Two Dimensional Visual Tracking in Construction Scenarios

Bo Xiao

A Thesis in

The Department of

Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Master of Applied Sciences (in Building Engineering) at

Concordia University

Montreal, Quebec, Canada

May 2017

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Bo Xiao

Entitled: Two Dimensional Visual Tracking in Construction Scenarios

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

Master of Applied Science in Building Engineering

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

Signed by the final Examining Committee:

______________________________ Chair
Dr. Ciprian Alecsandru

______________________________ Examiner
Dr. Sang Hyeok Han

______________________________ Examiner
Dr. Ciprian Alecsandru

______________________________ Examiner (External)
Dr. Amin Hammad

______________________________ Supervisor
Dr. Zhenhua Zhu

Approved by ____
Chair of Department

____
Dean of Faculty

Date ____

ABSTRACT

Two Dimensional Visual Tracking in Construction Scenarios

The tracking of construction resources (e.g. workforce and equipment) in videos, i.e., two dimensional (2D) visual tracking, has gained significant interests in the construction industries. There exist lots of research studies that relied on 2D visual tracking methods to support the surveillance of construction productivity, safety, and project progress. However, few efforts have been put on evaluating the accuracy and robustness of these tracking methods in the construction scenarios. Meanwhile, it is noticed that state-of-art tracking methods have not shown reliable performance in tracking articulated equipment, such as excavators, backhoes, and dozers etc.

The main objective of this research is to fill these knowledge gaps. First, a total of fifth (15) 2D visual tracking methods were selected here due to their excellent performances identified in the computer vision field. Then, the methods were tested with twenty (20) videos captured from multiple construction jobsites at day and night. The videos contain construction resources, including but not limited to excavators, backhoes, and compactors. Also, they were characterized with the attributes, such as occlusions, scale variation, and background clutter, in order to provide a comprehensive evaluation. The tracking results were evaluated with the sequence overlap score, center error ratio, and tracking length ratio respectively. According to the quantitative comparison of tracking methods, two improvements were further conducted. One is to fuse the tracking results of individual tracking methods based on the non-maximum suppression. The other is to track the articulated equipment by proposing the idea of tracking the equipment parts respectively.

The test results from this research study indicated that 1) the methods built on the local sparse representation were more effective; 2) the generative tracking strategy typically outperformed the discriminative one, when being adopted to track the equipment and workforce in the construction scenarios; 3) the fusion of the results from different tracking methods increased the tracking performance by 10% in accuracy; and 4) the part-based tracking methods improved the tracking performance in both accuracy and robustness, when being used to track the articulated equipment.

ACKNOWLEDGMENTS

Words cannot express the respect and appreciation I feel towards my family, supervisors, and friends. You resembled the kind-hearted father to me, the strict architecture to reach perfection and the willing friend to cheer me up. It was a pleasure working my thesis with you and sharing with your non-everlasting experience.

Special thanks to my Parents and the faithful friends who stood by my side. You have been exceptionally helpful and extraordinary throughout my journey.

My precious gratitude is addressed to my supervisor Dr. Zhenhua Zhu for his endless support, precious advice, and wisdom. I wouldn’t be here without his supervision. In addition, I would like to thank my examiners, Dr. Amin Hammad, Dr. Ciprian Alecsandru, and Dr. Sang Hyeok Han for their time, advice and effort in reviewing my thesis.

Lastly, I would like to show my gratitude to the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

2D	Two Dimensional
OTB	Object Tracking Benchmark
ALOV	Amsterdam Library of Ordinary Videos
OCC	Occlusion
IV	Illumination Variation
MB	Motion Blur
BC	Background Clutter
SV	Scale Variation
OS	Overlap Score
AOS	Average Overlap Score
CE	Center Location Error
TL	Tracking Length
CER	Center Location Error Ratio
NMS	Non-maximum Suppression
CCTV	Closed-circuit Television
CAT	Computerized Axial Tomography
MRI	Magnetic Resonance Imaging
HOG	Histograms of Oriented Gradients
SVM	Support Vector Machine
RFID	Radio-Frequency Identification
CNN	Convolutional Neural Network

