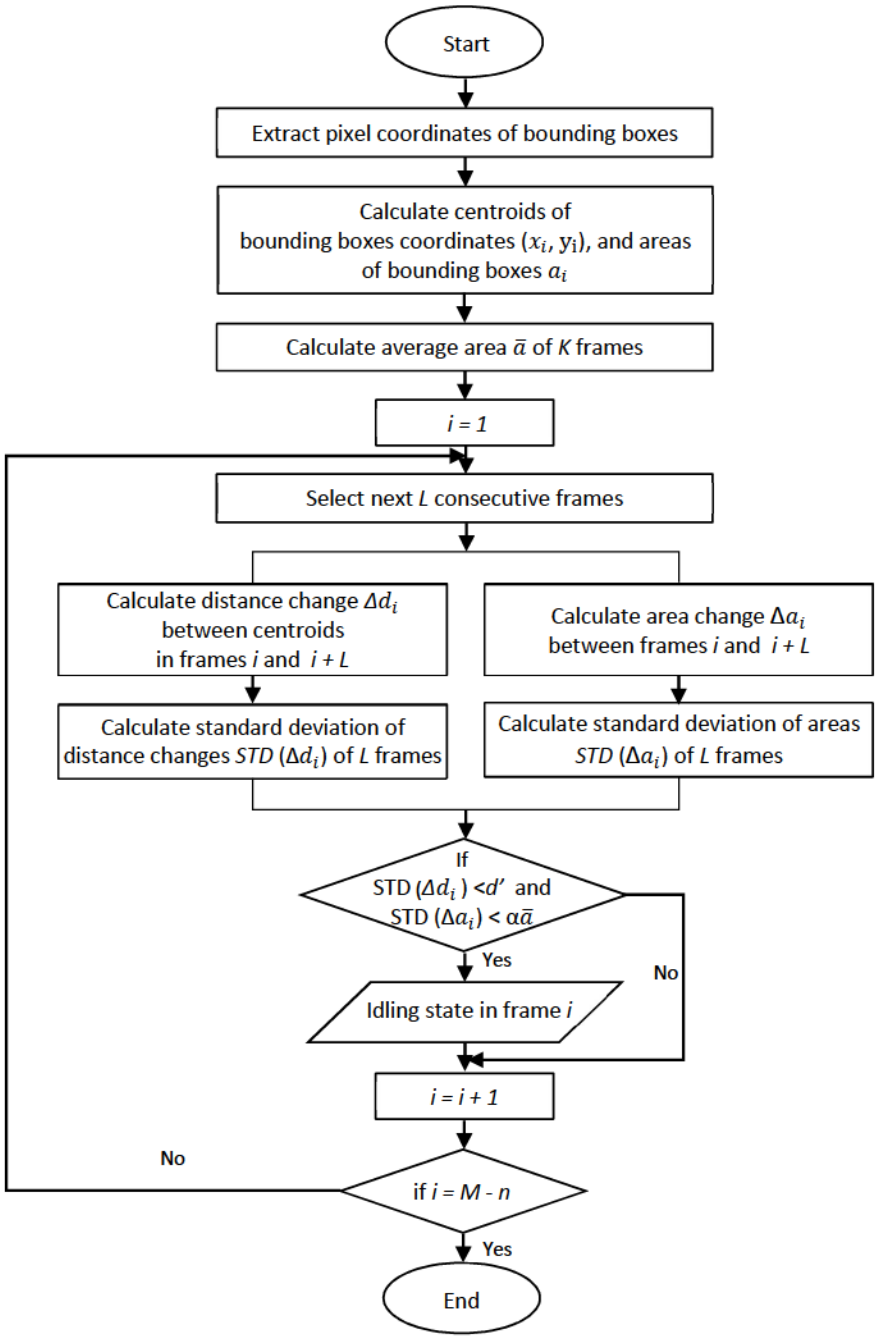
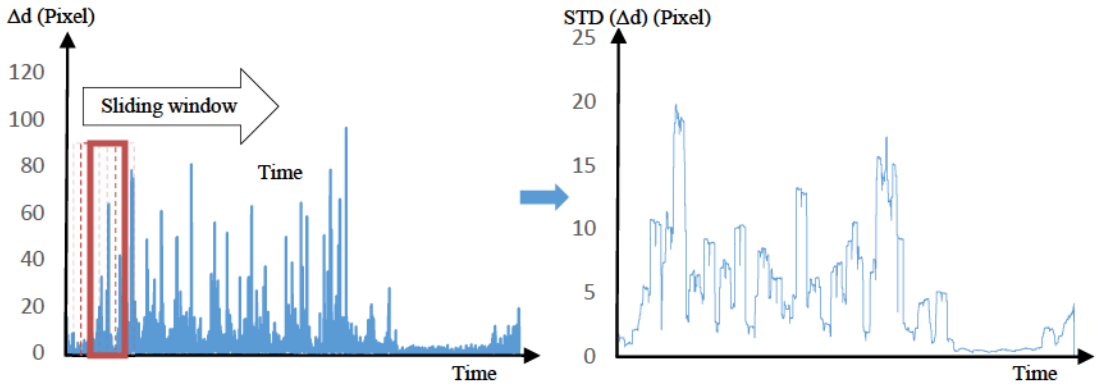


Figure 4-3 Method of idling state identification



(a) *Δd* (b) *STD (Δd)*

Figure 4-4 Examples of changes of distance (*Δd*) between centroids of bounding boxes and the concept of sliding window for calculating *STD (Δd)*

4.2.4 Activity recognition

In the activity recognition task, the latest 3D ResNet (Hara et al. 2018) is used to recognize the excavator's activities (e.g. digging, loading, and swinging) in video sequences. The 3D ResNet is selected since it achieved the best performance of activity recognition task on UCF-101dataset (Soomro et al. 2012) with 86.5% accuracy, and it could directly get the activity recognition results from RGB input video frames. The model has the same residual block architecture as the ResNet (He et al. 2017) but performs the convolution and polling with 3D kernels. It preserves both the spatial and the temporal information of the activities in the video. All the inputs, kernels, and outputs in the network are 3D tensors with the dimensions of temporal × height × width (L × H × W) (Tran et al. 2017). Specifically, the network takes 16 × 112 × 112 video frames as input. The sizes of the convolutional kernels are 3 × 3 × 3, and the temporal strides are 1 for the first convolutional layer and 2 for the other layers. First, the tracking results are taken as the input of the neural network. Accordingly, convolutional kernels are applied on every 16 frames to extract the spatiotemporal information of the activity. The selection of 16 frames is mainly due to the GPU memory limit, although the selection of a longer duration may improve the recognition accuracy of the model (Varol et al. 2016; Hara et al. 2017). At last, each frame is labeled to indicate the excavator activity after correcting the recognition errors with majority voting, as shown in Figure 4-5. This is because excavators usually work continuously and cannot change their activity states in a short period of time. It was observed that each activity lasts at least 2 seconds during the operation.

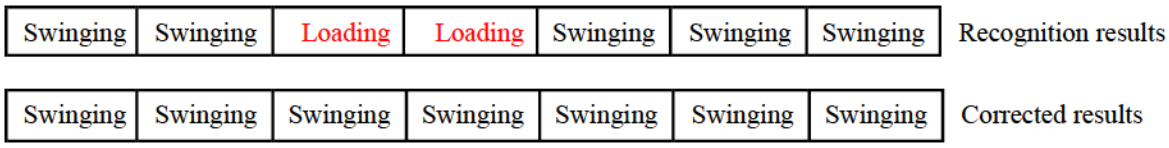


Figure 4-5 Correction of abnormal activity recognition

The training of the activity recognition model started from fine-tuning the 3D ResNet work developed by Kay et al. (2017). The fine-tuning is to efficiently initialize the parameters of the model and avoid overfitting. In the training process, the data augmentation technique is used to extend the size of the dataset by flipping the video frames, shifting the image channels, and shearing the frame size. In addition, all video clips are fixed at 25 FPS for training, the batch size of the model was set to 16, and the learning rate was set to 0.001.

4.2.5 Productivity calculation

In this module, a method is created to automatically calculate the productivity of the excavator based on the activity recognition results. In the earthmoving work, excavators usually work with other equipment, such as trucks and scrapers. For example, an excavator digs the soil and loads it onto the bed of a truck. When fully loaded, the truck moves the soil to the dumping area and returns to be loaded again. If the excavator is considered as an independent equipment, its working process could be broken down into several activities such as soil digging, swinging and loading the truck. Therefore, a single earthmoving cycle of an excavator could be broken down into four types of activities: digging, swinging after digging, loading into the dump truck, and swinging for digging. During the work time, these four types of activities are repeated sequentially, which is defined as a production cycle, as shown in Figure 4-6.

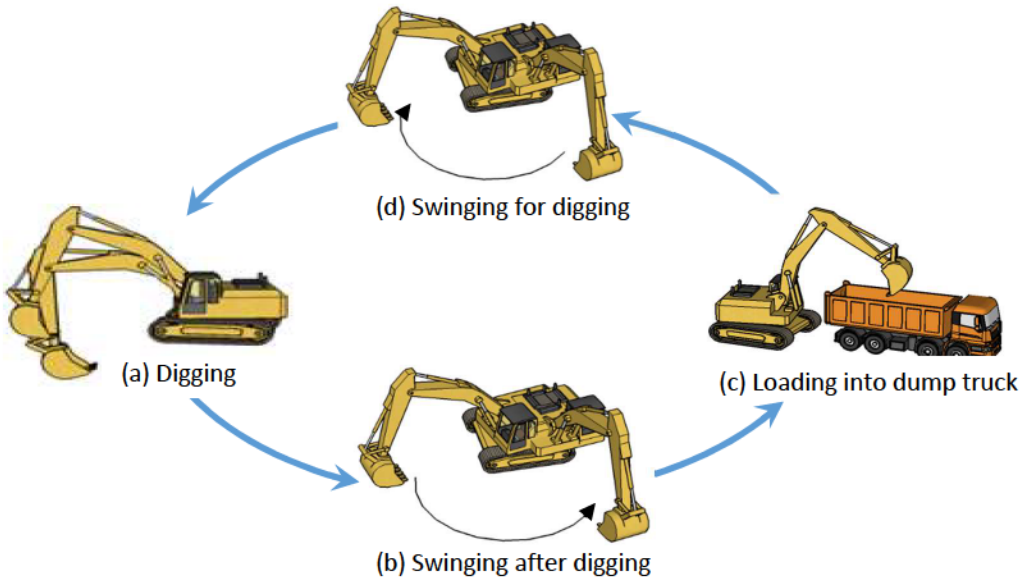


Figure 4-6 Excavator working cycle

In earthmoving, the excavator's productivity can be calculated with the cycle time and the bucket payload, as shown in Equation 4-3. Since the bucket payload is given by the manufacturer, the target of the productivity calculation becomes to determine the cycle time of the excavator. To simplify the procedure, the two types of swinging (swinging after digging and swinging for digging) are not distinguished in this research. This way, one excavator working cycle is broken down into digging, swinging, and loading.

$$Productivity\ (LCY/hr) = \frac{Cycles}{hr} \times \frac{Average\ bucket\ payload\ (LCY)}{Cycle} \qquad \text{Equation 4-3}$$

The time for each cycle is measured following the workflow in Figure 4-7. After the activity recognition, each video frame is labeled to indicate the activity of the excavator in the frame. The labels are shown as L_i in Figure 4-7. Then, the labels of two consecutive frames are compared. If they are the same, it means that the activity continues. Therefore, the time for this activity is increased by 1/FPS (frame per second). If the labels are different, it means that a new activity has started. The time of the newly recognized activity will increase by 1/FPS. The total time of one cycle is the difference between the start times of two adjacent digging activities.

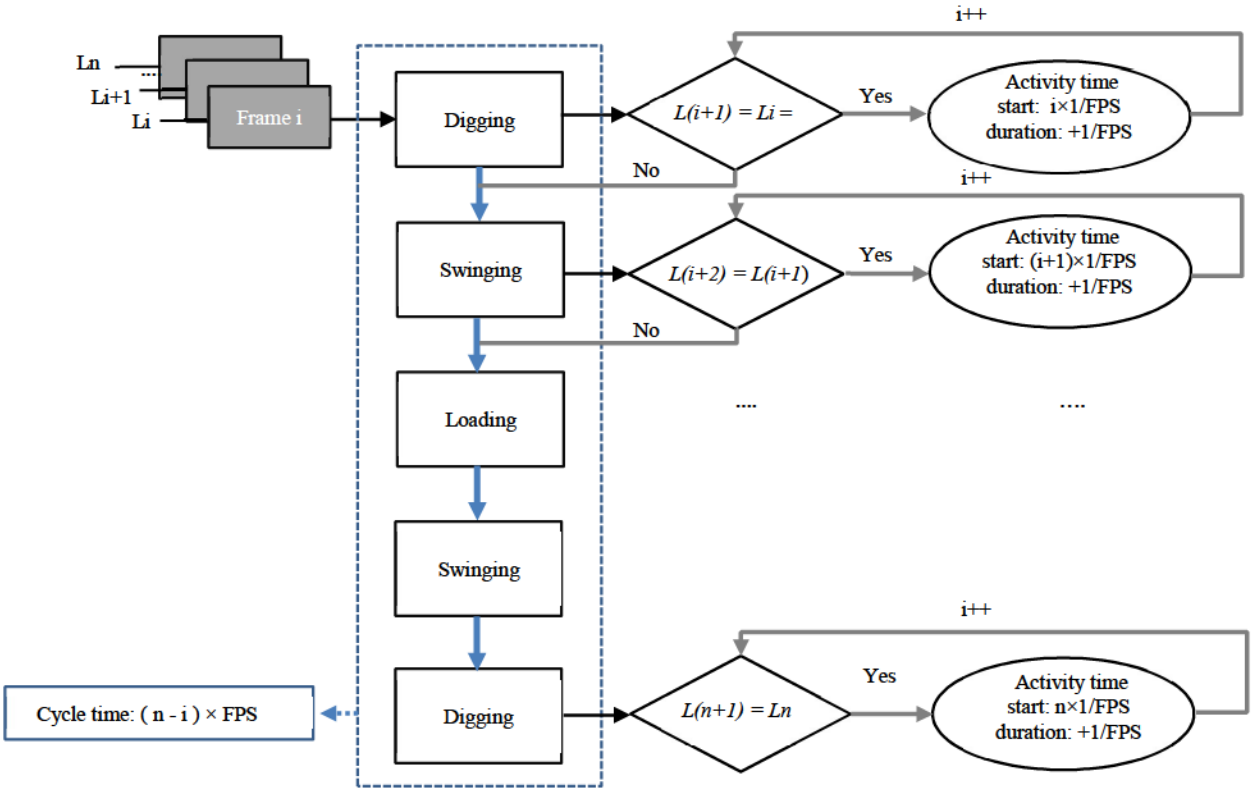


Figure 4-7 Workflow of cycle time calculation

4.3 Implementation and Results

In this section, the implementation of the proposed framework is introduced. The activity recognition model is trained in the Python 3.6 environment, 64-bit Ubuntu 16.04 system with the hardware configuration of an NVIDIA GeForce GTX 1080GPU and 32 gigabytes. The proposed framework is tested in the Python 3.6 environment, 64-bit Windows 10 system with the hardware configuration of an NVIDIA GeForce GTX 1070GPU and 32 gigabytes. It could process 15 frames per second with this configuration.

4.3.1 Training and testing

First, a dataset of videos containing excavators is manually collected. Each video contains one of the three types of activities: digging, loading and swinging. In order to avoid bias, 351 video clips were collected from 21 different construction sites, considering site conditions, equipment viewpoints, and scales and colors of excavators. The detailed information of the dataset is shown in Table 4-1. First, each video sample in the dataset is annotated with the activities of this video. Figure 4-8 illustrates the example images extracted from the videos in the dataset.

Table 4-1 Statistic information of the dataset

Activity type	Number of videos	Total time (s)	Average of video length (s)	Number of excavators
Digging	122	651	5.3	19
Swinging	119	440	3.7	21
Loading	110	490	4.5	19

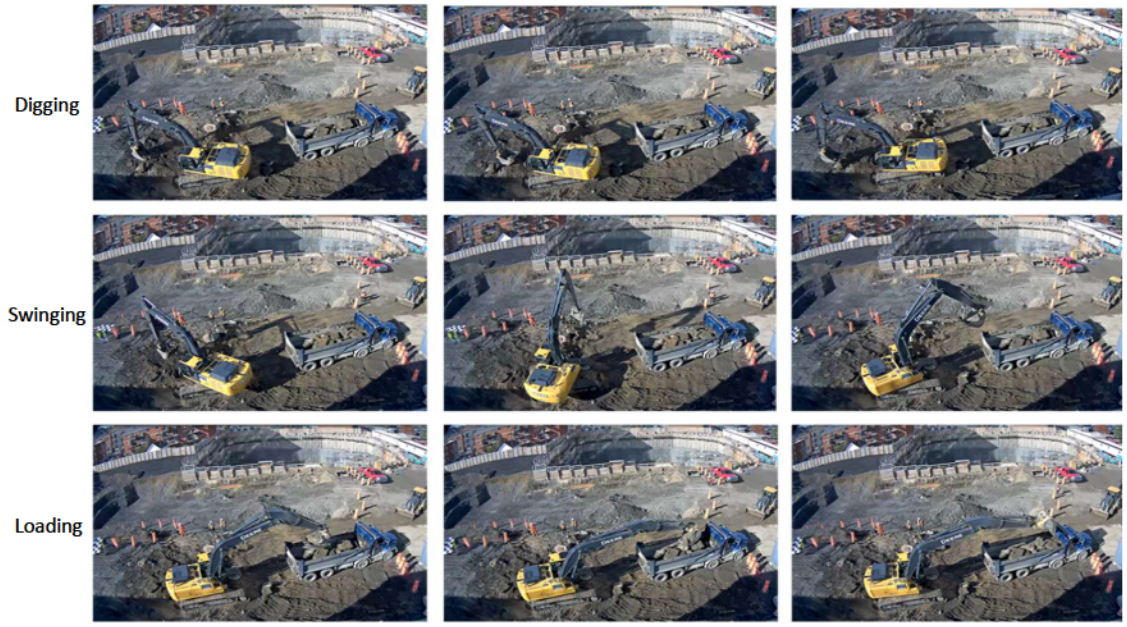


Figure 4-8 Example images extracted from the dataset

Then, 30 video clips with the resolution of 1280 × 720 pixels and the duration of 264 seconds were used to test the performance of the activity recognition model. Each video clip in the test dataset contains at least one type of excavator's activity. Two indicators, precision and recall, were applied to measure the performance of the activity recognition model as shown in Equations 4-4 and 4-5. These two indicators are widely used to validate the performance of object classification and detection methods in the computer vision domain.

$$Precision = \frac{TP}{TP+F} \qquad \text{Equation 4-4}$$

$$Recall = \frac{TP}{TP+F} \qquad \text{Equation 4-5}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. More details about the concepts of precision and recall could be found in the work of Kim et al. (2018d). The comparison of activity recognition results with the ground truth is shown in Figure 4-9, and the performance of the pre-trained model is presented in Table 4-2. The recognition precisions for the three activities are: 95% for digging, 86% for swinging, and 84% for loading. The recall rates are 86%, 93%, and 80%, respectively. The average accuracy of the recognition is 87.6%. The high activity recognition accuracy on test videos, which have various viewpoints, background noises from other movements, and partial occlusions, validated the robustness of the proposed method.