CHAPTER 1: INTRODUCTION

1.1. Backgrounds

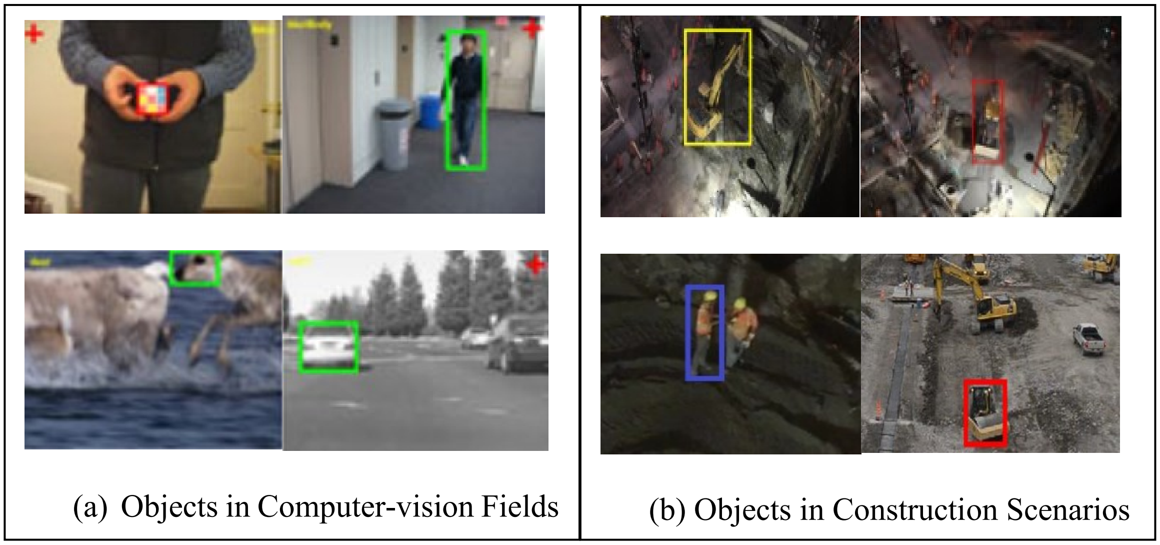
Visual Tracking is one of the most attractive research fields in modern industry and has different practical applications, in terms of human-computer interaction (HCI), surveillance, and medical imaging. Adopting visual tracking for human movements, such as manipulation, gestures, and even exercise, is the fundamental of HCI (Heckenberg 2006), which is performed to provide the scientific understanding of the interaction between humans and the computer technology. For example, tracking human fingers with webcam could be used to automatically execute specific commands on the computer. Also, visual tracking could be used to observe people from a distance by means of electronic equipment, such as Closed-circuit television (CCTV) cameras, for the purpose of managing, directing, or protecting (Cannons 2008). Apart from this, monitoring the position of a medical instrument could combine with Computerized Axial Tomography (PET) and Magnetic Resonance Imaging (MRI), in order to provide assistance during surgery.

Recently, the use of visual tracking in construction has been promoted to facilitate construction automation. Vision-based tracking was applied to detect the pothole in pavement assessment (Koch et al. 2012), recognize dirt loading cycles during excavation (Azar and McCabe 2012), identify construction cumulative trauma disorders (Rempel et al. 1992), and manage equipment and workers in real-time (Weerasinghe and Ruwanpura 2009). Another important usage of visual tracking in construction is safety monitoring. It is feasible to locate workers and equipment in order to protect workers from potential collisions (Han and Lee 2013). In addition, visual tracking technologies also helped to improve the safety of workers when they were working at heights (Auvinet et al. 2013).

Thanks to the development of computer science, the tracking methods have progressed rapidly in recent years. Lucas and Kanade (1981) firstly adopted holistic templates in tracking. In order to seek better templates, many visual features have been used recently, such as histograms of oriented gradients (HOG) (Dalal and Triggs 2005), Haar-like features (Viola and Jones 2004) and co-variance region descriptor (Tuzel et al. 2006). Then, the sparse-representation-based methods (Mei and Ling 2009) were proposed and have been improved which showed the high performance in tracking blurry objects. Meanwhile, the context information (Zhang et al. 2014) were widely employed in this domain and it has achieved significant performance than traditional tracking methods in dealing with occlusions. Thanks to the exploitation of machine learning, several effective and completed methods were proposed and applied in visual tracking, such as multiple-instance learning (Babenko et al. 2011), Gaussian process regression (Gao et al. 2014), and structures output SVM (SO-SVM) (Hare et al. 2011). Recently, deep learning methods adopted a deep graph of multiple processing layers to abstract and model data in high level (Wang et al. 2013).

1.2. Problem Statement and Motivation

Although several benchmarks were conducted to evaluate the performance of existing 2D tracking methods, few efforts have been put on evaluating the accuracy and robustness of these tracking methods in the construction scenarios. Most of existing benchmark studies have used sequences captured from general scenarios and these sequences have much differences from construction sequences. As an example, the Figure 1.1 displays the different sequences tested in computer vision fields and construction fields. The common tracking objects in computer vision community include: person, animals, cars and toys, while the excavators, backhoes, compactors and workers are frequently tracked in construction scenarios. Comparing with computer-vison community, it is noticed the sequences captured from construction sites have larger illumination changes, more complex contents and similar backgrounds. These differences result in that the well performed tracking methods may not have reliable performance when tracking construction targets.



(a) Objects in Computer-vision Fields

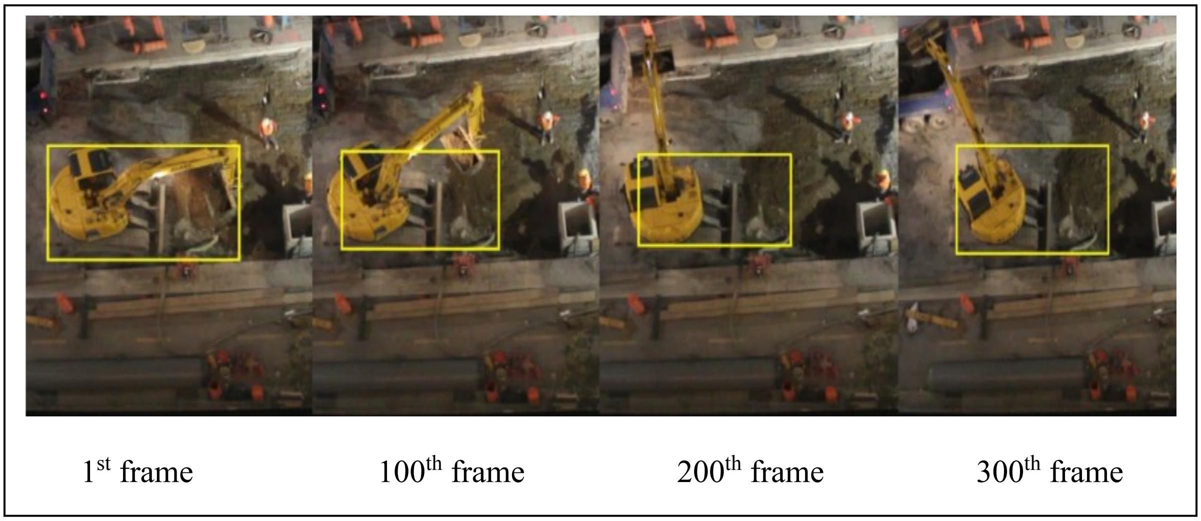
(b) Objects in Construction Scenarios

Figure 1.1: Tracking Objects from Computer-vision (Wu et al. 2015) and Construction Scenarios

It is necessary to conduct a comparative study of tracking methods in construction scenarios in order to help researchers, which adopted vision-based tracking in construction management, understand the weaknesses and strengths of state-of-art methods. At the meantime, an effective comparison study could enhance the successes by selecting the proper methods when tracking different construction targets. Park et al. (2011) presented a comparison of 2D tracking methods for construction resources. But the novel tracking methods which adopted different schemes and achieved promising performance were not considered. It is important to update existing comparative studies with novel tracking methods.

Moreover, it is common in visual tracking field that an average good tracking methods perform unstable in different sequences or scenarios. And an average worse tracking methods may perform very well in specific sequences. For example, in the *lemming* sequence of OTB benchmark (Wu et al. 2013), the overall second worst method SMS showed the best performance in this sequence. It lacks a generic fusion methods to combine arbitrary tracking results into one result and build a stronger.

On the other hand, the articulated equipment, such as excavators, backhoes, and dozers, have been widely employed for digging foundations, drilling piles, and handling materials in earthmoving works. The tracking of excavators is time-saving, cost-assuming, and convenient in calculating excavation productivities, such as dirt-loading cycles, excavation capability, and working time. But the state-of-art methods have not shown reliable performance in dealing with these articulated equipment, especially in dirt-loading operations. The frequent rotation and movement of bucket make successful tracking difficult. The Figure 1.2 shows the tracking results of an excavator, which is doing dirt-loading activity. This figure indicates that the tracking method can predict excavator's position correctly in the beginning and then it fails from the 200th frame when the bucket moved in high speed.



1st frame 100th frame 200th frame 300th frame

Figure 1.2: Example of tracking an excavator

1.3. Research Objective and Scope

The main objective of this research include:

1) Proposing a comparative study to evaluate current tracking methods under construction scenarios, which contain construction resources, including but not limited to excavators, backhoes, workers, and compactors. And, the performance of each tracking method during different challenge factors, including occlusions, illumination variations, motion blur, scale variation and background clutters, are also taken in consideration.