In addition, it is worth noting that the model was also applied on a 60.2 min video to recognize excavator's activities in the implementation stage and achieved the accuracy of 92.5%. The results indicate that the proposed method can effectively recognize the consecutive activities of the excavator in the long video sequences.

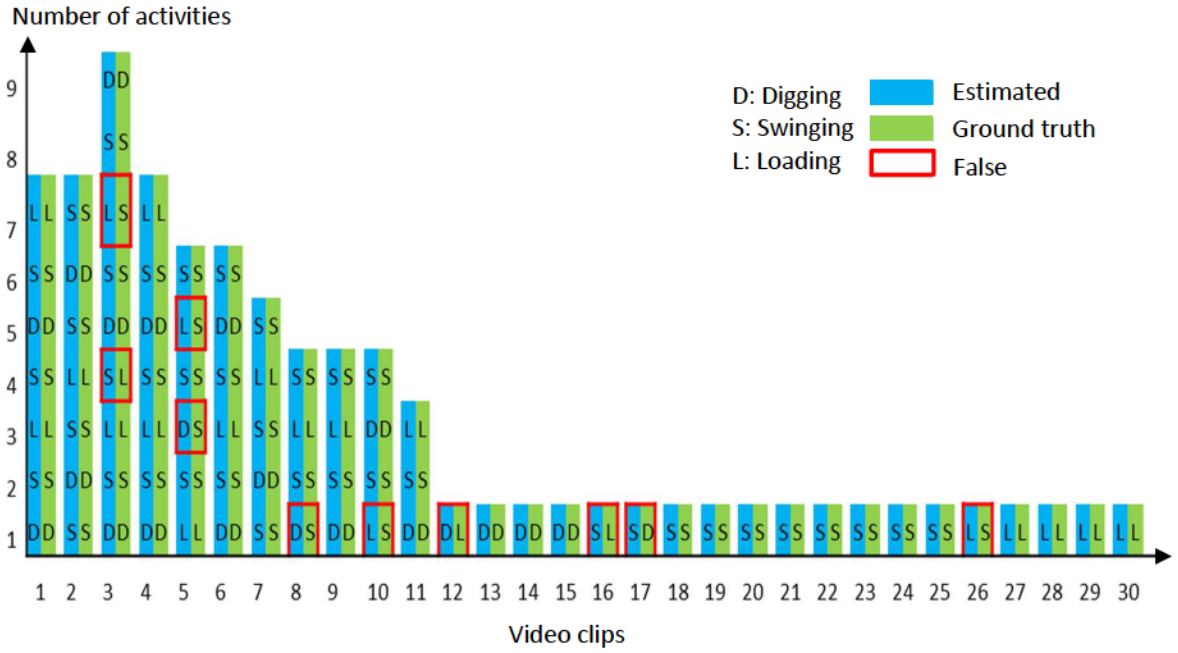


Figure 4-9 Comparison of test results and ground truth values

Table 4-2 Activity recognition results

Activity	Number of activities in the videos		Performance metrics		
	Total	Incorrect recognition	Precision (%)	Recall (%)	Accuracy (%)
Digging	21	3	95	86	85.7
Swinging	40	3	86	93	92.5
Loading	20	4	84	80	80
Average/Total	81	10	88	88	87.6

4.3.2 Implementation and results

In this section, two case studies were conducted to validate the feasibility of the proposed method for multiple excavators' activity recognition.

(1) Case 1: Two excavators without overlapping

In Case 1, the proposed method was used to recognize two excavators' activities in a 1 min video (Figure 4-10). In this video, Excavator 1 has 8 activities within 2 cycles, and Excavator 2 has 10 activities within 2.5 cycles. Figure 4-11 shows examples of the activity recognition. The test results are shown in Table 4-3.

Figure 4-10 Activity recognition under contrast lighting conditions in Case 1

Figure 4-11 Examples of activity recognition in Case 1

Table 4-3 Test results and ground truth values of Case 1

Excavator 1	Cycle 1	**Ground truth**	**D**	**S**	**L**	**S**
		Estimation	D	S/D	L	S
	Cycle 2	**Ground truth**	**D**	**S**	**L**	**S**
		Estimation	D/L	S	D	S
Excavator 2	Cycle 1	**Ground truth**	**D**	**S**	**L**	**S**
		Estimation	D	S	L	S
	Cycle 2	**Ground truth**	**D**	**S**	**L**	**S**
		Estimation	D	S	L	S
	Cycle 3 (partial)	**Ground truth**	**D**	**S**		
		Estimation	D/L	S		

D: Digging; L: Loading; S: Swinging

The results show that there are four wrongly recognized activities: three for Excavator 1 and one for Excavator 2. In fact, the errors are mainly due to the bad light conditions as can be seen in Figure 4-10. The light condition changes drastically in the second half of the video.

(2) Case 2: Two excavators with overlapping

In Case 2, the proposed method was used to recognize two excavators' activities in a 40 s video (Figure 4-12). Excavator 1 has five activities within 1.25 cycle, and the Excavator 2 has six activities within 1.5 cycle. Figure 4-13 shows examples of the activity recognition. The test results are shown in Table 4-4.

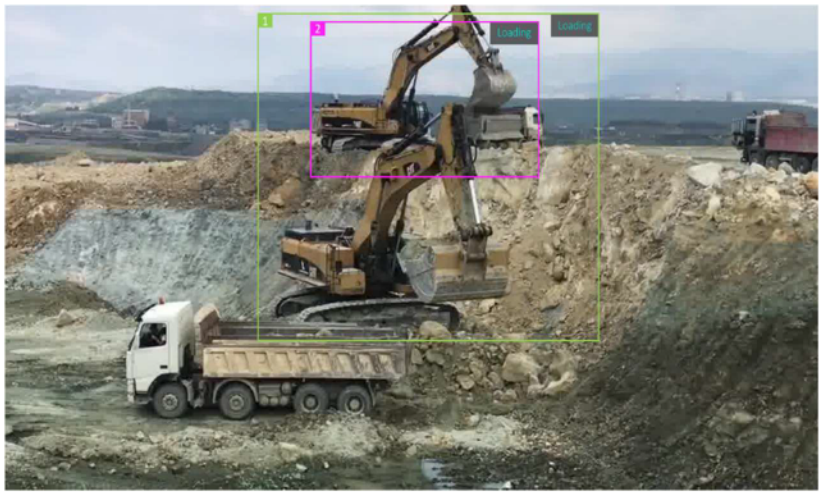


Figure 4-12 Activity recognition with overlapping bounding boxes in Case 2

Figure 4-13 Examples of activity recognition in Case 2

The results indicate that one activity is wrongly recognized for Excavator 1 because of the overlapping of the bounding boxes as shown in Figure 4-12. The deep SORT tracker identified a large bounding box for Excavator 1 fully overlapping with Excavator 2 in 500 frames, which affected the activity recognition results.

Table 4-4 Test results and ground truth values of Case 2

Excavator 1	Cycle 1	**Ground truth**	**D**	**S**	**L**	**S**
		Estimation	D	S	L	L
	Cycle 2 (partial)	**Ground truth**	**D**			
		Estimation	D			
Excavator 2	Cycle 1	**Ground truth**	**D**	**S**	**L**	**S**
		Estimation	D	S	L	S
	Cycle 2 (partial)	**Ground truth**	**D**	**S**		
		Estimation	D	S		

D: Digging; L: Loading; S: Swinging

4.4 Productivity Estimation

Additionally, the proposed framework was also tested to estimate the productivity of excavators using a long video sequence, which contains 60.2 min of excavator's operation. Based on the method of idling state identification explained in Section 4.2.3, the width of the sliding window L is selected as 100 frames (equivalent to 4 s with 25 FPS video). The threshold d' is selected as 1 pixel, and α is selected as 1.5% considering possible small changes in the bounding box even when the excavator is in idling state. The test outputs are shown in Figure 4-14. The activity information is added on bottom of the image including: (1) the activity identified in the current frame; and (2) the total time of each activity compiled till the current frame. In addition, the last frame provides the productivity information, which include: (1) the total time of the video; (2) the number of working cycles; and (3) the productivity of the excavator with the unit of LCY/h.

Figure 4-14 Examples of productivity calculation results

In the video, the Deere 290G excavator (bucket payload of 1.4 LCY) completed 94 work cycles in 60.2 min with the idling time of 9.7 min and the actual productivity of 131.16 LCY/h. The estimated total idling state duration is 9.1 min, and the estimated productivity is 153.48 LCY/h with 110 cycles. The accuracy of productivity calculation is 83%. The test results showed the feasibility of using the developed framework to analyze real videos of construction projects and to monitor the operation of excavators.

4.5 Summary and Conclusions

This chapter presented a novel fully automatic vision-based framework for multiple excavators' operation monitoring and productivity calculation. The framework integrates the detection, tracking, activity recognition and productivity calculation modules. The detection module identifies all the excavators in video frames. The tracking module correlates the same excavators and extracts their regions across the frames in the video sequences. After detecting the idling state, the activities of each excavator are recognized based on the tracking results. Finally, the productivity is automatically calculated based on the activity information of the excavators. The

effectiveness of the proposed framework has been tested on construction surveillance videos, and the results have shown the feasibility of the proposed framework.

The contributions of this work mainly consist of three perspectives. First, the advanced technology in computer vision is integrated to recognize continuous activities of excavators in long site surveillance videos. Specifically, compared with the recent work of Kim et al. (2018b; 2018d) which could just identify working and idling states of the excavator based on its relative location with respect to the truck, the proposed method can recognize detailed activities such as digging, loading and swinging from activities spatial and temporal features. In addition, the proposed method is also superior to another related work of Kim and Chi (2019) in long videos analysis, since they did not consider the temporal features of the activities. In Kim and Chi's method, activities are recognized based on their sequence relations. The method in this chapter can recognize activities based on both spatial and temporal features that are directly learned from videos regardless of their sequential relations. In the model training stage, the proposed method relied on several training strategies, such as fine-tuning, data augmentation and batch size optimization, to train the detection, tracking and activity recognition models with a limited dataset. The results of the activity recognition on different excavator sizes and types, various viewpoints and lighting conditions, and background movements have proved its robustness in construction environments and provided a solid basis to automate the calculation of the equipment productivity.

Second, the detection, tracking and recognition techniques are integrated to measure multiple excavators' operations. In a recent construction activity recognition research, Luo et al. (2018b) manually specified bounding boxes to start the tracking process, which is not efficient. In this chapter, a fully automated framework is conducted for excavators' activity recognition without the manual input.

Third, a sliding window method is applied to identify the idling state of excavators by comparing the changes of centroid distance and areas of bounding boxes of excavators in consecutive video frames.

Forth, an automatic method is developed to directly analyze the productivity based on the activity recognition results. The experimental results on the video with the duration of nearly one hour supported the feasibility and applicability of the proposed method in practice. Moreover, the detailed operation information provided by this method, such as the time of each activity and the

number of work cycles can be further used to help construction managers identify the causes of excavator's low productivity from the video.

CHAPTER 5: AUTOMATIC IDENTIFICATION OF IDLING REASONS BASED ON EXCAVATOR-TRUCK RELATIONSHIPS

5.1 Introduction

As cameras are recently installed to monitor construction sites, an increasing number of research studies have been focused on monitoring equipment work productivity by automatically analyzing surveillance videos with CV-based methods. The recent research work has been focused on monitoring equipment operations by recognizing the activities or estimating equipment's productivity by identifying the durations of activities, such as working, moving and idling. However, there is no CV-based method to investigate the reasons that cause low productivity automatically. There are different reasons that cause excavators idling. One important reason is that the jobsite is not properly planned to allow enough maneuvering space for trucks; thus, the excavator has to wait for the trucks, which should spend long time maneuvering near the excavator. Therefore, by analyzing the interactive work states of excavators and trucks, the idling reasons can be identified. Accordingly, based on the specific reason, measurements can be taken to reduce the idling time, and consequently to improve the productivity of the excavator.

This chapter aims to develop a CV-based method for identifying idling reasons based on the interaction analysis between excavators and trucks from construction surveillance videos. First, the activities of the excavators and trucks are identified using CNN. Then, work groups of excavators and trucks are clustered. Finally, the relationships between each excavator and the surrounding truck(s) are analyzed to identify potential reasons that cause the idling.

5.2 Methodology of Idling Reasons Identification

The methodology for idling reasons identification is shown in Figure 5-1, which contains three main steps. First, the excavators and trucks are detected and tracked to get their locations and activities in video frames. Second, the excavators and trucks are clustered to analyze their interactive work states. Third, the potential reasons of idling are identified and justified. Specifically, the idling reasons of the excavators and trucks are classified into five different cases based on the number, activities and locations of trucks, as well as the interactive work states of the

excavators and trucks, which are calculated in the previous steps. The details of these three steps are introduced in the following sub-sections. The methodology of this chapter is based on the assumption that the equipment does not have mechanical problems and all the operators have no health issues that may cause idling. These potential reasons of idling are beyond the scope of this chapter.

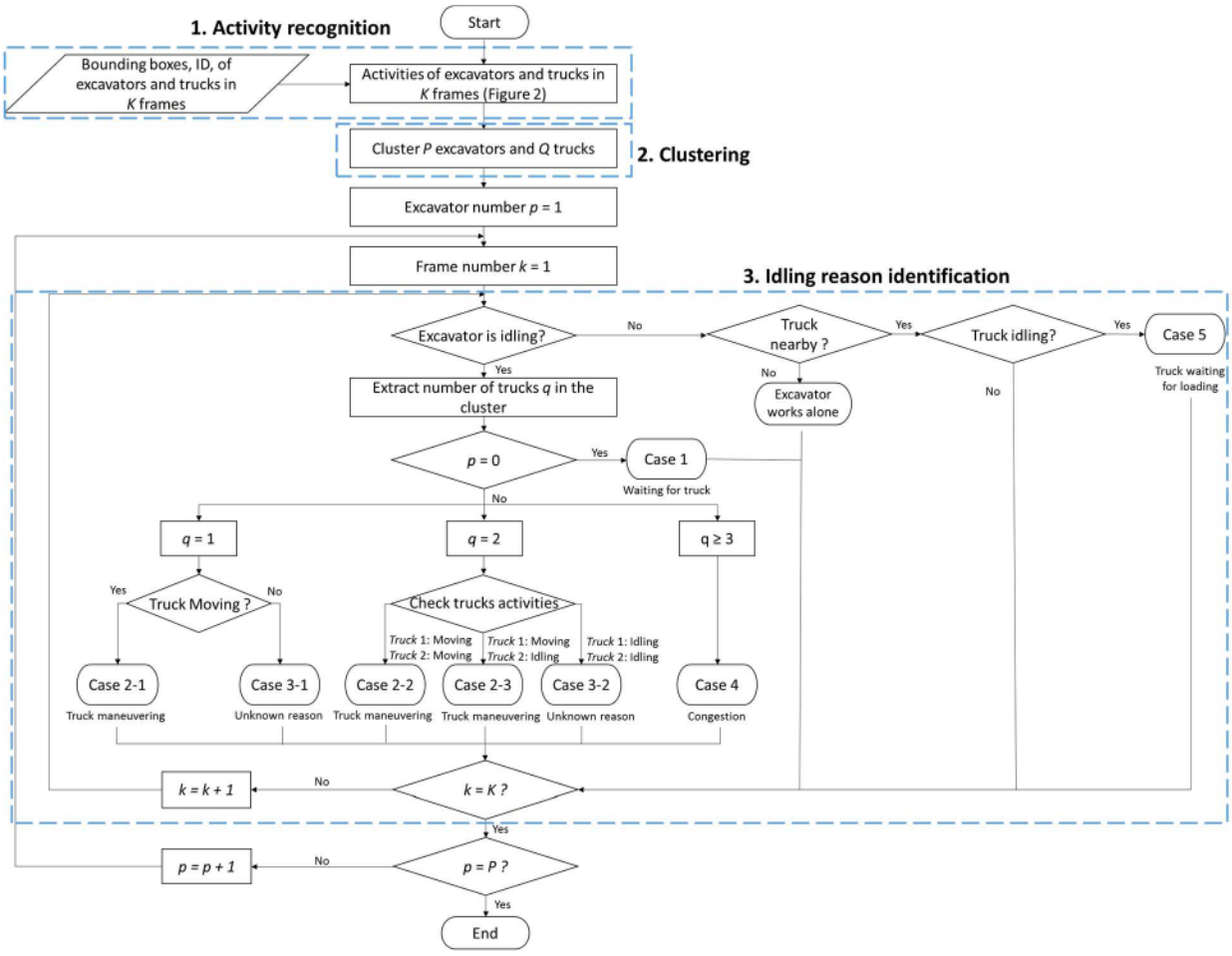


Figure 5-1 Workflow of the working group identification

5.2.1 Identification of excavators and trucks locations and activities

In the first step, detection and tracking methods are used to extract equipment's types and coordinates of bounding boxes in *K* video frames. You Only Look Once-v3 (YOLO-v3) (Redmon and Farhadi 2018) detector and multi-object deep SORT tracker (Nicolai et al. 2016) are applied in this study for equipment detection and tracking, respectively. The YOLO-v3 and deep SORT are selected for their performance of high accuracy and speed in both computer vision and applications in the construction domain (Chen et al. 2020a; Luo et al. 2020).

Following the detection and tracking, the working and idling states of excavators and trucks are recognized using the method proposed in Section 4.2.3. Excavators work at a fix position, but they have obvious shape changes. On the contrary, trucks move between loading and dumping locations, but have less changes in their shape when they are not moving. Thus, the activities of trucks are identified based on Δd only without considering the changes of the bounding box area. The thresholds are selected according to the different characteristics of excavators and trucks working process. Finally, the activities of equipment are recognized as working or idling.

5.2.2 Excavator and truck clustering

The second step is to cluster excavators and trucks into different groups. In the real earthwork operations, excavators usually work with nearby trucks. Therefore, excavators and trucks are clustered based on their distances in video frames. First, the number of excavators *P* and trucks *Q* are obtained from detection results. Second, the distances between excavators and trucks in frame *k* (d_k) are calculated. In order to calculate the distance, the camera is calibrated to get the intrinsic and extrinsic parameters to transfer the pixel coordinates of equipment to the global coordinates using Equation 5-1 (Heikkila and Silven, 1997). Where *M* is the intrinsic parameter of the camera; *R* is the rotation matrix and *t* is the translation vector of the camera, which are extrinsic parameters; *u* and *v* are the pixel coordinates, *s* is scale factor, which is calculated from *M*, *R* and pixel coordinates as in Equation 5-2. After getting the global centroid coordinates, d_k is calculated using Equation 5-3 (Heikkila and Silven, 1997).

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = R^{-1}(M^{-1}s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} - t) \qquad \text{Equation 5-1}$$

$$R^{-1}M^{-1}\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}a_1\\a_2\\a_3\end{bmatrix} \qquad R^{-1}t=\begin{bmatrix}b_1\\b_2\\b_3\end{bmatrix} \qquad s=\frac{b_3}{a_3} \qquad \text{Equation 5-2}$$

$$d_k=\sqrt{(y_k^e-y_k^t)^2+(x_k^e-x_k^t)^2+(z_k^e-z_k^t)^2} \qquad \text{Equation 5-3}$$

where (x_k^e, y_k^e, z_k^e), (x_k^t, y_k^t, z_k^t) are the global centroid coordinates of an excavator and a truck in frame k, respectively. Accordingly, each truck is grouped with the nearest excavator. A threshold γ is set for clustering. If the distance between the truck and the excavator is larger than γ, the truck will not be included in the group. γ is selected based on the size of the excavator and truck.

5.2.3 Idling reasons identification

The third step is to identify the potential reasons of why the excavators and trucks are idling. These reasons are summarized into four cases for excavator idling and one case for truck idling, as shown in Table 5-1 and Figure 5-2. All these cases are summarized based on the literature (Zou and Kim 2007; Peurifoy et al. 2009; Ibrahim and Moselhi 2014; Song et al. 2017; Manyele 2017; Kim et al. 2018a), and observing site surveillance videos. In Case 1, there is no truck working with the excavator, and the excavator is waiting. In Case 2, the excavator is waiting for the truck, which is maneuvering or moving to the right loading position. In Case 3, both of the excavator and truck are idling. In this case, the reason of idling cannot be easily observed from videos. For instance, the operator of the excavator may be communicating with other workers or workers may be passing nearby. For Case 2 and Case 3, there are subcases depending on the number of trucks. In Case 4, there are too many trucks (more than two) near the excavator, which may cause site congestion. In Case 5, excavator is loading a truck, meanwhile, one or more trucks are idling nearby to reduce the waiting time of the excavator to load to the next truck. It should be noticed that an excavator working alone without a truck is considered in a partially productive working state. For each idling excavator, the number of trucks q in the same group is calculated. If there is no truck in the group, the reason of the idling is classified as Case 1, which indicates that the excavator is waiting for a truck. When there is only one truck in the group, the idling reason is determined by the activity of the truck. If the truck is moving, the reason is classified as Case 2, which indicates that the excavator is waiting for the truck maneuvering to the loading position. Otherwise, there could be different reasons that cause excavator's idling as explained above. Therefore, in this condition, the reason of excavator's idling is classified as Case 3 (unknown reason). When there are two trucks

identified in the group, the activities of trucks have three conditions: both trucks are moving, one is moving and the other is idling, or both trucks are idling. These different conditions of trucks' activities could lead to two reasons of excavator's idling. If at least one of the two trucks is moving, the reason is classified as Case 2 (i.e. truck maneuvering). If both trucks are idling, the reason of excavator's idling is unknown, which is Case 3. When there are more than three trucks in the group, the idling of the excavator is classified as Case 4, which indicates too many trucks causing site congestion around the excavator. If the excavator is working and two or more trucks are identified idling, this indicates that one truck is waiting aside to reduce the idling time of excavator, this case is classified as Case 5.

Table 5-1 Potential reasons and managerial suggestions of the excavator idling

Case		Potential reason	Managerial actions suggestions
Excavator idling	Case 1	Excavator is waiting and there is no truck	More truck to keep the excavator working at capacity $Balanced\ number\ of\ trucks = \frac{Truck\ cycle\ time\ (\text{min})}{Loading\ time\ per\ truck\ (\text{min})}$
	Case 2	One or more trucks are maneuvering	Schedule planning
	Case 3	Unknown reasons (e.g. operator, mechanical problem, safety issue)	Observing the video for specific suggestion
	Case 4	Congested site with many trucks	a. Equipment path planning and optimization b. Site organization optimization
Truck idling	Case 5	One or more trucks waiting to be loaded	Use less trucks

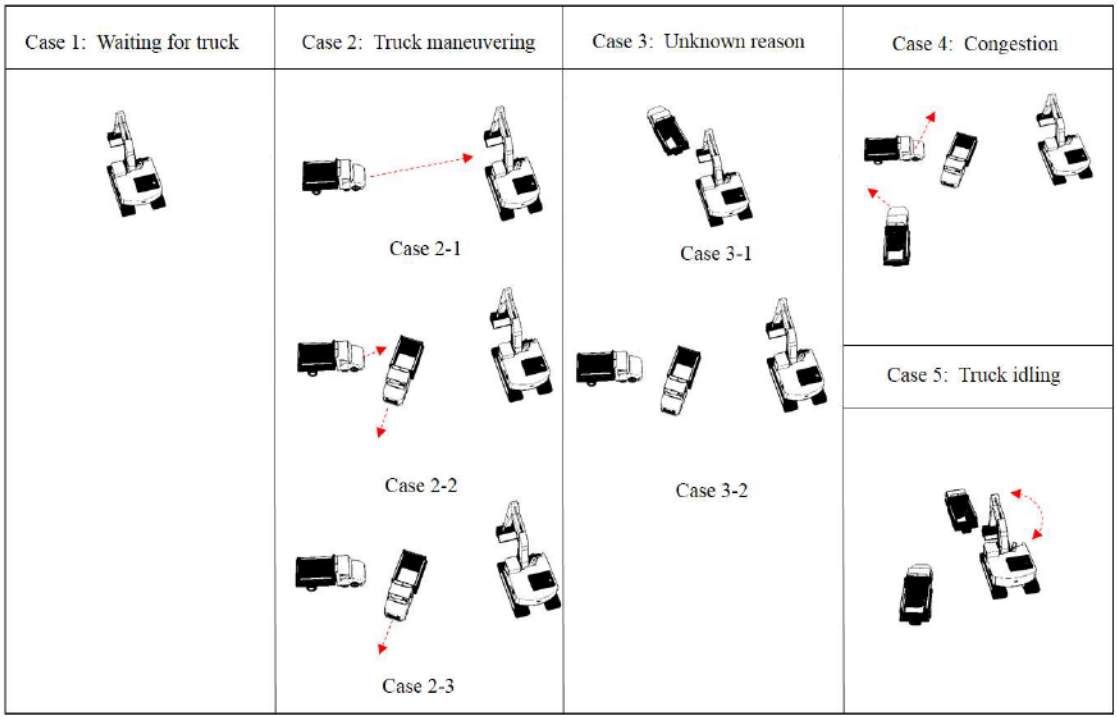


Figure 5-2 Conceptual figures for different cases (The arrows indicate the equipment is moving)

After recognizing the reasons that caused idling, construction managers can take different strategies to reduce the idling time, as shown in Table 5-1. For Case 1, the idling time can be reduced by controlling the number of trucks that work with an excavator. In this case, more trucks should be arranged to cooperate with the excavator to reduce its idling time, since the utilization cost of the excavator is higher than the truck. Therefore, to keep the excavator working at capacity, more trucks are required based on Equation 5-4 (Peurifoy et al. 2009). For Case 2, the optimization of the schedule, site layout and path of the equipment are needed to reduce the maneuvering time of the truck. For Case 3, different managerial actions should be taken based on a specific reason. For example, if there is a mechanical problem, the inspection of the equipment should be arranged more frequently. If workers are passing nearby, the safety education of the workers should be improved. For Case 4, the site organization and path planning should be optimized to reduce site congestion. For Case 5, to reduce the idling time of the truck, fewer trucks should be arranged to work with an excavator based on Equation 5-4 (Peurifoy et al. 2009).

$$Balanced\ number\ of\ trucks = \frac{Truck\ cycle\ time\ (\text{min})}{Loading\ tim\ \ per\ truck\ (\text{min})} \qquad \text{Equation 5-4}$$

5.3 Implementation and Case Studies

In this section, the implementation of the proposed method is introduced, and three case studies are provided to demonstrate the performance of the proposed method. A computer with two NVIDIA GeForce GTX 1070 GPUs @ 3.4 GHz, 64 GB DDR, and Windows 10 system was used for the implementation.

5.3.1 Training and testing

First, to get the locations of the excavators and trucks in video frames, the YOLO-v3 detection model was trained to detect excavators and trucks in the video frames. A dataset containing 1,191 images of excavators and trucks (1,071 excavators, 871 trucks) was created to train the detector. In the training process, the learning rate is set to 0.1, and an Adam optimizer was used to adjust the learning rate during each epoch. The batch size was set to 6. It took about 10 hours to reach to the validation loss not decreasing after 350 epochs. Then, the detection model was tested on the test dataset with 300 images (362 excavators, 421 trucks). The test results are shown in Table 5-2. Three indicators, precision, recall and accuracy were calculated to measure the performance of the activity recognition model. The detection precisions for the excavators and trucks are 98% and 99%, respectively. The recall rates are 93% and 72%, respectively. The average accuracy of the detection is 82%, which shows that the model has a good ability to identify excavators and trucks in video frames.

Table 5-2 Detection results

Confusion matrix	Predict class			Model performance		
	Excavator	Truck	None	Precision (%)	Recall (%)	Accuracy (%)
Excavator	337	2	23	98	93	93
Truck	6	303	112	99	72	72
Average	_					82

5.3.2 Case Studies

In this section, three case studies were conducted to validate the feasibility of the proposed method for idling reasons identification. Case Studies 1 and 2 have one excavator working with several trucks. Case Study 3 has two excavators working with four trucks and was used to test the proposed clustering method. The test videos used in these case studies contain different idling reasons.

(1) **Case Study 1**

In Case Study 1, a video of about 62 min of earthmoving work was used for testing. The video has the resolution of 1920 × 1080 pixels and the frame rate of 30 fps (110,914 image frames). In this video, one excavator has 2,645 s idling time and 1,052 s working time. The idling and working states of the excavator and trucks were identified based on the proposed method. Sensitive analysis has been done to select the thresholds' values based on the closest matching results of the idling time compared to ground truth value. The results are shown in Table 5-3 and Table 5-4.

Table 5-3 Sensitive analysis to choose the threshold L and α (Case Study 1)

L	α (pixel)	Estimated idling time from $STD(\Delta d_i)$ (s)	Ground truth (s)
80	5	2,328	2,645
	7	2,489	
	9	2,762	
100	5	2,413	
	7	**2,705**	
	9	2,886	
120	5	2,278	
	7	2,488	
	9	2,720	

Table 5-4 Sensitive analysis to choose the threshold L and β (Case Study 1)

L	β (%)	Estimated idling time from $STD(\Delta a_i)$ (s)	Ground truth (s)
80	1	1,387	2,645
	2	2,783	
	3	2,918	
100	1	1,277	
	2	**2,612**	
	3	2,881	
120	1	1,264	
	2	2,208	
	3	2,836	

From the table, the width of the sliding window was selected as L = 100 frames for excavators idling states identification. The thresholds d' and α of the excavator were selected as 7 pixels and 2% of average bounding box areas. The activity recognition of the truck is the same as the excavator as explained in the previous section. The threshold of trucks was selected as L = 100 frames, and d' =10 pixels. The comparison of ground truths and estimated results are shown in Figure 5-3. The estimated idling and working times are 2,612 s and 1,085 s, respectively. The error rates are 1.2% and 3.1%, respectively.

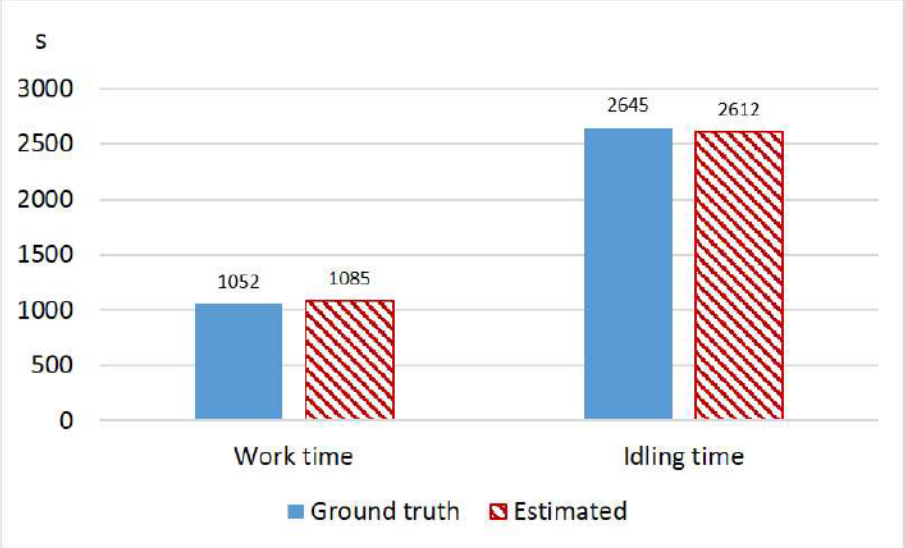


Figure 5-3 Estimated excavator idling and working time with ground truth (Case Study 1)

The idling time of the excavator was further analyzed to identify the reasons that caused excavator's idling. In this video, there are three reasons of excavator's idling: Case 1 (i.e. excavator is waiting for a truck), Case 2 (i.e. excavator is waiting for truck maneuvering), and Case 3 (i.e. unknown reason). The accuracy of the estimated results of these three cases are 99%, 82%, and 98%, respectively, as shown in Figure 5-4. Examples of the results are shown in Figure 5-5.