2) Based on the comparative results, proposing two improvements to enhance tracking performance, which are tracking fusion and part-based tracking. The tracking fusion is fusing results of different tracking methods in one sequence in order to get better results than existing ones, while the part-based tracking is to track the different parts of articulated equipment and then combine the tracking results together for seeking the better tracking performance.

This research is applied on both day and night time construction scenarios and the captured sequences were converted into the resolution of 1920*1080 pixels. Meanwhile, the tested scenarios are limited to open-ground earthmoving works in out-door environment and on-site construction period in in-door environment. Also, the tested sequences were captured under normal weather, which have not included raining, snowing, fogy, hail, and other extreme weather.

1.4. Proposed Methodology and Main Conclusion

A total of fifteen (15) visual tracking methods were selected for the proposed comparison, and all of these methods have been validated successful in the computer vision field. These tracking methods were tested with twenty (20) videos, which were captured from different construction sites. As mentioned before, the sequences were characterized with challenging attributes of occlusions, illumination variations, motion blur, scale variation and background clutters. In addition, the construction resources of interest were manually annotated as ground truths. Furthermore, the tracking results of different methods have been measured in a quantitative manner, in terms of accuracy and robustness. And these tracking methods were compared from an overall view and attribute-based view respectively.

After the comparison works, the first improvement is to fuse the tracking results of individual tracking method. For each sequence, the fifth tracking results were put together at the first. Then, the proposed attraction function has been adopted to remove five unreliable results and the similarity function based on image structure was applied to gain weights for the rest of ten bounding boxes. The non-maximum suppression has been implemented in order to get a new bounding box to represent the object. The fusion method was tested on the datasets, which was captured in comparison works, and compared with fifth tracking methods with the criteria of accuracy (bounding box overlap and location error ratio) and robustness (tracking length ratio).

The other improvement is the part-based tracking, which is the idea of tracking the two parts of articulated equipment respectively. Taking the excavator as an example, the robustness and accuracy tracking methods were selected out and have been applied to track buckets and 'cab' respectively. Then, two tracking boxes have been combined in order to get a new tracking box, which represents the prediction result of the excavator. This method was also compared with the same criteria of comparison works in accuracy and robustness separately.

The test results of this research indicated that: 1) the methods based on the sparse representation were more effective than other methods. Furthermore, the local sparse representation had better performance than holistic sparse representation. 2) in construction scenarios, the generative tracking strategy outperformed the discriminative one, while the discriminative methods performed better in computer vision benchmarks for general scenarios. 3) the proposed fusion methods have increased the tracking performance over 10% in accuracy when compare with fifteen tested methods. 4) the tracking performance has been improved in both accuracy and robustness for tracking articulated equipment with part-based tracking.

1.5. Expected Contribution

The outcomes of this research reveal that there is huge difference on performance between different visual tracking methods. Moreover, this research shows that deep learning methods have large potentials in visual tracking technologies, and this method even outperformed sparse representation method in some sequences. In addition, the ensemble of various tracking results is an efficient way in order to improve the final tracking performance. Implementing visual tracking could be very helpful to automation construction. For example, tracking trucks and excavators and could be used to estimate excavation productivities automatically, which is time and labor consuming.

Research finding are encouraged for researchers from construction management to enhance the future applications of using visual tracking technologies. The comparison results could be used to select proper tracking methods when conducting certain applications, which could improve the precision and robustness of existing applications. The proposed fusion method is an offline and generic method, which is easy to achieve and has no requirement for tracking method. This method can be used to overcome challenging factors by integrating the advantages from individual methods. On the other hand, the proposed part-based method could indicate excavators' position with higher accuracy, which is useful to construction safety by calculating the distance between buckets and workers. In this study, only two parts were considered for tracking, and it is possible to track three, four, or even more parts respectively in order to estimate more complex articulated equipment.

1.6. Thesis Structure

The outline of this research is summarized below in Table 1.1 and it highlights the main ideas of each chapter in short sentences.

Chapter	Summary
1) Introduction	This chapter introduces the main idea of this thesis, points out the research gaps and objectives. Meanwhile, the proposed methodology and main conclusion have been stated briefly. At the end, the expected contribution and thesis structure have been presented.
2) Literature Review	This chapter reviews the current practice in vision-based technologies, especially the 2D visual tracking. The existing benchmarks and tracking applications in construction have also been introduced.
3) Proposed Methodology	This chapter explains the proposed methodology step by step, which is followed the introduction in order to achieve the objectives of this thesis.
4) Results and Discussion	In this chapter, research results have been presented and analyzed from multiple views. Findings are highlighted and discussed.
5) Conclusion and Future Works	This chapter concludes the outcomes of this research. Also, future works have been defined.