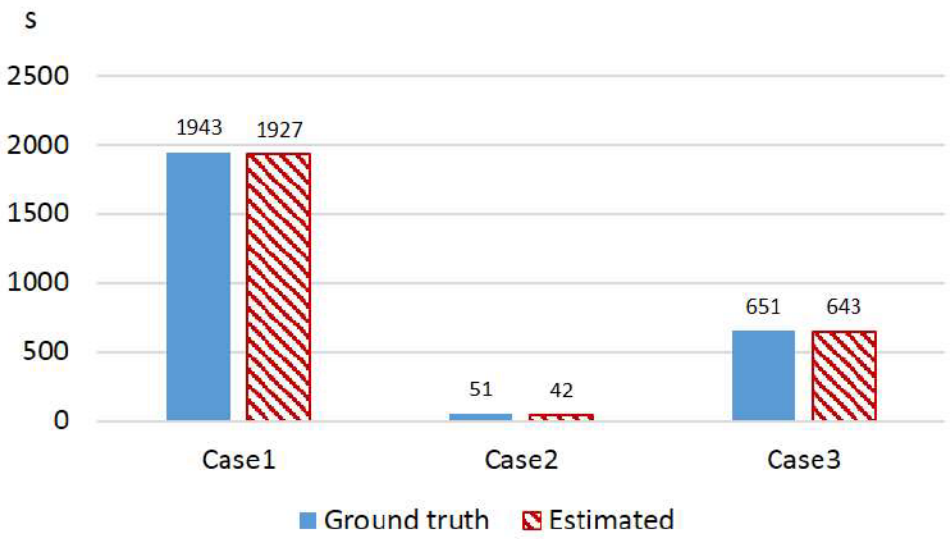
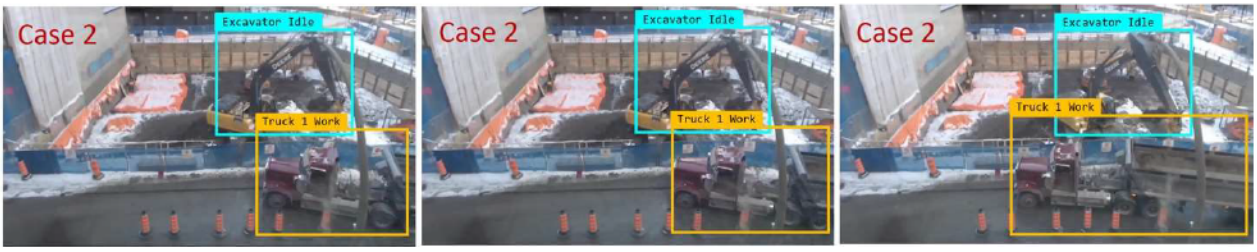
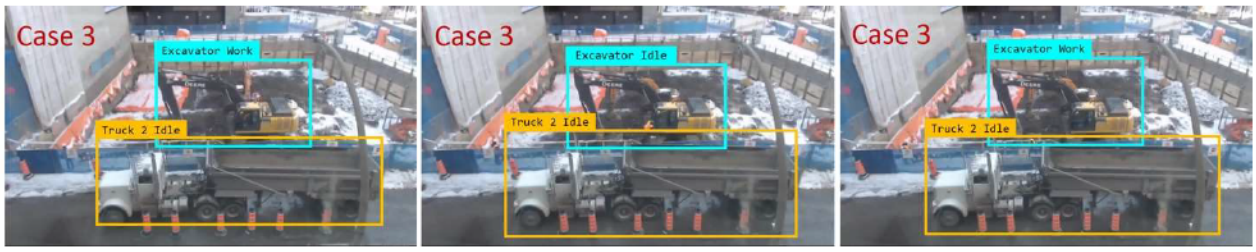


Figure 5-4 Results of idling reasons analysis with ground truth (Case Study 1)

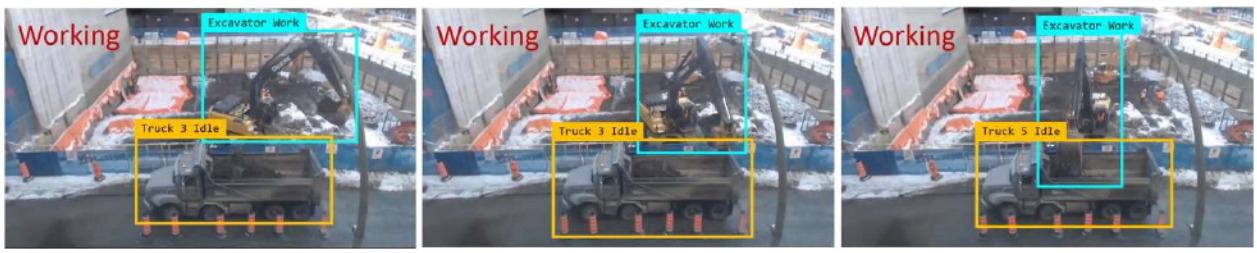
(a) Case 1: Excavator is waiting for truck



(b) Case 2: Excavator is waiting for truck maneuvering



(c) Case 3: Excavator and truck are idling (unknown reason)



(d) Excavator is working

Figure 5-5 Examples of the results of Case Study 1

The results show that the identification of Case 2 has the maximum error rate of 18%. The errors are mainly due to the failure of detection of partial appearances of trucks as can be seen in Figure 5-6. This condition appeared from time T = 2,535 s to T = 2,540 s and T = 3,153 s to T = 3,156 s, which decreased the estimated moving time of trucks. If the errors of the detection results are

excluded, the accuracy of Cases 1-3 are 100%, 97%, and 99%, respectively. The confusion matrix is shown in Figure 5-7.

Figure 5-6 Example of lost detection for the partial truck (Case Study 1)

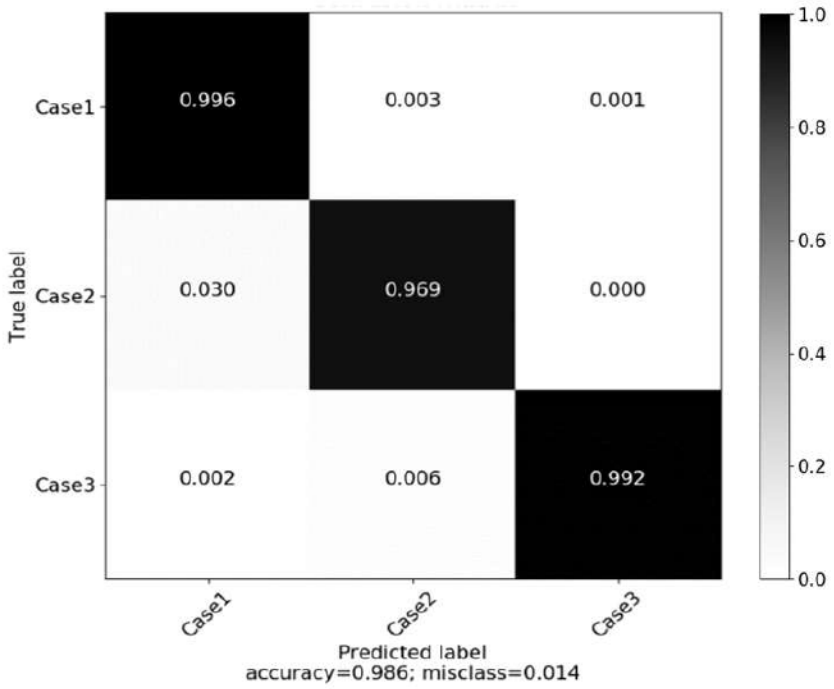


Figure 5-7 Confusion matrix for idling reasons identification (Case Study 1)

The results of Case Study 1 show that during 62 min earthmoving work, the excavator's idling time is 60% of total operation time. The proportion of each case is shown in Figure 5-8. Among these

three cases, Case 1 consumes 74% of the total idling time, which indicates that a limited number of trucks were arranged to work with the excavator. By observing the video, it could be noticed that the average cycle time of trucks is about 20 min, and the loading time per truck is about 4 min. As explained in the proposed method, the balanced number of trucks should be 5 according to Equation 5-4.

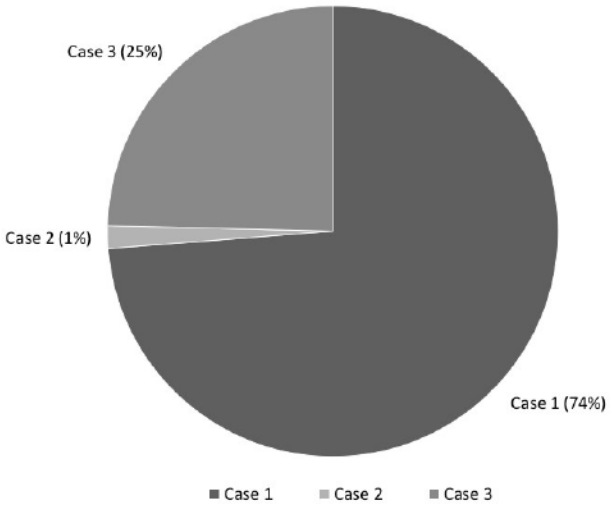
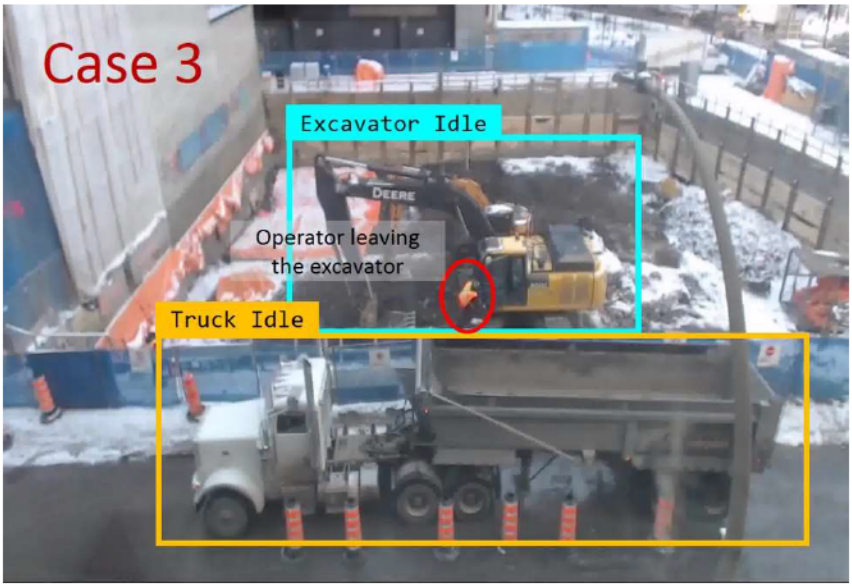
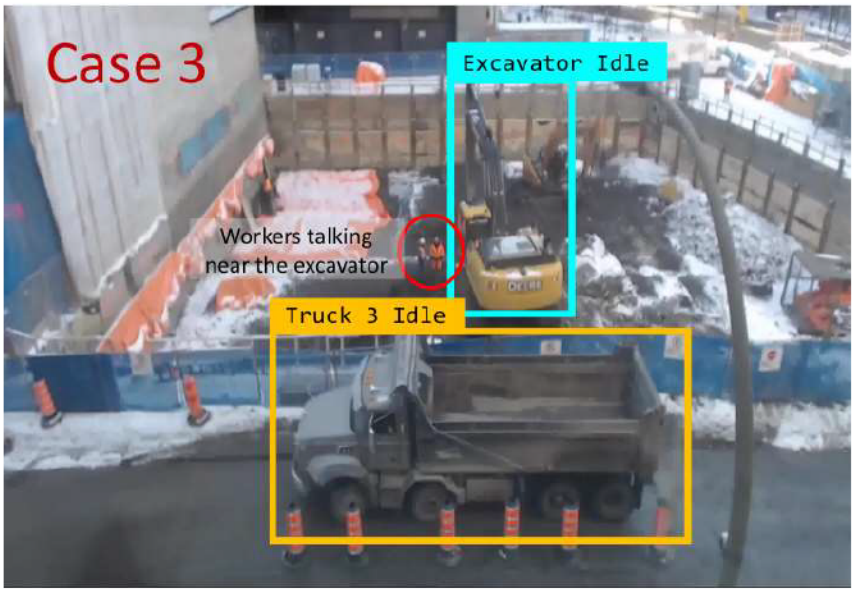


Figure 5-8 Percentage of the reasons for idling (Case Study 1)

Case 3 consumes 25 % of the total idling time. From the video time $T = 51$ s to $T = 520$ s, it can be observed that the operator of the excavator left the equipment, as shown in Figure 5-9(a). From $T = 3{,}159$ s to $T = 3{,}358$ s, it can be observed that two persons were talking near the excavator, and the excavator started to work after they left, as shown in Figure 5-9(b). In order to reduce the idling time caused by Case 3, the workers should not be allowed to do non-productive works near the equipment.



(a) Operator leaving the excavator



(b) People talking near the excavator

Figure 5-9 Examples of Case 3 (Case Study 1)

(2) Case Study 2

In Case Study 2, a video of 10 min of earthwork was used for testing. The video has the resolution of 1280 × 720 pixels and the frame rate of 25 fps (15,000 image frames). In this video, Excavator 1 has 139 s idling time and 461 s working time, and Excavator 2 has 600 s working time. The method of equipment activity identification is the same as Case study 1. The thresholds of

excavator are selected $d' = 1.5$ pixel and $\alpha = 2\%$, and the threshold of truck was selected as $d' = 2$ pixels based on the sensitivity analysis. As explained in the proposed method, the thresholds are based on the resolution of the video and the distance of the equipment from the camera. The width of the sliding window L was selected as 100 frames. The comparison of ground truths and estimated results are shown in Figure 5-10. The estimated idling and working times are 143 s and 1057 s, respectively. The error rates are 3% and 1%, respectively.

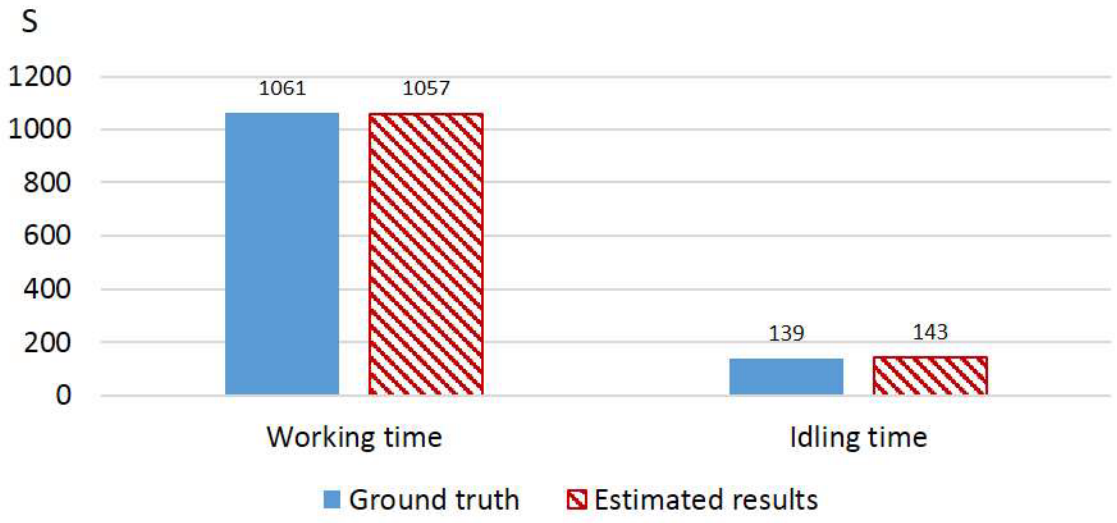


Figure 5-10 Estimated excavator idling and working time with ground truth (Case Study 2)

The idling time of the excavator was further analyzed to identify the reasons that caused excavator's idling. In this video, there are three reasons for excavator's idling: Case 2 (i.e. excavator is waiting for truck maneuvering), Case 3 (i.e. unknown reason) and Case 4 (i.e. congested site with many trucks). There is also Case 5 for truck's idling. All the idling cases appeared during Excavator 1's operation. Excavator 2 was working alone. The accuracy of the estimated results of the four cases are 98%, 75%, 67% and 95%, respectively, as shown in Figure 5-11. Examples of the results are shown in Figure 5-12.

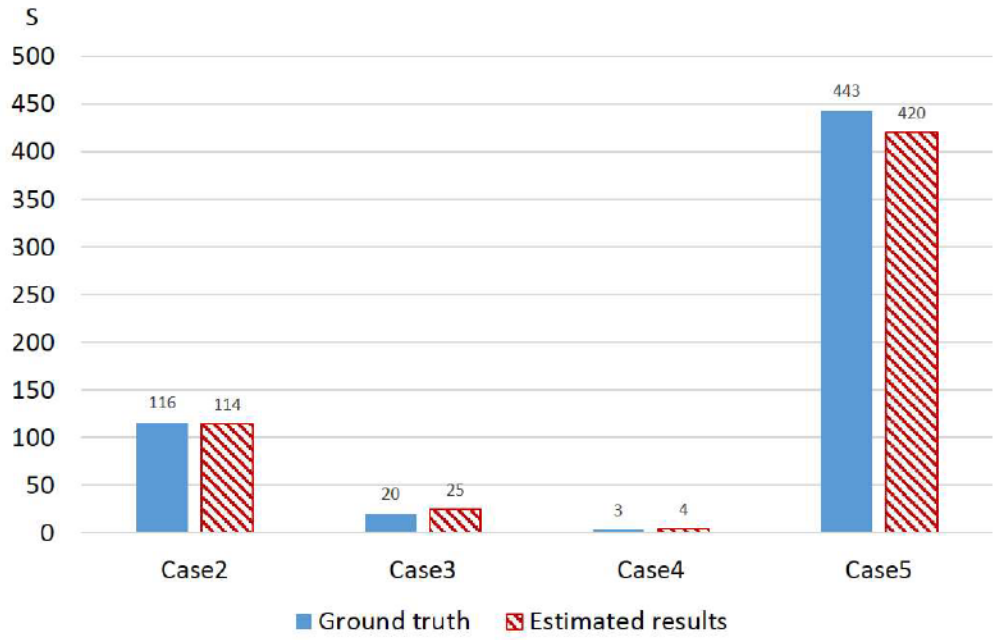


Figure 5-11 Results of idling reasons analysis with ground truth (Case Study 2)

Figure 5-12 Examples of the results of Case Study 2

The results show that the identification of Case 4 has the maximum error rate of 33%. The high error rate is mainly due to the failure of detecting trucks under occlusion as shown in Figure 5-13 (truck occluded behind excavator). This situation appeared from time $T = 281$ s to $T = 282$ s which decreased the estimated number of trucks. Excluding the detection failures, the accuracy rates of Cases 2-5 are 96%, 86%, 100% and 99%, respectively, as shown in the confusion matrix in Figure 5-14.

Figure 5-13 Example of trucks' occlusion (Case Study 2)

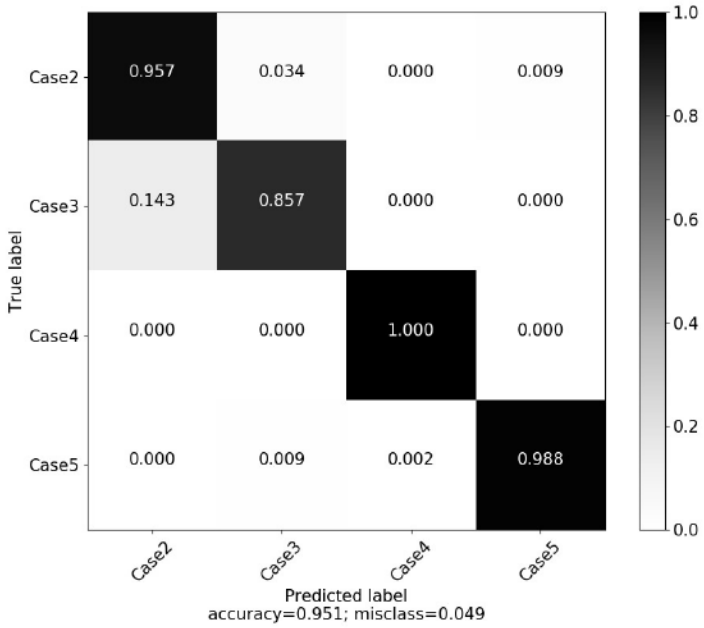


Figure 5-14 Confusion matrix for idling reasons identification (Case Study 2)

The results of Case Study 2 show that during 10 min earthmoving work, the excavator's idling time is 31% of total operation time. The proportion of each case is listed in Figure 5-15. Among the three cases, Case 2 accounts for 80% of total idling time, which indicates the idling is mainly caused by trucks maneuvering to the loading position. Moreover, there is 420 s idling time of Case 5, which indicates that many trucks were waiting to be loaded. In order to reduce the idling time, less trucks should be arranged to work with one excavator. In addition, to reduce the idling time caused by the congested site, more detailed site planning should be made for trucks' paths.

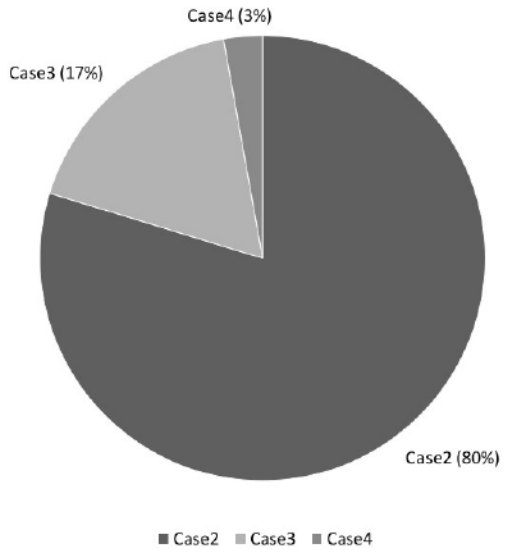


Figure 5-15 Percentage of the reasons for excavator idling (Case Study 2)

(3) Case Study 3

In Case Study 3, a video of about 11 min of earthmoving work was used for testing. The video has the resolution of 1920 × 1080 pixels and the frame rate of 20 fps (14140 image frames). In this video, there are two excavators working with four trucks.

In the activity recognition, the width of the sliding window was selected as L = 100 frames and the thresholds d' and $\alpha$of the excavator were selected as 0.7 pixels and 5% of average bounding box areas, respectively. The threshold of trucks was selected as L = 100 frames, and d' =1 pixel. The method of equipment activity recognition is the same as Case study 1. The comparison of the ground truths and estimated results are shown in Figure 5-16. The average error rates of idling and working states estimation are 1% and 2%, respectively.

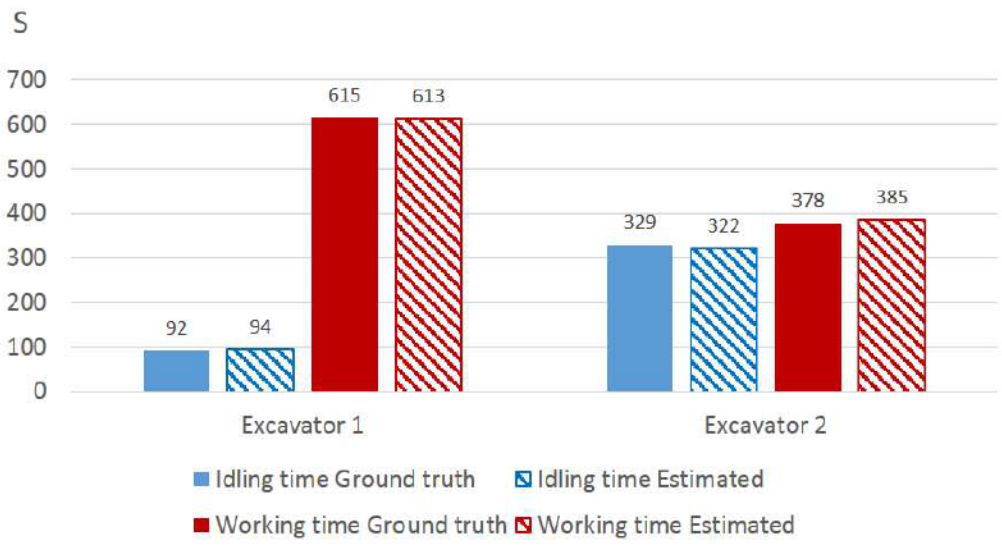


Figure 5-16 Estimated excavator idling and working time with ground truth (Case Study 3)

In order to cluster the trucks with the excavators, the camera has to be calibrated to calculate the real distance between excavators and trucks. Two checkerboards were printed with square sizes of 20 × 20 cm and 10 × 10 cm for calibration, as shown in Figure 5-17. In the tests, each checkerboard was placed at different orientations and locations, and recorded by the camera. Finally, 16 images were selected, based on the visibility of the board, to calculate the intrinsic and extrinsic parameters of the camera. The parameters were automatically calculated using Camera Calibration Toolbox for MATLAB developed by Bouguet (2004). The checkerboard with the squares size of 20 × 20 cm achieved higher accuracy. The results of M, and R are $\begin{bmatrix} 0.0012 & 0 & -0.8266 \\ 0 & 0.0014 & 0.2704 \\ 0 & 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 0.9958 & 0.0875 & 0.2739 \\ -0.0087 & -0.2071 & 0.9782 \\ 0.0913 & -0.9744 & -0.2056 \end{bmatrix}$, respectively.

(a) Checkerboard with 20 × 20 cm squares

(b) Checkerboard with 10 × 10 cm squares

Figure 5-17 Examples of two checkerboards for calibration

In the work group clustering step, considering the average size of the excavators (14.8 m × 4.5 m) and trucks (7 m × 3m), as well as the relative position of the truck with the excavator in this video, the threshold γ was selected as 20 m. The clustering of the excavators and trucks achieved the accuracy of 100%. Examples of the clustering results are shown in Figure 5-18. In the test video, Excavator 1 was clustered with four trucks, individually. Excavator 2 was not clustered with any truck.

Figure 5-18 Examples of the clustering result

The idling time of the excavators was further analyzed to identify the reasons that caused excavator's idling. In this video, Excavator 1 had two reasons of excavator's idling: Case 2 (i.e. excavator is waiting for truck maneuvering), and Case 3 (i.e. unknown reason). Excavator 2 had one reason of idling: Case 1 (i.e. excavator is waiting for a truck), and it worked alone for 378 sec. The results of idling reasons analysis are shown in Figure 5-19. The accuracy of the estimated results of these three cases are 96%, 91% and 98%, respectively.

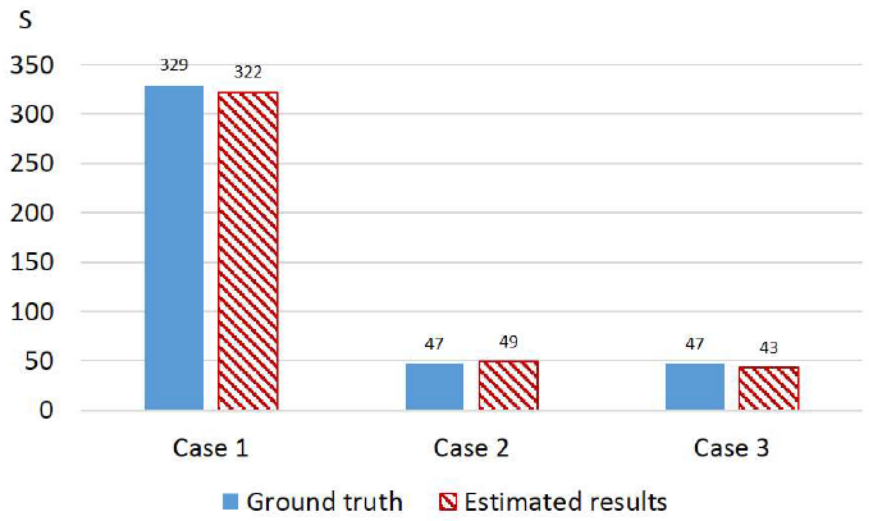


Figure 5-19 Results of idling reasons analysis with ground truth (Case Study 3)

Examples of the test results are shown in Figure 5-20. From the test results, it can be noticed that Case 2 and Case 3 each consists of 50% of the total idling time of Excavator 1. The idling time consists 15% of total operation time. Case 2 was caused by the site condition, because the truck path was too narrow for positioning. Case 3 occurred in the last 43 seconds, which may due to operator rest. Excavator 2 had only Case 1, which requires more trucks to be arranged to work with the excavator.

Figure 5-20 Example of the results of Case Study 3

5.4 Summary and Conclusions

This chapter presented a CV-based method to identify the potential reasons that cause excavator's idling. The proposed method consists of three main steps: equipment working and idling states identification, excavators and trucks group clustering, and idling reasons analysis. The effectiveness of the proposed method has been tested on three construction surveillance videos. The implementation results illustrated that the proposed method has a good ability to analyze the reasons of equipment's idling from site videos. The summary of the results of the proposed method on test videos is shown in Table 5-5. The method is able to identify four kinds of reasons that cause excavator's idling, and one reason of truck's idling with the average accuracy of 90%. The method can also quantify the idling time associated with each cause.

The main contribution of this chapter is developing a novel CV-based method to automatically identify the idling reasons of excavators and trucks based on their interactive work states. To the best knowledge of the authors, this is the first attempt to automatically classify the reasons of equipment's idling into detailed categories using CV. The proposed method provides an efficient solution to explore the reasons of low productivity from site surveillance videos, which could contribute to better understanding of earthmoving productivity under the dynamic and complex construction site conditions. Furthermore, the in-depth understanding of the low productivity reasons can further support the cost control and resource allocation in future construction project management.

Table 5-5 Summary of the test results

Case studies	Accuracy of detection of idling reason				
	Case 1 (%)	Case 2 (%)	Case 3 (%)	Case 4 (%)	Case 5 (%)
Case Study 1	99	82	98	_	_
Case Study 2	_	98	75	67	95
Case Study 3	98	96	91	_	_
Average	90%				

CHAPTER 6: SUMMARY, CONTRIBUTIONS, AND FUTURE WORK

6.1 Introduction

This chapter first summarizes all the works that have been done in this research. Then, the contributions and conclusions are introduced in detail. Accordingly, the limitations and future work of this research is introduced.

6.2 Summary of Research

This research covered the review of the related literature, the existing research gaps, the overview of the proposed framework, and a detailed explanation of the proposed methods followed by the implementation to validate the feasibility of the proposed framework.

In Chapter 2, a roadmap is proposed to show the technology path forward for automatic equipment productivity monitoring and provide future directions that will support the development of full automation in monitoring construction equipment.

In Chapter 4, a novel fully automatic CV-based framework was proposed for multiple excavators' operation monitoring and productivity calculation. The framework integrates the detection, tracking, activity recognition and productivity calculation modules. The detection module identifies all the excavators in video frames. The tracking module correlates the same excavators and extracts their regions across the frames in the video sequences. After detecting the idling state, the activities of each excavator are recognized based on the tracking results. Finally, the productivity is automatically calculated based on the activity information of the excavators. The effectiveness of the proposed framework has been tested on construction surveillance videos, and the results have shown the feasibility of the proposed framework.

In Chapter 5, a CV-based method was proposed to identify the potential reasons that cause excavator's idling. The proposed method consists of three main steps: equipment working and idling states identification, excavators and trucks group clustering, and idling reasons analysis. The effectiveness of the proposed method has been tested on three construction surveillance videos. The implementation results illustrated that the proposed method has a good ability to analyze the reasons of equipment's idling from site videos. The summary of the results of the proposed method

on test videos is shown in Table 5-5. The method is able to identify four kinds of reasons that cause excavator's idling, and one reason of truck's idling with the average accuracy of 90%. The method can also quantify the idling time associated with each cause.

6.3 Research Contributions and Conclusions

This research work made the following contributions to the body of knowledge：

(1) A roadmap is proposed to show the advances in each automatic equipment productivity monitoring method, and future research directions of this domain are proposed. The roadmap shows the research paths towards more comprehensive equipment productivity management by integrating different research areas and leveraging new advancements in computer science.

(2) The advanced technology in computer vision is integrated to recognize continuous activities of excavators in long site surveillance videos. With regard to this contribution the following conclusions can be drawn:

- Compared with the recent work of Kim et al. (2018b) which could just identify working and idling states of the excavator based on its relative location with respect to the truck, the proposed method can recognize detailed activities such as digging, loading and swinging from activities spatial and temporal features.
- The proposed method is also superior to another related work of Kim and Chi (2019) in long videos analysis, since they did not consider the temporal features of the activities. In Kim and Chi (2019) method, activities are recognized based on their sequence relations. The method in this research can recognize activities based on both spatial and temporal features that are directly learned from videos regardless of their sequential relations.
- In the model training stage, the proposed method relied on several training strategies, such as fine-tuning, data augmentation and batch size optimization, to train the detection, tracking and activity recognition models with a limited dataset. The results of the activity recognition on different excavator sizes and types, various viewpoints and lighting conditions, and background movements have proved its robustness in construction environments and provided a solid basis to automate the calculation of the equipment productivity.

(3) The detection, tracking and recognition techniques are integrated to measure multiple excavators' operations. In a recent construction activity recognition research, Luo et al. (2018) manually specified bounding boxes to start the tracking process, which is not efficient. In this research, a fully automated framework is conducted for excavators' activity recognition without the manual input.

(4) A sliding window method is applied to identify the idling state of excavators by comparing the changes of centroid distance and areas of bounding boxes of excavators in consecutive video frames.

(5) An automatic method is developed to directly analyze the productivity based on the activity recognition results. The experimental results on the video with the duration of nearly one hour supported the feasibility and applicability of the proposed method in practice. Moreover, the detailed operation information provided by this method, such as the time of each activity and the number of work cycles, can be further used to help construction managers identify the causes of excavator's low productivity from the video.

(6) A novel CV-based method was developed to automatically identify the idling reasons of excavators and trucks based on their interactive work states. To the best knowledge of the authors, this is the first attempt to automatically classify the reasons of equipment's idling into detailed categories using CV. The proposed method provides an efficient solution to explore the reasons of low productivity from site surveillance videos, which could contribute to better understanding of earthmoving productivity under the dynamic and complex construction site conditions. Furthermore, the in-depth understanding of the low productivity reasons can further support the cost control and resource allocation in future construction project management.

6.4 Limitations and Future Work

Despite the contributions of this work, there are still some limitations to be addressed in future work as follows:

(1) The activity recognition performance is affected by the detection and tracking results. When the bounding boxes of two excavators are fully overlapping in the detection and tracking results, the activity of one excavator maybe affected by the other excavator as shown in Figure 4-12.

- To improve the robustness of the method, videos taken from more than one camera should be aligned and applied to recognize the activities.

(2) The light condition of the video also affects the activity recognition result. When the light is too bright or too dark, it is difficult to recognize the moving features in video frames, as shown in Figure 4-10.

- To avoid the influence of the bad light condition, the video should be preprocessed by adding a filter.

(3) It is noted that the diversity of the activity dataset is limited, which has negative influences on both recognition and productivity results. For instance, some of the videos in the dataset are taken from low heights. However, most of the construction surveillance videos are overlooking from high points.

- In order to train a better performance model for activity recognition, various data collected from the construction site surveillance video should be added to the dataset.

- To improve the efficiency and optimize excavator's productivity, the activities of the truck that works with the excavator should also be recognized. By analyzing the relationship of excavator's cycle time with the truck's position, excavator's productivity can be optimized (e.g. by improving the maneuvering). Moreover, by analyzing the cycle time of the truck, the number of trucks required to keep the excavator working at full capacity can be calculated. This could be done based on identifying the work groups of excavators and trucks.

- For the purpose of productivity calculation, having long videos (i.e. about an hour) is important to capture the variation in cycle times. However, the process of validation long videos is time and labor intensive. Therefore, more long videos will be tested in the future.

(4) From the case studies' results, it could be noticed that the error of idling reasons identification mainly came from previous steps. For example, the failure in detecting partially occluded appearance trucks has decreased the estimated idling time in Case Study 2 (Section 5.2). If the equipment is not detected, it would be difficult to apply the subsequent idling identification method.

- To improve the robustness of the proposed method, more advanced CV-based methods should be tested in the future (e.g. using multiple cameras).
- To avoid the error caused by overlapped equipment, more than one camera should be applied.
- To improve the accuracy, the future work can calibrate and align two cameras to detect excavators and trucks and calculate the 3D geolocation.

(5) The proposed method can just monitor the operation processes between excavators and trucks, which is not enough for earthwork productivity analysis.

- The method will be expanded to analyze the work states and interactive work between other types of equipment, such as dozers and loaders.

(6) The activity recognition of the excavator did not include all work states of excavators, which is not efficient for monitoring the operation process of the excavator.

- The activity recognition of the excavator can be extended to recognize more detailed activities, such as the equipment relocation and inspection, which will contribute to a better understanding of equipment productivity.

(7) In the identification of the idling reasons, there are some unknown reasons in Case 3 that cannot be determined, such as safety related reasons.

- By combining with the worker detection, the proposed method can estimate safety distance between workers with equipment and explore more detailed information about the idling reasons.

REFERENCES

AbouRizk, S. (2010). "Role of Simulation in Construction Engineering and Management." *Journal of Construction Engineering and Management*, American Society of Civil Engineers, 136(10), 1140–1153.

Ahn, C., Lee, S., Peña-Mora, F., and Schapiro, A. (2012). "Monitoring System for Operational Efficiency and Environmental Performance of Construction Operations, Using Vibration Signal Analysis." *Construction Research Congress 2012*, pp. 1879-1888.