Table 1.1: Structure of the thesis

CHAPTER 2: LITERATURE REVIEW

In this chapter, it will conduct a review of the recent technologies and practice related to the computer vision field. Then, 2D visual tracking will be introduced in details. After that, benchmark works of visual tracking methods from computer-vision field are also presented. Besides, specific tracking applications in construction management will be summarized at the end.

2.1. Computer Vison Technologies

Computer vision is an interdisciplinary research field, which is aiming to gain higher level understanding of digital images and videos with computers. And it was seeking to achieve the automatically human visual system. The sub-domains of computer vision consist of the scene reconstruction, event detection, object tracking, pose estimation, motion analysis, and image restoration. The development of computer vision do not have long history. In the late 1980s, the computer vision technologies have been developed at universities that were pioneering the computer intelligence. Studies in the 1980s have built the early foundations of many computer vision algorithms, which exist today. For example, extraction of edges from images, mean-shift, optical flow and labeling of lines. By the 1990s, research in 3-D reconstructions resulted in the better understanding of camera calibration. Also, with the advance of optimization methods, it was recognized that the exploring in bundle adjustment was used to improve the field of photogrammetry. In addition, variations of graph cut were developed to solve image segmentation problem. In the next decade, it was the first time statistical techniques were adopted in recognizing faces in images. Recent work has been focused on the resurgence pf feature-based methods, which used in interface with machine learning techniques. A classification of computer vision tasks would help researchers to reduce the difficulty of problems. The tasks of computer vision can be categorized into following classes:

1) Imaging processing

In order to improve the quality of imaging, imaging processing is aiming to decline noise of images. This is the fundamental of advance level of computer vision tasks.

2) Feature extraction

Extracting feature, for example, lines, textures, edges and regions from processed images. And feature extraction provides different representation of the world object information.

3) Object recognition

The main idea of object recognition is to identify objects in the world after being given the models or patterns of known objects.

4) Motion analysis

The aim of motion analysis is to retrieve properties or status of objects, such as structure, position, and velocity, according to the given motion information based on images.

5) 3D reconstruction

The 3D reconstruction is retrieving shape, sized of objects in the world by gaining the 3D coordinates of each vertex of the objects.

2.1.1. Technologies in computer vision

In order to achieve different tasks of computer vision, different technologies have been developed to simplify the state of art problems

Image processing: Imaging processing takes raw pixels as input and produces another pixels as output, which have higher level of quality than original images under certain views. After image processing, noise, such as blurring caused by cameras and geometrical distortion caused by lens, could be reduced or removed. The technologies of image processing could be divided into two categories: real domain and Fourier domain. The main idea of real domain processing manipulate the images pixel by pixel directly and the smoothing templates convolve each pixel in order to remove noise. Fourier domain processing performs a Fourier transform in order to gain spectrum representation of images. The spectrum image has been processed then. In general the Fourier space processing is much faster than the real domain processing. And this is because of the real domain processing often needs to do convolution works pixel by pixel, which are time-consuming.

According to the motivation of image processing, the technologies of image processing can be classified into image improvement and image restoration. The image improvement utilizes certain performance index to measure the quality of images, and adapt the image to higher quality. The purpose of image restoration is aiming to restore a better image based on known degeneration models. In general, the purpose of image improvement is to deal with images and make human more comfortable. And the purpose of image restoration is to process image and make it more understandable for computers.

Feature extraction: In general, image feature is properties of pixels, which can be edge, region, texture, lines, curves, etc. There also exists complicated features, which combine basic features, and upgrade them with mathematic expression. The feature extraction have been applied in advance research fields of computer vision, for example, the image retrieval, image registration, object recognition, object categorization, and robot localization.

In order to represent the pixels, a large variety of feature have been developed. The earliest works could be the local derivatives (Koenderink and Van 1987). Then, Florack et al. adopted a series of local derivatives and constructed differential invariants for local feature representation in 1994. The local derivatives have been extended to the local gray value invariants for the purpose of image retrieval (Scihmid and Mohr 1997). Because Gabor functions (Daugman 1980) have the ability to represent the receptive field profiles in cortical simple cells, Marcelja (1980) worked on the responses of the mammalian visual cortex based on a series of Gabor functions. Textons (Leung and Malik 2001) and the Varma–Zisserman model (Van et al. 1996), have been proved to have reliable performance in the task of texture classification. The SIFT feature, which is a 3D histogram of gradient magnitudes representation, have demonstrated the effectiveness. The advantages of SIFT is that its invariance in challenging factors in terms of illumination variation, background clutter, occlusion, etc. Carneiro and Jepson (2003) proposed phase-based local features in order to enhance the invariance to illumination changes. Ke and Sukthankar (2004) have simplified the SIFT by utilizing principal component analysis (PCA), which normalized gradient patches and shown good performance on image deformations. Lazebnik et al. (2005) divided each circular normalized patch into concentric rings, which put forward the rotation invariant. Then, the gradient location and orientation histogram have shown its significance (Mikolajczyk and Schmid 2005), which has applied PCA to decrease the dimension of the representation. The color information is very important in visual representations, Van and Schmid (2006) developed four color descriptors, which include histograms of RGB, hue, opponent angle, and spherical angle.