Ahn, C. R., Lee, S., and Peña-Mora, F. (2015). "Application of Low-Cost Accelerometers for Measuring the Operational Efficiency of a Construction Equipment Fleet." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 29(2), 04014042.

Ahn, S., Han, S., and Al-Hussein, M. (2020). "Improvement of transportation cost estimation for prefabricated construction using geo-fence-based large-scale GPS data feature extraction and support vector regression." *Advanced Engineering Informatics*, 43, 101012.

Akhavian, R., and Behzadan, A. H. (2012). "An integrated data collection and analysis framework for remote monitoring and planning of construction operations." *Advanced Engineering Informatics*, *EG-ICE 2011 + SI: Modern Concurrent Engineering*, 26(4), 749–761.

Akhavian, R., and Behzadan, A. H. (2013). "Knowledge-Based Simulation Modeling of Construction Fleet Operations Using Multimodal-Process Data Mining." *Journal of Construction Engineering and Management*, American Society of Civil Engineers, 139(11), 04013021.

Akhavian, R., and Behzadan, A. H. (2015). "Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers." *Advanced Engineering Informatics*, Collective Intelligence Modeling, Analysis, and Synthesis for Innovative Engineering Decision Making, 29(4), 867–877.

Alshibani, D. A., and Moselhi, D. O. (2016). "Productivity based method for forecasting cost & time of earthmoving operations using sampling GPS data." *Journal of Information Technology in Construction (ITcon)*, 21(3), 39–56.

Arditi, D., and Mochtar, K. (2000) "Trends in productivity improvement in the US construction industry." *Construction Management and Economics*, 18 (2000), 15-27.

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

Azar, E., and McCabe, B. (2012b). "Automated visual recognition of dump trucks in Construction Videos." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 26(6), 769–781.

Azar, E., Dickinson, S., and McCabe, B. (2013). "Server-Customer interaction tracker: computer vision–based system to estimate dirt-loading cycles." *Journal of Construction Engineering and Management*, 139, 785–794.

Azar, E. R. (2016). "Construction equipment identification using marker-based recognition and an active zoom camera." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 30(3), 04015033.

Azar, E. (2017). "Semantic annotation of videos from equipment-intensive construction operations by shot recognition and probabilistic reasoning." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 31(5), 04017042.

Bae, J., Kim, K., and Hong, D. (2019). "Automatic identification of excavator activities using joystick signals." *International Journal of Precision Engineering and Manufacturing*, 20(12), 2101–2107.

Bastan, M., Yap, K., and Chau, L. (2018). "Idling car detection with ConvNets in infrared image sequences." *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5.

Bewley, A., Ge, Z., Ott, L., Ramos, F., and Upcroft, B. (2016). "Simple online and realtime tracking." *arXiv:1602.00763 [cs.CV]*. https://arxiv.org/abs/1602.00763.

Bouet, M., and Santos, A. L. dos. (2008). "RFID tags: Positioning principles and localization techniques." *2008 1st IFIP Wireless Days*, 1–5.

Bouguet, J. (2000). "Pyramidal implementation of the Lucas Kanade feature tracker." *Intel Corporation, Microprocessor Research Labs*. http://robots.stanford.edu/cs223b04/algo_affine_tracking.pdf.

Bouguet, J. Y. (2004). Camera calibration toolbox for MATLAB. Santa Clara, *CA: Intel Corporation*.

Bügler, M., Borrmann, A., Ogunmakin, G., Vela, P. A., and Teizer, J. (2017). "Fusion of Photogrammetry and Video Analysis for Productivity Assessment of Earthwork Processes." *Computer-Aided Civil and Infrastructure Engineering*, 32(2), 107–123.

Bügler, M., Ongunmakin, G., Teizer, J., Vela, P., and Borrmann, A. (2014). "A comprehensive methodology for vision-based progress and activity estimation of excavation processes for productivity assessment." *in: The 21st International Workshop: Intelligent Computing in Engineering, EG-ICE 2014*.

Cai, J., and Cai, H. (2020). "Robust hybrid approach of vision-based tracking and radio-based identification and localization for 3D tracking of multiple construction workers." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 34(4), 04020021.

Chen, C., Zhu, Z., and Hammad, A. (2020a). "Automated excavators activity recognition and productivity analysis from construction site surveillance videos." *Automation in Construction*, 110, 103045.

Chen, C., Zhu, Z., and Hammad, A. (2020b). "Automatic analysis of idling in excavator's operations based on excavator-truck relationships." *ISARC Proceedings*, IAARC, 1307–1313.

Chen, H., Luo, X., Zheng, Z., and Ke, J. (2019). "A proactive workers' safety risk evaluation framework based on position and posture data fusion." *Automation in Construction*, 98, 275–288.

Cheng, C.F., Rashidi, A., Davenport, M. A., and Anderson, D. V. (2017). "Activity analysis of construction equipment using audio signals and support vector machines." *Automation in Construction*, 81, 240–253.

Cheng, C.-F., Rashidi, A., Davenport, M. A., and Anderson, D. V. (2019). "Evaluation of Software and Hardware Settings for Audio-Based Analysis of Construction Operations." *International Journal of Civil Engineering*, 17(9), 1469–1480.

Cheng, T., Venugopal, M., Teizer, J., and Vela, P. A. (2011). "Performance evaluation of ultra wide band 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.

Dai, J., Li, Y., He, K., and Sun, J. (2016). "R-FCN: Object detection via region-based fully convolutional networks." *arXiv:1605.06409 [cs]*.

Dalal, N., Triggs, B., Schmid, C., Dalal, N., Triggs, B., Schmid, C., Detection, H., and Oriented, U. (2006). "Human detection using oriented histograms of flow and appearance." *European Conference on Computer Vision (ECCV '06)*, 428–441.

Diba, A., Fayyaz, M., Sharma, V., Karami, A. H., Arzani, M. M., Yousefzadeh, R., and Van Gool, L. (2017). "Temporal 3D ConvNets: new architecture and transfer learning for video classification." *arXiv:1711.08200 [cs.CV]*. https://arxiv.org/abs/1711.08200.

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2017). " Long-term recurrent convolutional networks for visual recognition and description." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 677–691.

Edwards, D., and Griffiths, I. (2000). "Artificial intelligence approach to calculation of hydraulic excavator cycle time and output." *Mining Technology: Transactions of the Institutions of Mining and Metallurgy*, 109, 23–29.

Edwards, D., and HOLT, G. (2000). "ESTIVATE: A model for calculating excavator productivity and output costs." *Engineering, Construction and Architectural Management*, 7, 52–62.

Eleonora, V., and David, C. (2012). "Space-variant descriptor sampling for action recognition based on saliency and eye movements." *Eccv2012*, 84–97.

Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M., and An, W. (2018a). "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos." *Automation in Construction*, 85, 1–9.

Fang, W., Ding, L., Zhong, B., Love, P. E. D., and Luo, H. (2018b). "Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach." *Advanced Engineering Informatics*, 37, 139–149.

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). "Convolutional two-stream network fusion for video action recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1933-1941.

Feng, C., Dong, S., Lundeen, K. M., Xiao, Y., and Kamat, V. R. (2015). "Vision-based articulated machine pose estimation for excavation monitoring and guidance." *ISARC Proceedings*, IAARC, 1–9.

Fukunaga, K. (2013). "Introduction to statistical pattern recognition." *Academic Press*, 2013,

Girdhar, R., and Ramanan, D. (2017). "Attentional pooling for action recognition." *arXiv:1711.01467 [cs.CV]*. https://arxiv.org/abs/1711.01467.

Girshick, R. (2015). "Fast R-CNN." *arXiv:1504.08083 [cs]*.

Golparvar-Fard, M., Heydarian, A., and Niebles, J. C. (2013). "Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers." *Advanced Engineering Informatics*, 27(4), 652–663.

Gong, J., and Caldas, C. H. (2011). "An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations." *Automation in Construction*, 20(8), 1211–1226.

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*, Special Section: Advances and Challenges in Computing in Civil and Building Engineering, 25(4), 771–782.

Goodrum, P. M., Haas, C. T., and Glover, H. W. (2002). "The divergence in aggregate and activity estimates of US construction productivity." *Construction Management and Economics*, 20(2002), 415-423.

Haag, S., and Anderl, R. (2018). "Digital twin – Proof of concept." *Manufacturing Letters*, Industry 4.0 and Smart Manufacturing, 15, 64–66.

Han, S. H., Hong, T. H., and Lee, S. L. (2008). "Production prediction of conventional and global positioning system–based earthmoving systems using simulation and multiple regression analysis." *Canadian Journal of Civil Engineering*, 35(6):574–87.

Hannus, M. Construction ICT Roadmap, *ROADCON Project Deliverable* Rep. No. D5, IST-2001-37278.

Hara, K., Kataoka, H., and Satoh, Y. (2018). "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" *arXiv:1711.09577 [cs]*.

He, K., Zhang, X., Ren, S., and Sun, J. (2017). " Deep Residual Learning for Image Recognition." *arXiv:1512.03385 [cs.CV]*. https://arxiv.org/abs/1512.03385.

Heikkila, J., Silven, O. (1997) "A four-step camera calibration procedure with implicit image correction." *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1106 – 1112. DOI: 10.1109/CVPR.1997.609468.

Henriques, J. F., Caseiro, R., Martins, P., and Batista, J. (2012). "Exploiting the circulant structure of tracking-by-detection with kernels." *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science, Berlin, Heidelberg, 702–715.

Henriques, J. F., Caseiro, R., Martins, P., and Batista, J. (2015). "High-Speed tracking with kernelized correlation filters." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3), 583–596.

Hernandez, C., Slaton, T., Balali, V., and Akhavian, R. (2019). "A Deep learning framework for construction equipment activity analysis." *ASCE International Conference on Computing in Civil Engineering 2019*, 479–486.

Hola, B., and Schabowicz, K. (2010). "Estimation of earthworks execution time cost by means of artificial neural networks." *Automation in Construction*, Building Information Modeling and Collaborative Working Environments, 19(5), 570–579.

Holt, G., and Edwards, D. (2015). "Analysis of interrelationships among excavator productivity modifying factors." *International Journal of Productivity and Performance Management*, 64, 853–869.

Ibrahim, M., and Moselhi, O. (2014). "Automated productivity assessment of earthmoving operations." *Journal of Information Technology in Construction (ITcon)*, 19(9), 169–184.

Jain, A., Sharma, A., Borana, s. L., Ravindra, B., and Mangalhara, J. P. (2018). "Study and analysis of exhaust emission of diesel vehicles using thermal IR imagers." *Defence Science Journal*, 68, 533–539.

Jiang, Y. G., Dai, Q., Xue, X., Liu, W., and Ngo, C. W. (2012). "Trajectory-based modeling of human actions with motion reference points." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7576 LNCS(PART 5), 425–438.

Kalal, Z., Mikolajczyk, K., and Matas, J. (2010). "Forward-backward error: automatic detection of tracking failures." *2010 20th International Conference on Pattern Recognition*, 2756–2759.

Kalal, Z., Mikolajczyk, K., and Matas, J. (2012). "Tracking-Learning-Detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7), 1409–1422.

Kang, S. M., and Wildes, R. P. (2016). "Review of action recognition and detection methods." *arXiv:1610.06906 [cs.CV]*. https://arxiv.org/abs/1610.06906.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. (2017). "The kinetics human action video dataset." *arXiv:1705.06950 [cs.CV]*.https://arxiv.org/abs/1705.06950.

Khajavi, S. H., Motlagh, N. H., Jaribion, A., Werner, L. C., and Holmström, J. (2019). "Digital twin: vision, benefits, boundaries, and creation for buildings." *IEEE Access*, 7, 147406–147419.

Kim, H., Ahn, C. R., Engelhaupt, D., and Lee, S. (2018a). "Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement." *Automation in Construction*, 87, 225–234.

Kim, H., Bang, S., Jeong, H., Ham, Y., and Kim, H. (2018b). "Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation." *Automation in Construction*, 92, 188–198.

Kim, H., Ham, Y., Kim, W., Park, S., and Kim, H. (2019). "Vision-based nonintrusive context documentation for earthmoving productivity simulation." *Automation in Construction*, 102, 135–147.

Kim, H., Kim, H., Hong, Y. W., and Byun, H. (2018c). "Detecting construction equipment using a region-based fully convolutional network and transfer learning." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 32(2), 04017082.

Kim, H., Kim, K., and Kim, H. (2016). "Vision-based object-centric safety assessment using fuzzy inference: monitoring struck-by accidents with moving objects." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 30(4), 04015075.

Kim, J., and Chi, S. (2017). "Adaptive detector and tracker on construction sites using functional integration and online learning." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 31(5), 04017026.

Kim, J., and Chi, S. (2019). "Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles." *Automation in Construction*, 104, 255–264.

Kim, J., and Chi, S. (2020). "Multi-camera vision-based productivity monitoring of earthmoving operations." *Automation in Construction*, 112 (2020), 103121.

Kim, J., Chi, S., and Seo, J. (2018d). "Interaction analysis for vision-based activity identification of earthmoving excavators and dump trucks." *Automation in Construction*, 87, 297–308.

Kim, J., Hwang, J., Chi, S., and Seo, J. (2020). "Towards database-free vision-based monitoring on construction sites: A deep active learning approach." *Automation in Construction*, 120, 103376.

Kim, K., Kim, H., and Kim, H. (2017). "Image-based construction hazard avoidance system using augmented reality in wearable device." *Automation in Construction*, 83, 390–403.

Kim, Y. B., Ha, J., Kang, H., Kim, P. Y., Park, J., and Park, F. C. (2013). "Dynamically optimal trajectories for earthmoving excavators." *Automation in Construction*, 35, 568–578.

Klaser, A., Marszalek, M., Schmid, C., Klaser, A., Marszalek, M., Schmid, C., Based, A. S. D., Everingham, G. M., Needham, C., Fraile, R., British, B., Kl, A., Schmid, C., and Grenoble, I. (2010). "A spatio-temporal descriptor based on 3d-gradients" *In Proceedings of the British Machine Vision Conference 2008, Leeds,* 00514853.

Köpüklü, O., Wei, X., and Rigoll, G. (2020). "You Only Watch Once: A unified CNN architecture for real-time spatiotemporal action localization." *arXiv:1911.06644 [cs]*. https://arxiv.org/abs/1911.06644.

Kuhn, H. W. (1955). "The Hungarian method for the assignment problem." *Naval Research Logistics Quarterly*, 2(1–2), 83–97.

Laptev, I., and Pérez, P. (2007). "Retrieving actions in movies." *Proceedings of IEEE 11th International Conference on Computer Vision*. 9849125.

Li, J., Li, H., Umer, W., Wang, H., Xing, X., Zhao, S., and Hou, J. (2020). "Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology." *Automation in Construction*, 109, 103000.

Li, J., Li, H., Wang, H., Umer, W., Fu, H., and Xing, X. (2019). "Evaluating the impact of mental fatigue on construction equipment operators' ability to detect hazards using wearable eye-tracking technology." *Automation in Construction*, 105, 102835.

Li, L., Huang, W., Gu, I., and Tian, Q. (2003). "Foreground object detection from videos containing complex background." *Proceedings of the ACM International Multimedia Conference and Exhibition*, 10.

Li, X., Yi, W., Chi, H.-L., Wang, X., and Chan, A. (2017). "A Critical Review of Virtual and Augmented Reality (VR/AR) Application in Construction Safety." *Automation in Construction*, 86.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). "SSD: Single shot MultiBox detector." *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, International Publishing, Cham, 21–37.

Liu, Y., Tu, Z., Lin, L., Xie, X., and Qin, Q. (2020). "Real-time spatio-temporal action localization via learning motion representation." *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Lowe, D. (2004) "Diatinctive image features from scale-invariant keypoints." *International Journal of Computer Vision*, 60(2), 91-110.

Louis, J., and Dunston, P. S. (2018). "Integrating IoT into operational workflows for real-time and automated decision-making in repetitive construction operations." *Automation in Construction*, 94, 317–327.

Lu, W., Huang, G. Q., and Li, H. (2011). "Scenarios for applying RFID technology in construction project management." *Automation in Construction*, Building Information Modeling and Changing Construction Practices, 20(2), 101–106.

Lucas, B., and Kanade, T. (1981). "An iterative image registration technique with an application to stereo vision." *Proceedings DARPA Image Understanding Workshop, April 1981,* pp. 121-130.

Lundeen, K. M., Dong, S., Fredricks, N., Akula, M., Seo, J., and Kamat, V. R. (2016). "Optical marker-based end effector pose estimation for articulated excavators." *Automation in Construction*, 65, 51–64.

Luo, X., Li, H., Cao, D., Dai, F., Seo, J., and Lee, S. (2018a). "Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks." *Journal of Computing in Civil Engineering*, 32(3), 04018012.

Luo, X., Li, H., Cao, D., Yu, Y., Yang, X., and Huang, T. (2018b). "Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks." *Automation in Construction*, 94, 360–370.

Luo, X., Li, H., Yang, X., Yu, Y., and Cao, D. (2018c). "Capturing and understanding workers' activities in far-field surveillance videos with deep action recognition and Bayesian nonparametric learning." *Computer-Aided Civil and Infrastructure Engineering*, 0 (2018) 1-9.

Luo, X., Li, H., Yu, Y., Zhou, C., Cao, D. (2020) "Combining deep features and activity context to improve recognition of activities of workers in groups." *Computer-Aided Civil and Infrastructure Engineering*, 35, 965-978.

Lustbader, J. A., Venson, T., Adelman, S., Dehart, C., Yeakel, S., and Castillo, M. S. (2012). "Application of sleeper cab thermal management technologies to reduce idle climate control loads in long-haul trucks." *SAE 2012 Commercial Vehicle Engineering Congress,* Warrendale, PA, 2688-3627.

Manyele, S. V. (2017). "Investigation of excavator performance factors in an open-pit mine using Loading cycle time." *Engineering*, Scientific Research Publishing, 9(7), 599–624.

Mathur, N., Aria, S. S., Adams, T., Ahn, C. R., and Lee, S. (2015). "Automated cycle time measurement and analysis of excavator's loading operation using smart phone-embedded IMU sensors." *2015 International Workshop on Computing in Civil Engineering*, 215–222.

Memarzadeh, M., Golparvar-Fard, M., and Niebles, J. C. (2013). "Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors." *Automation in Construction*, 32, 24–37.

Montaser, A., and Moselhi, O. (2012). "RFID+ for tracking earthmoving operations." *Proceedings of Construction Research Congress (CRC 2012),* West Lafayette, IN, US(2012), pp. 1011-1022..

Montaser, A., and Moselhi, O. (2014). "Truck+ for earthmoving operations." *Journal of Information Technology in Construction*, 19, 412–433.

Montaser, BakryIbrahim, AlshibaniAdel, and MoselhiOsama. (2012). "Estimating productivity of earthmoving operations using spatial technologies." *Canadian Journal of Civil Engineering*, 39, 1072-1082.

Moselhi, O., Gong, D., and El-Rayes, K. (2011). "Estimating weather impact on the duration of construction activities." *Canadian Journal of Civil Engineering*, 24, 359–366.

Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). "Beyond short snippets: Deep networks for video classification." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June, 4694–4702.

Ni, L. M., Yunhao Liu, Yiu Cho Lau, and Patil, A. P. (2003). "LANDMARC: indoor location sensing using active RFID." *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, (PerCom 2003).* 407–415.

Nicolai, W., Alex, B., Paulus, D. (2016) "Simple online and realtime tracking with a deep association metric." *2016 IEEE International Conference on Image Processing (ICIP)*, 3464–3468.

Oloufa, A.A., Ikeda, M., and Oda, H. (2002) " GPS based wireless collision detection of construction equipment." *Proceedings of 19th ISARC Gaithersburg*, Maryland, September 23–25, pp. 461–466.

Ok, S. C. and Sinha, S. K. (2006) "Constructio equipment productivity estimation using artificial neural network model." *Construction Management and Economics* 24(2006), 1029-1044.

Park, M.-W., Makhmalbaf, A., and Brilakis, I. (2011). "Comparative study of vision tracking methods for tracking of construction site resources." *Automation in Construction*, 20(7), 905–915.

Pazhoohesh, M., and Zhang, C. (2015). "Automated construction progress monitoring using thermal images and wireless sensor networks" *in: Proceedings of CSCE Annual Conference,* Regina, Canada, 2015, 957-963.

Peurifoy, R. L., Schexnayder, C. J., Shapira, A., Schmitt, R. L. (2009) "Construction Planning, Equipment, and Methods." New York, NY: McGraw-Hill, 324.

Pradhananga, N., and Teizer, J. (2013). "Automatic spatio-temporal analysis of construction site equipment operations using GPS data." *Automation in Construction*, 29, 107–122.

Quebec Wood Export Bureau. (2019). "Présentation-chantierVision2019." https://quebecwoodexport.com/assets/uploads/Pre%CC%81sentation-chantierVision2019.pdf

Rashid, K. M., and Louis, J. (2019). "Times-series data augmentation and deep learning for construction equipment activity recognition." *Advanced Engineering Informatics*, 42, 100944.

Rashid, K. M., and Louis, J. (2020). "Automated activity identification for construction equipment using motion data from articulated members." *Frontiers in Built Environment* 5 (1): 144.

Reddy, V., Sanderson, C., and Lovell, B. C. (2009). "An efficient and robust sequential algorithm for background estimation in video surveillance." *2009 16th IEEE International Conference on Image Processing (ICIP)*, 1109–1112.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You Only Look Once: Unified, Real-Time Object Detection." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Redmon, J., Farhadi, A. (2018) "YOLOv3: An incremental improvement." arXiv:1804.02767 [cs.CV] https://arxiv.org/abs/1804.02767

Roberts, D., and Golparvar-Fard, M. (2019). "End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level." *Automation in Construction*, 105, 102811.

Sabillon, C., Rashidi, A., Samanta, B., Davenport, M. A., and Anderson, D. V. (2020). "Audio-based Bayesian model for productivity estimation of cyclic construction activities." *Journal of Computing in Civil Engineering*, 34(1), 04019048.

Salem, A., Salah, A., and Moselhi, O. (2018). "Fuzzy-based configuration of automated data acquisition systems for earthmoving operations." *Journal of Information Technology in Construction*, 23, 122-137.

Santosh, K. G., Bharti, W. G., and Yannawar, P. (2010). "A Review on Speech Recognition Technique." *International Journal of Computer Applications*, 10.

Scovanner, P., Ali, S., and Shah, M. (2007). "A 3-dimensional SIFT descriptor and its application to action recognition." *Proceedings of the 15th International Conference on Multimedia*, Augsburg, Germany, September 24-29, 2007, PP 357-360.

Seresht, N., and Fayek, A. (2019). "Factors influencing multifactor productivity of equipment-intensive activities." *International Journal of Productivity and Performance Management*, Vol. 69 No. 9, pp. 2021-2045.

Shahandashti, S. M., Razavi, S. N., Soibelman, L., Berges, M., Caldas, C. H., Brilakis, I., Teizer, J., Vela, P. A., Haas, C., Garrett, J., Akinci, B., and Zhu, Z. (2011). "Data-fusion approaches and applications for construction engineering." *Journal of Construction Engineering and Management*, 137(10), 863–869.

Sherafat, B., Rashidi, A., Lee, Y.-C., and Ahn, C. R. (2019). "A hybrid kinematic-acoustic system for automated activity detection of construction equipment." *Sensors*, Multidisciplinary Digital Publishing Institute, 19(19), 4286.

Shi, Y., Tian, Y., Wang, Y., and Huang, T. (2017). "Sequential deep trajectory descriptor for action recognition with three-stream CNN." *IEEE Transactions on Multimedia*, 19(7), 1510–1520.

Simonyan, K., and Zisserman, A. (2014). "Two-stream convolutional networks for action recognition in videos." *arXiv:1406.2199 [cs.CV],* https://arxiv.org/abs/1406.2199.

Slaton, T., Hernandez, C., and Akhavian, R. (2020). "Construction activity recognition with convolutional recurrent networks." *Automation in Construction*, 113, 103138.

Smith, S. D. (1999). "Earthmoving productivity estimation using linear regression techniques." *Journal of Construction Engineering and Management*, 125(3), 133–141.

Soltani, M. M., Zhu, Z., and Hammad, A. (2017). "Skeleton estimation of excavator by detecting its parts." *Automation in Construction*, 82, 1–15.

Soltani, M. M., Zhu, Z., and Hammad, A. (2018). "Framework for Location Data Fusion and Pose Estimation of Excavators Using Stereo Vision." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 32(6), 04018045.

Song, L., and Eldin, N. N. (2012). "Adaptive real-time tracking and simulation of heavy construction operations for look-ahead scheduling." *Automation in Construction*, 27, 32–39.

Song, S., Marks, E., and Pradhananga, N. (2017). "Impact variables of dump truck cycle time for heavy excavation construction projects." *Journal of Construction Engineering and Project Management*, 7, 11–18.

Soomro, K., Zamir, A. R., and Shah, M. (2012). "UCF101: A dataset of 101 human actions classes from videos in the wild." *arXiv:1212.0402 [cs.CV]*. https://arxiv.org/abs/1212.0402.

Statista. (2018). U.S. Construction Industry - Statistics & Facts. Retrived Febrary 25th 2019,

from https://www.statista.com/topics/974/construction/.

Statistics Canada (2011). Canadia Year Book. Retrive Febrary 27th 2019, from https://www150.statcan.gc.ca/n1/pub/11-402-x/2011000/chap/construction/construction-eng.htm.

Statistics Canada (2016). Canadia Year Book. Retrive Febrary 27th 2019, from https://www.guichetemplois.gc.ca/content_pieces-eng.do?cid=11247.

Statistics Canada. (2018). Capital and repair expenditures, non-residential tangible assets, by industry and geography. Retrived Febrary 25th 2019, from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3410003501.

Tajeen, H., and Zhu, Z. (2014). "Image dataset development for measuring construction equipment recognition performance." *Automation in Construction*, 48, 1–10.

Tam, C. M., Tong, T. K. L., and Tse, S. L. (2002). "Artificial neural networks model for predicting excavator productivity." *Engineering Construction and Architectural Management*, 9(5–6), 446–452.

Teizer, J. (2015). "Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites." *Advanced Engineering Informatics*, Infrastructure Computer Vision, 29(2), 225–238.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3D convolutional networks." *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter, 4489–4497.

Tran, D., Ray, J., Shou, Z., Chang, S.-F., and Paluri, M. (2017). "ConvNet Architecture Search for Spatiotemporal Feature Learning." *arXiv:1708.05038 [cs.CV],* https://arxiv.org/abs/1708.05038.

Vahdatikhaki, F., and Hammad, A. (2014). "Framework for near real-time simulation of earthmoving projects using location tracking technologies." *Automation in Construction*, 42, 50–67.

Vahdatikhaki, F., Hammad, A., and Siddiqui, H. (2015). "Optimization-based excavator pose estimation using real-time location systems." *Automation in Construction*, 56, 76–92.

Varol, G., Laptev, I., and Schmid, C. (2016). "Long-term Temporal Convolutions for Action Recognition." *arXiv:1604.04494 [cs.CV]*. https://arxiv.org/abs/1604.04494.

Varol, G., Laptev, I., and Schmid, C. (2018). "Long-term temporal convolutions for action recognition" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), 1510–1517.

Vasenev, A., Hartmann, T., and Dorée, A. G. (2014). "A distributed data collection and management framework for tracking construction operations." *Advanced Engineering Informatics*, 28(2), 127–137.

Wang, H., Kläser, A., Schmid, C., and Liu, C. L. (2013). "Dense trajectories and motion boundary descriptors for action recognition." *International journal of Computer Vision,* 103, 60–79.

Wang, H., and Schmid, C. (2013). " Action recognition with improved trajectories." *ICCV-IEEE International Conference on Computer Vision*, Dec 2013, Sydney, Australia. IEEE, pp.3551-3558,

Wang, L., Qiao, Y., and Tang, X. (2015a). "Action recognition with trajectory-pooled deep-convolutional descriptors." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June, 4305–4314.

Wang, L., Xiong, Y., Wang, Z., and Qiao, Y. (2015b). "Towards good practices for very deep two-stream convnets." *arXiv:1507.02159 [cs.CV]*. https://arxiv.org/abs/1507.02159.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and van Gool, L. (2016). "Temporal segment networks: Towards good practices for deep action recognition." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9912 LNCS, 20–36.

Willems, G., Tuytelaars, T., and An, L. G. (2008). " Efficient Dense and Scale-Invariant Spatio-Temporal." *European Conference on Computer Vision (ECCV),* 2008. pp. 650–663.

Wojke, N., Bewley, A., and Paulus, D. (2017). "Simple online and realtime tracking with a deep association metric." *2017 IEEE International Conference on Image Processing (ICIP),* pp. 3645-2649.

Wu, S., Zhong, S., and Liu, Y. (2017). "Deep residual learning for image steganalysis." *Multimedia Tools and Applications*, 77, 10437–10453.

Xiao, B., and Zhu, Z. (2018). "Two-dimensional visual tracking in construction scenarios: A comparative study." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 32(3), 04018006.

Yang, J., Edwards, D. J., and Love, P. E. D. (2003). "A computational intelligent fuzzy model approach for excavator cycle time simulation." *Automation in Construction*, Design e-ducation: Connecting the Real and the Virtual, 12(6), 725–735.

Yang, J., Shi, Z., and Wu, Z. (2016a). "Vision-based action recognition of construction workers using dense trajectories." *Advanced Engineering Informatics*, Elsevier Ltd, 30(3), 327–336.

Yang, J., Shi, Z., and Wu, Z. (2016b). "Vision-based action recognition of construction workers using dense trajectories." *Advanced Engineering Informatics*, 30(3), 327–336.

Yoon, J., Kim, J., Seo, J., and Suh, S. (2014). "Spatial factors affecting the loading efficiency of excavators." *Automation in Construction*, 48, 97–106.

Yuan, C., Li, S., and Cai, H. (2017). "Vision-based excavator detection and tracking using hybrid kinematic shapes and key nodes." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 31(1), 04016038.

Zang, J., Wang, L., Liu, Z., Zhang, Q., Niu, Z., Hua, G. and Zheng, (2018). "Attention-based temporal weighted convolutional neural network for action recognition." *arXiv:1803.07179 [cs.CV],* https://arxiv.org/abs/1803.07179.

Zhang, H. (2008) "Multi-objective simulation-optimization for earthmoving operations." *Automation in Construction*, 18, 79–86.

Zhang, B., Wang, L., Wang, Z., Qiao, Y., and Wang, H. (2016). "Real-time action recognition with enhanced motion vector CNNs." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2718-2726

Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., and Li., Z. (2017). "A review on human activity recognition using vision-based method." *Journal of Healthcare Engineering*, 2017, 31.

Zhu, Y., Lan, Z., Newsam, S., and Hauptmann, A. G. (2017). "Hidden Two-Stream Convolutional Networks for Action Recognition." 1–12. *arXiv:1704.00389 [cs.CV]*. https://arxiv.org/abs/1704.00389

Zhu, Z., Ren, X., and Chen, Z. (2016). "Visual tracking of construction jobsite workforce and equipment with particle filtering." *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, 30(6), 04016023.

Zolfaghari, M., Oliveira, G. L., Sedaghat, N., and Brox, T. (2017). "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection."

Proceedings of the IEEE International Conference on Computer Vision, 2017-Octob, 2923–2932.

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*, American Society of Civil Engineers, 21(4), 238–246.

APPENDICES

Appendix A. Python Code of Excavator's Activity Recognition

```
import torch
from torch.autograd import Variable
import time
import os
import sys
import json
from utils import AverageMeter

def calculate_video_results(output_buffer, video_id, test_results, class_names):
    video_outputs = torch.stack(output_buffer)
    average_scores = torch.mean(video_outputs, dim=0)
    sorted_scores, locs = torch.topk(average_scores, k=10)
    video_results = []
    for i in range(sorted_scores.size(0)):
        video_results.append({'label': class_names[locs[i]], 'score': sorted_scores[i]})
    test_results['results'][video_id] = video_results
def test(data_loader, model, opt, class_names):
    print('test')
    model.eval()
    batch_time = AverageMeter()
    data_time = AverageMeter()
    end_time = time.time()
    output_buffer = []
    previous_video_id = ''
    test_results = {'results': {}}
    for i, (inputs, targets) in enumerate(data_loader):
        data_time.update(time.time() - end_time)
        inputs = Variable(inputs, volatile=True)
        outputs = model(inputs)
        for j in range(outputs.size(0)):
            if not (i == 0 and j == 0) and targets[j] != previous_video_id:
                calculate_video_results(output_buffer, previous_video_id,   test_results, class_names)
                output_buffer = []
            output_buffer.append(outputs[j].data.cpu())
            previous_video_id = targets[j]
        if (i % 100) == 0:
            with open(os.path.join(opt.result_path, '{}.json'.format(opt.test_subset)), 'w') as f:
                json.dump(test_results, f)
        batch_time.update(time.time() - end_time)
        end_time = time.time()
        print('[{}/{}]\t'
              'Time {batch_time.val:.3f} ({batch_time.avg:.3f})\t'
              'Data {data_time.val:.3f} ({data_time.avg:.3f})\t'.format(
                  i + 1, len(data_loader), batch_time=batch_time, data_time=data_time))
    with open(os.path.join(opt.result_path,
                           '{}.json'.format(opt.test_subset)),
              'w') as f:
        json.dump(test_results, f)
```