Object recognition: As one of the fundamental challenges in computer vision, object recognition is considered as the problem of detecting and localizing objects from given categories such as people or cars in static images (Felzenszwalb et al. 2010). The difficulty of object recognition comes from the huge differences in appearance in such categories. Even in the same categories, the variability in shapes and other visual properties could be enormous, for example, different cars come in various colors and shapes.

There are significant works on object recognition, which including various kinds of deformable template methods (Cootes et al. 2001; Coughlan et al. 2000), and a large number of part-based methods (Amit and Trouve 2007; Burl et al. 1998; Crandall et al. 2005). Fergus (2003) proposed an object recognition method based on the constellation models, the images were separated into parts, which are constrained to be a set of locations determined by interest point operators, and a Gaussian distribution was used to capture the geometric arrangement. In contrast, Felzenszwalb and Huttenlocher (2005) proposed the pictorial structure models, which defined a matching problem where each part have an individual match cost in set of locations. And in this model, the geometric arrangement was captured by "springs", which connecting pairs of parts.

Significant variations, such as caused by viewpoint changes, are not were recognized by deformable models. The aspect graphs (Plantigna and Dyer 1986) has significant performance for capturing the changes from the viewpoint changes. Another approach is to use multiple models. It is widely adopted to use multiple templates to encode different views of cars or faces (Schneiderman and Kanade 2000). Discriminative training methods adopt the strategy to select model parameters in order to minimize the mistake of object recognition. This approach directly optimizes the boundary of positive and negative images (Dalal and Triggs 2005). On the other hand, the information of context for object recognition has received great attention in recent years. Torralba (2003) proposed a method which use low-level holistic image for defining similar candidates. Some methods (Hoiem et al. 2008) have adopted a coarse representation of a scene, which including its 3D geometry. The discriminatively trained part based model (Felzenszwalb et al. 2010) is good at data-mining, and this approach requires relatively few passes through the complete set of training examples. This method is extremely suitable for training very large date-sets.

Motion analysis: Motion analysis from a sequence of images can be used to extract a lot of information which is difficult from static images. The most common motion analysis is image differencing, which means differencing neighboring few frames in a sequence, and it is possible to detect the edge of moving objects in order to subtract these objects from static background. The single Gaussian model was used in background subtraction (Wren et al. 1997) at the first, but this method did not consider the pixel value of images and more elaborate model was needed. After that, the Gaussian mixture model was extend with a hysteresis threshold (Power and Schoonees 2002). The topology and the number of components of a Markov model was selected in a training procedure (Stenger et al. 2001).

So far, human motion analysis has been a major topic because it has huge potential of applications. For the aim to understand the behaviors of humans, a higher level of understanding is required (Mao et al. 2013). The pipeline of human motion analysis involves three steps: feature extraction, dimension reduction and classification (Aggarwal and Ryoo 2011). In Aggarwal and Ryoo's work, they concluded motion analysis scheme into two categories: Single-layered approach and sequential approach. The single-layered approaches recognized human activities from video directly, and these approaches considered each kind of activity as a specific class and recognize the activity based on given classes. Most single-layered methods adopted the sliding windows to classify candidate subsequences. Sequential approaches are extensions of single-layered approaches by analyzing sequences of features. The sequential approaches firstly convert images into a sequence of feature vectors, such as degrees of joint angles, which could be used to describe the status of a person.

3D reconstruction: 3D information in terms of size, position and shape of objects can be retrieved from images by giving the explicit geometry of the objects. The 3D reconstruction of an object can be summarized as the process which begins from the data acquisition and ends at the output of 3D virtual model on a computer (Remondino and EI-Hakim 2006). Three are two main classes of methods for vision-based 3D reconstruction and they are 3D modeling from videos, 3D modeling from oriented images.

Videos to 3D modeling approaches aim to obtain a 3D model from uncalibrated images or videos. In 1998, Fitzgibbon and Zisserman reported an automated procedure in computer vision community. This system extracts corner points automatically and then matches them across views. Some methods have been proposed for the works of extraction of image correspondences (Ferrari et al. 2003). These methods are heavily relied on feature extractions, which results in the results are affected by occlusions, illumination variations and so on. The invariant point detector overcame this problem. The methods based on SIFT operator (Lowe 2004) have improved the robustness under image variations.

Oriented images to 3D modeling approaches automatically orient and calibrate images and then perform the semi-automated modeling (El-Hakim 2002; Guarnieri et al. 2004). This is a more common approach in 3D reconstruction especially in the case of dealing with complex geometric objects. In 1999, Liebowitz et al. proposed a method, which creating 3D graphical models from limited number of images. Then line-photogrammetric mathematical models were employed to recover the 3D shapes of polyhedral objects (Van et al. 1999). Dick et al. (2001) have employed a recognition technique based on models in order to extract higher level models from a single image. D'Apuzzo (2003) has developed an automated surface measurement procedure in order to match the homologous points.

2.1.2. Applications in computer vision

Computer vision technologies have been widely applied in many industries, including medical application, human computer interfacing application, transport and traffic application, etc. ***Medial application:*** Thanks to the development of high-resolution cameras, enhanced computer-based has become a powerful tool for disease discovery. High-content screening cameras are able to capture high resolution images for cells or organisms. This technique would promise to enhance drug discovery pipeline (Gosai et al. 2010). On the other hand, the bright-field images can provide effective data of many kinds of properties (e.g. shape, size and motility), which have been used to detect helminths (Ramot et al. 2008).

The counting and classification of medical objects, such as blood cells is time-consuming and labor-consuming. However, this task could be easily solved by computer vision technology. Ramoser et al. (2006) proposed an automated system to classify white blood cells by adopting computer vision concepts, which involved different features and classifiers. Ongun et al. (2001) proposed a system, which using active contours in order to track the boundaries of white blood cells. Lezoray et al. (1999) has introduced a region-based white blood cells segmentation using extracted markers, while this method relies on proper seed extraction. Kumar et al. (2002) applied a cell edge detector which is trying to determine the boundary of the nucleus. Sinha and Ramakrishnan (2003) proposed a segmentation framework using k-means clustering for a neural network classifier.

A lot of effort has been devoted to developing automated image segmentation techniques in 3D. Recently, Yin et al. (2010) have reported a layered graph approach for optimal segmentation of single and multiple interacting surfaces of human bones. Tu and Xiang (2010) have introduced an auto-context algorithm for 3D brain image segmentation, which could learn the low-level appearance, implicit shape, and context information through a sequence of discriminative models.

In chromosome analysis, the properties of chromosomes can be gained using computer vision techniques from an interactive way. Such approaches have been proved to increase current practice (Jahne B, 2000).

Human computer interfacing application: Applications based on computer vision technology have enhanced persons' mobility and some applications even have the ability to support social interaction. S. Schörnich et al. (2013) have used visual classification technologies to recognize impaired individuals' ability to navigate. Lanigan et al. (2006) have proposed a prototype system named Trinetra, which used barcodes of products to help users recognize supermarket objects. However, this system has a limitation that the user has to be able to adjust the camera view toward the barcode, which could be tedious. To overcome this shortcoming, Tekin and Coughlan (2009) proposed an algorithm which can help users to locate the barcode by giving left or right indications. Another prototype application system (Winlock et al. 2010) could recognize objects in supermarkets that users have put in the shopping list of the phone. In this system, it detects the objects (using an internal database) automatically and then compares the detected objects with the ones on the list. This system will notify the user when it finds a match.

Computer vison also opens a new branch of methods for fall detection. Rougier et al. (2011) have extracted information from the captured video in order to determine if there is a fall or not. This method extracted the head's velocity and shape change information and then differentiate fall or non-fall activities by setting proper thresholds. But the performance of this method is strongly related to the threshold. Auvinet et al. (2011) also proposed a threshold-based method for fall detection, which used calibrated cameras to reconstruct 3D shape of people. Then, fall events were analyzed by the volume distribution along the vertical axis. The system would alarm when the distribution was abnormally changed over a period of time. Juang and Chang (2007) proposed a posture recognition-based fall detection system. In this system, it defined a neural network to do the posture classification. Liu et al. (2010) also focused on the classifier and they replaced with a more common k-nearest neighbor classifier. Belshaw et al. (2011) have adopted three pattern recognition methods for fall detection and compared three methods (logistic regression, neural network, and support vector machine) and the neural network achieved the best performance.