Appendix B. Python Code of Excavator's Productivity Analysis

```
import os
import sys
import json
import subprocess
import numpy as np
from PIL import Image, ImageDraw, ImageFont
# retrive the information of the video
def get_fps(video_file_path, frames_directory_path):
    p = subprocess.Popen('ffprobe {}'.format(video_file_path),
                         shell=True, stdout=subprocess.PIPE, stderr=subprocess.PIPE)
    _, res = p.communicate()
    res = res.decode('utf-8')
    duration_index = res.find('Duration:')
    duration_str = res[(duration_index + 10):(duration_index + 21)]
    hour = float(duration_str[0:2])
    minute = float(duration_str[3:5])
    sec = float(duration_str[6:10])
    total_sec = hour * 3600 + minute * 60 + sec
    n_frames = len(os.listdir(frames_directory_path))
    fps = round(n_frames / total_sec, 2)
    return fps
#load the activity recognition results
if __name__ == '__main__':
    result_json_path = sys.argv[1]
    video_root_path = sys.argv[2]
    dst_directory_path = sys.argv[3]
    if not os.path.exists(dst_directory_path):
        subprocess.call('mkdir -p {}'.format(dst_directory_path), shell=True)
    class_name_path = sys.argv[4]
    temporal_unit = int(sys.argv[5])
    with open(result_json_path, 'r') as f:
        results = json.load(f)
    with open(class_name_path, 'r') as f:
        class_names = []
        for row in f:
            class_names.append(row[:-1])
    for index in range(len(results)):
        video_path = os.path.join(video_root_path, results[index]['video'])
        clips = results[index]['clips']
        unit_classes = []
        if temporal_unit == 0:
            unit = len(clips)
        else:
            unit = temporal_unit
        for i in range(0, len(clips), unit):
            n_elements = min(unit, len(clips) - i)
            scores = np.array(clips[i]['scores'])
            for j in range(i, min(i + unit, len(clips))):
                scores = np.array(clips[i]['scores'])
            unit_classes.append(class_names[np.argmax(scores)])
        new_segments = []
        major_list = []
        for i in range(0, len(unit_classes), 4):
```

```
            elements = min(4, len(unit_classes) - i)
            major = 0
            count = 0
            for j in range(i, min(i + 4, len(unit_classes))):
                if count == 0:
                    major = unit_classes[j]
                    count = count + 1
                elif major == unit_classes[j]:
                    count = count + 1
                else:
                    count = count - 1
            major_list.append(major)
            new_segments.append([clips[i]['segment'][0], clips[i+elements-1]['segment'][1]])
        print(new_segments)
        print(major_list)
        new_acti_class = []
        for i in range(len(major_list)):
            for j in range(new_segments[i][0], new_segments[i][1]+1):
                new_acti_class.append(major_list[i][2:])
        subprocess.call('ffmpeg -i {} tmp/image_%05d.jpg'.format(video_path), shell=True)
        fps = get_fps(video_path, 'tmp')
        per_duration = round(1/fps,2)
    #calculate cycle times
        dict_all = {"Digging": [0, 0], "Loading": [0, 0], "Swing": [0, 0]}
        dict_all.setdefault('{}'.format(new_acti_class[0], []))[0] = 1
        for i in range(len(new_acti_class) - 1):
            if new_acti_class[i] == new_acti_class[i + 1]:
                dict_all.setdefault('{}'.format(new_acti_class[i], []))[1] = \
                dict_all.setdefault('{}'.format(new_acti_class[i], []))[1] + 1
            else:
                dict_all.setdefault('{}'.format(new_acti_class[i], []))[1] = \
                dict_all.setdefault('{}'.format(new_acti_class[i], []))[1] + 1
                dict_all.setdefault('{}'.format(new_acti_class[i + 1]), [])[0] = \
                dict_all.setdefault('{}'.format(new_acti_class[i + 1]), [])[0] + 1
        filename = './writetotxt.txt'
        dig_count=[]
        with open(filename, 'w') as f:
            for i in range(len(new_acti_class) - 1):
                if new_acti_class[i] != new_acti_class[i + 1]:
                    a = new_acti_class[i] + str(i + 1)
                    dig_count.append(a)
                    f.write("{} {} \n".format(i + 1, new_acti_class[i]))
            f.write("{} {} \n".format(len(new_acti_class) - 1, new_acti_class[len(new_acti_class) - 1]))
            dig_count.append(new_acti_class[len(new_acti_class)-1] + str(len(new_acti_class)))
        file_name='./writecycle'
        c=0
        cycle_time =[0]
        with open (file_name, 'w') as f:
            for i in range(1, len(dig_count)):
                if dig_count[i][:7] =='Digging':
                    c=c+1
                    cycle_time.append(dig_count[i-1][5:])
                    _dur = int(cycle_time[c])-int(cycle_time[c-1])
                    f.write("cycle number: {} cycle time: {} \n".format(c, str(_dur*0.04)))
            cyc = c
        print(cycle_time)
```

```
    #calculate activity time
        num_loa = 0
        num_swi = 0
        num_dig = 0
        dig_list = []
        swi_list = []
        loa_list = []
        total_duration = []
        for i in range(len(new_acti_class)):
            total_duration.append(round((i + 1) * per_duration, 2))
            time_dig = round(num_dig * per_duration, 2)
            time_swi = round(num_swi * per_duration, 2)
            time_loa = round(num_loa * per_duration, 2)
            if new_acti_class[i] == 'Digging':
                num_dig = 1 + num_dig
            else:
                if new_acti_class[i] == 'Swing':
                    num_swi = 1+num_swi
                else:
                    num_loa = 1 + num_loa
            dig_list.append(time_dig)
            swi_list.append(time_swi)
            loa_list.append(time_loa)
#calculate productivity and print on the frames
        productivity = round((cyc * 60 * 1.74) / 1.1, 2)
        for i in range(len(new_acti_class)):
            image = Image.open('tmp/image_{:05}.jpg'.format(i + 1)).convert('RGBA')
            lab_img = Image.new('RGBA', image.size)
            font = ImageFont.truetype(os.path.join(os.path.dirname(__file__), 'SourceSansPro-Regular.ttf'), 25)
            d = ImageDraw.Draw(lab_img)
            textsize = d.textsize(new_acti_class[i], font=font)
            d.text((10, 10), new_acti_class[i], font=font, fill=(0, 255, 255))
            d.text((10, 40), 'Digging:' + str(dig_list[i]) + 's', font=font, fill=(255, 0, 255))
            d.line([(150, 60), (150 + dig_list[i], 60)], fill=(255, 0, 255), width=10)
            d.text((10, 70), 'Swing:' + str(swi_list[i]) + 's', font=font, fill=(255, 0, 255))
            d.line([(150, 90), (150 + swi_list[i], 90)], fill=(255, 0, 255), width=10)
            d.text((10, 100), 'Loading:' + str(loa_list[i]) + 's', font=font, fill=(255, 0, 255))
            d.line([(150, 120), (150 + loa_list[i], 120)], fill=(255, 0, 255), width=10)
            if i == len(new_acti_class) - 1:
                d.text((250, 40), 'Total time:' + str(total_duration[-1]) + 's', font=font, fill=(0, 255, 255))
                d.text((250,70), 'Cycle:' + str(cyc), font=font, fill=(0,255,255))
                d.text((250,100), 'Productivity:' + str(productivity)+' LCY/h', font=font, fill=(0,255,255))
            image.paste(lab_img, (10, 10), lab_img)
            image = image.convert("RGB")
            image.save('tmp/image_{:05}_pred.jpg'.format(i + 1))
        dst_file_path = os.path.join(dst_directory_path, video_path.split('/')[-1])
        subprocess.call('ffmpeg -y -r {} -i tmp/image_%05d_pred.jpg -b:v 1000k {}'.format(fps, dst_file_path),
                        shell=True)
```

Appendix C. Python Code of Excavator and Truck Detection and Tracking

```
from __future__ import division, print_function, absolute_import
import argparse
import os
from timeit import time
import warnings
import sys
import cv2
import numpy as np
from PIL import Image
from yolo import YOLO
from deep_sort import preprocessing
from deep_sort import nn_matching
from deep_sort.detection import Detection
from deep_sort.tracker import Tracker
from tools import generate_detections as gdet
from deep_sort.detection import Detection as ddet
warnings.filterwarnings('ignore')

def main(yolo,read_type):
  # Definition of the parameters
   max_cosine_distance = 0.3
   nn_budget = None
   nms_max_overlap = 1.0
  # deep_sort
   model_filename = 'model_data/mars-small128.pb'
   encoder = gdet.create_box_encoder(model_filename,batch_size=1)
   metric = nn_matching.NearestNeighborDistanceMetric("cosine", max_cosine_distance, nn_budget)
   tracker = Tracker(metric)
   #geneate a video object
   video_dir="E:\\Chen\\deep_sort_yolov3-master\\short_out.mp4"
   output_path = 'E:\\Chen\\deep_sort_yolov3-master\\GH010016_out.mp4'
   video=video_open(read_type,video_dir)
   video_capture = video.generate_video()
   video_FourCC    = int(video_capture.get(cv2.CAP_PROP_FOURCC))
   video_fps = int(video_capture.get(cv2.CAP_PROP_FPS))
   video_size = (int(video_capture.get(cv2.CAP_PROP_FRAME_WIDTH)),
                 int(video_capture.get(cv2.CAP_PROP_FRAME_HEIGHT)))
   out = cv2. VideoWriter(output_path, video_FourCC, video_fps, video_size)
   fps=0
   while True:
      ret, frame = video_capture.read()  # frame shape 640*480*3
#       out = cv2.VideoWriter('output.mp4', -1, 20.0, (640,480))
      if ret != True:
         break;
      t1 = time.time()
      image = Image.fromarray(frame)
      time3=time.time()
      boxs = yolo.detect_image(image)
      time4=time.time()
      print('detect cost is',time4-time3)
     # print("box_num",len(boxs))
      time3=time.time()
      features = encoder(frame,boxs)
```

```
        # score to 1.0 here.
        detections = [Detection(bbox, 1.0, feature) for bbox, feature in zip(boxs, features)
        # Run non-maxima suppression.
        boxes = np.array([d.tlwh for d in detections])
        scores = np.array([d.confidence for d in detections])
        indices = preprocessing.non_max_suppression(boxes, nms_max_overlap, scores)
        detections = [detections[i] for i in indices]
        time4=time.time()
        print('features extract is',time4-time3)
        # Call the tracker
        tracker.predict()
        tracker.update(detections)
        for track in tracker.tracks:
            if track.is_confirmed() and track.time_since_update >1 :
                continue
            bbox = track.to_tlbr()
            cv2.rectangle(frame, (int(bbox[0]), int(bbox[1])), (int(bbox[2]), int(bbox[3])),(255,255,255), 2)
            cv2.putText(frame, yolo.class_names[1]+str(track.track_id),(int(bbox[0]), int(bbox[1])),0, 5e-3 * 200,
                        (255,255,255),2)
        for det in detections:
            bbox = det.to_tlbr()
            cv2.rectangle(frame,(int(bbox[0]), int(bbox[1])), (int(bbox[2]), int(bbox[3])),(255,0,0), 2)
        cv2.imshow('', frame)
        out.write(frame)
        fps  = ( fps + (1./(time.time()-t1)) ) / 2
        print("fps= %f"%(fps))
        if cv2.waitKey(1) & 0xFF == ord('q'):
            break
    video_capture.release()
    cv2.destroyAllWindows()
class video_open:
    def __init__(self,read_type,video_dir):
        #self.readtype=read_type
        if read_type=='video':
            self.readtype=0
        else:
            self.readtype=video_dir
    def generate_video(self):
        video_capture=cv2.VideoCapture(self.readtype)
        return video_capture
def parse_args():
    parser = argparse.ArgumentParser(description="Deep SORT")
    parser.add_argument(
        "--read_type", help="camera or video",
        default='camera', required=False)
    return parser.parse_args()
if __name__ == '__main__':
    args = parse_args()
    main(YOLO(),args.read_type)
```

Appendix D. Python Code of Idling State Identification

```
import codecs
import numpy as np
import xlwt

book = xlwt.Workbook(encoding='ucf-8')
sheet3 = book.add_sheet(u'Sheet3', cell_overwrite_ok=True)
set_style = xlwt.easyxf('font: name Arial, bold True, height 200;')
sheet3.col(0).width = 256*20
sheet3.col(1).width = 256*20

f = codecs.open('../MOTformat1.txt', mode='r', encoding = 'utf-8') # read coordinates of the equipment
line = f.readline()
line.rstrip('(')
x=[]
y=[]
while line:
    a = line.split(",")
    b1 = a[2].replace('(',")
    b2 = a[5].replace(')',")
    xmin = int(b1)+int(a[4])/2
    ymin = int(a[3])+int(b2)/2
    x.append(xmin)
    y.append(ymin)
    line = f.readline()
f.close()
dl = []
for i in range(0,len(x)-1):
    d = np.sqrt(pow(abs(y[i+1]-y[i]),2)+pow(abs(x[i+1]-x[i]),2)) #calculate the distance changes in consecutives
frames
    dl.append(d)
    i=i+1
std=[]
for i in range(1,len(dl)-100):
    dd = abs(round((dl[i+99]-dl[i-1]),3))
    std.append(dd)
xstd=[]
xarr = np.array(std)
for i in range(len(xarr-100)):
    astd = np.std(xarr[i:i+99], ddof=0) # calculate the standard deviation of distance changes
    xstd.append(astd)
    i=i+1
dd1=0
for item in dl:
    sheet3.write(dd1, 3, dl[dd1], set_style)
    dd1+=1
c1 = 0
for item in xstd:
     sheet3.write(c1, 1, xstd[c1], set_style)
     c1+=1
book.save('distance11.xls')
```

Appendix E. Python Code of Transfer 2D Coordinates to 3D Coordinates

```
import math
import numpy as np


class Perspective_transformation:
    def __init__(self):
        self.rotation_inverse = np.array([[0.9958, 0.087525, 0.027392937], [-0.008716768, -0.20712564, 0.978235441],
                                          [0.091272854, -0.974368931, -0.205555728]], dtype=np.float32)

        self.cam_inverse = np.array([[0.001190775, 0, -0.826559976], [0, 0.001045513, 0.270372976], [0, 0, 1]],
                                    dtype=np.float32)

        self.scale = 2989.978231

        self.transition_vector = np.array([[-778.2842947], [2923.387541], [2634.614225]], dtype=np.float32)

    def transfer_2d_to_3d(self, u, v):
        pixel_coordinates = np.array([[u*2], [v*2], [1]], dtype=np.float32)

        cam_times_pixel = np.matmul(self.cam_inverse, pixel_coordinates)
        cam_times_pixel_t = np.subtract(cam_times_pixel, self.transition_vector)
        scale_in = self.scale * cam_times_pixel_t
        self.point = np.matmul(self.rotation_inverse, scale_in)
        x_coordinate = (self.point[0], self.point[1])[0][0]
        y_coordinate = (self.point[0], self.point[1])[1][0]
        return x_coordinate, y_coordinate

    def get_real_distance(self, excavator_loc, truck_loc):
        real_distance = math.sqrt(((excavator_loc[0] - truck_loc[0]) ** 2) + ((excavator_loc[1] - truck_loc[1]) ** 2))
        print(real_distance)
        return float("%.2f" % real_distance) / 100
```

Appendix F. Python Code of Workgroup Identification

```
import json
import xlwt
import numpy as np
import cv2

#This script is to group trucks with excavator
book = xlwt.Workbook(encoding='ucf-8')
sheet1 = book.add_sheet(u'Sheet1', cell_overwrite_ok=True)
sheet3 = book.add_sheet(u'Sheet3', cell_overwrite_ok=True)
set_style = xlwt.easyxf('font: name Arial, bold True, height 200;')
sheet1.col(0).width = 256*20
sheet1.col(1).width = 256*20
sheet3.col(0).width = 256*20
sheet3.col(1).width = 256*20

exc_n =[]
tru_n =[]
with open ('e3.txt', 'r', encoding='utf-8') as f:
    fn=1
    for line in f:
        value = line[:-1]
        x = value.replace("'", '"')
        d1=json.loads(x)
        exc = d1['excavator']
        tru = d1['truck']

        centroid = []
        for item in tru:
            xt = int((item[0]+item[2])/2) #centroid of truck (yt,yx)
            yt = int((item[1]+item[3])/2)
            ex = int(exc[0][0])#centroid of exc (ex,ey)
            ey = int(exc[0][1])
            ts =0.5*(exc[0][2]+item[2]) # define threshold
            d = np.sqrt(pow((ex-yt), 2)+pow((ey-xt), 2)) # distance between excavator and truck
            if d < ts:
                centroid.append((yt,xt))
            if len(centroid) != 0:
                point = np.array(centroid)
                img = cv2.imread("/Users/chen/PycharmProjects/DeepSort/deep_sort/demo/img2/img1
              /{}.jpg".format(str(fn).zfill(5)))
                for item in centroid:
                    line = cv2.line(img, item, (ex, ey), (0, 255, 255), 2)
                cv2.imwrite("/Users/chen/PycharmProjects/DeepSort/deep_sort/demo/img3/{}.jpg".for
                            mat(str(fn).zfill(5)), line) # draw lines between grouped equipment
        fn=fn+1
        if fn == 141000:
            break

        print(fn)
```

Appendix G. Python Code of Idling Reasons Identification

```
import xlwt

book = xlwt.Workbook(encoding='ucf-8')
sheet1 = book.add_sheet(u'Sheet1', cell_overwrite_ok=True)
sheet3 = book.add_sheet(u'Sheet3', cell_overwrite_ok=True)
set_style = xlwt.easyxf('font: name Arial, bold True, height 200;')
sheet1.col(0).width = 256*20
sheet1.col(1).width = 256*20
sheet3.col(0).width = 256*20
sheet3.col(1).width = 256*20

file=open('C3det1.txt') # open the file of equipment's activities
dataMat=[]
labelMat=[]
fn=1
for line in file.readlines():
    curLine=line.strip().split("\t")
    curLine = [x for x in curLine if x !="]
# the following script is to identify the idling reasons of excavators
    if curLine[0] == 'Moving': # curLine[0] is the activity of the excavator
        fn=fn+1
        continue
    if len(curLine) >= 4:
        c = 'Case4'
    elif len(curLine) == 3:
        if curLine[1]==curLine[2]=='Idling':
            c = 'Case3'
        else:
            c = 'Case2'
    elif len(curLine) == 2:
        if curLine[1] == 'Idling':
            c = 'Case3'
        else:
            c = 'Case2'
    elif len(curLine) == 1:
            c = "Case 1"
    f = open('ttest.txt','a')
    f.write('\n{},{}'.format(fn,c))
    fn=fn+1
    if fn == 3566:
        break

# the following script is to identify the idling reasons of trucks
    if curLine[0] == 'Idling':
        fn=fn+1
        continue
    if len(curLine) <= 2:
        c = 'Working'
    elif len(curLine) ==3:
        if curLine[1]==curLine[2]=='Idling':
            c = "Case 5"
        else:
            c = "Working"
```

Appendix H. List of Publications

Journal Papers

Chen, C., Zhu, Z., and Hammad, A. (2021). "Critical review and road map of automated methods for earthmoving equipment productivity monitoring." *Automation in Construction* (Under Review).

Chen, C., Zhu, Z., and Hammad, A. (2020). "Automatic identification of idling reasons in earthwork operations based on excavator-truck relationships." *Computing in Civil Engineering*, (Under Review).

Chen, C., Zhu, Z., and Hammad, A. (2020). "Automated excavators activity recognition and productivity analysis from construction site surveillance videos." *Automation in Construction*, 110, 103045.

Conference Papers

Chen, C., Zhu, Z., and Hammad, A. (2020). "Automatic analysis of idling in excavator's operations based on excavator-truck relationships." *ISARC Proceedings*, IAARC, 1307–1313.

Chen, C., Zhu, Z., and Hammad, A. (2019) "Vision-based Excavator Activity Recognition and Productivity Analysis in Construction." *ASCE International Conference on Computing in Civil Engineering*, 241-248.