Transport and traffic application: Pedestrian detection is an essential task in transportation engineering due to its potential for enhancing human safety and there are a lot of efforts have been made to improve the performance of pedestrian detection (Dollar et al. 2012). On the other hand, road safety has become more and more crucial around the world and the traffic sign are important because it contains a lot of important information about current traffic environment. In 1987, Akatsuka and Imai proposed a study on traffic sign recognition system. This system could be used as an assistance for drivers for the purpose of alerting them when encountering some specific sign, such as the one-way street sign. Fu and Huang (2010) summarized the traffic sign recognition procedure as shown in the Figure 2.1. And they have divided the traffic sign recognition system into three stages: 1) detecting candidates of signs; 2) tracking these candidates; 3) classification of these candidates.

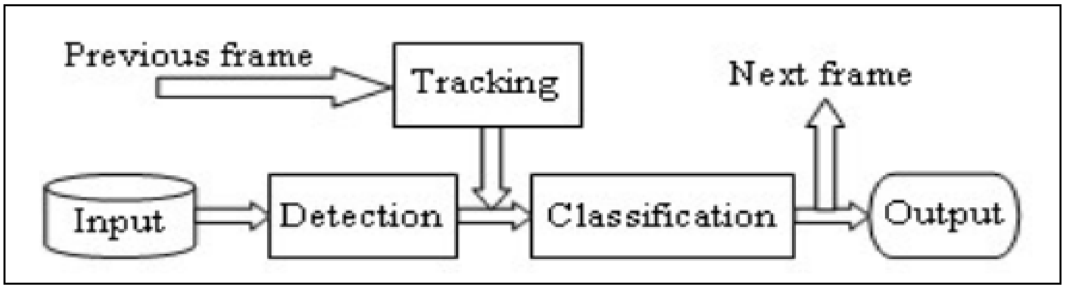


Figure 2.1: Procedure of traffic sign recognition (Fu and Huang 2010)

2.2. Two Dimensional (2D) Visual Tracking

2D visual tracking is one of the most important problems in the field of computer vision. The processing of 2D visual tracking is to estimate the states of objects (e.g., position and extent) in a sequence of 2D frames after given the initial state of the objects in the first frame (Wu et al. 2015). Wang et al. (2015) explained how 2D visual tracking system works, and they broke the 2D visual tracking system into five constituent parts: Motion Model, Feature Extractor, Observation Model, Model Updater, and Ensemble Post-processor (as showed in the Figure 2.2).

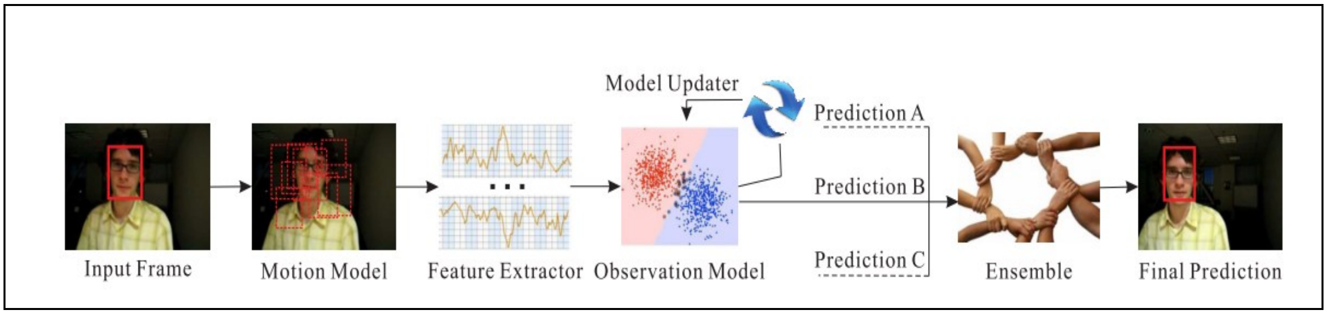


Figure 2.2: Pipeline of visual tracking system (Wang et al. 2015)

1) Motion Model: The motion model generates a large number of candidate bounding boxes, which may contain the expecting object. The motion model is working based on the analysis of the previous frame.
2) Feature Extractor: The feature extractor is used to represent each candidate with some features.
3) Observation Model: The observation model is used to justify if the candidate box is the object based on the features extracted before.
4) Model Updater: The model updater is used to decide when and how to update the observation model.

5) Ensemble Post-processor: If a tracking system contains multiple trackers, the ensemble post-processor combines the results of each tracker in order to get the better performance than individual trackers.

A tracking system usually starts from initializing the observation model after given the information of object bounding box in the first frame. Then, the motion model works and generates candidate proposals based on the estimation of the previous frame. These candidate proposals are sent to the observation model in order to compute their statistical probability of being the object. The candidate, which has the highest probability is then selected. Based on the output of the observation model, the model updater decides if the observation model should be updated. At the end, if there exists multiple trackers, the tracking results of each tracker will be combined by the ensemble post-processor to obtain a better estimation result.

In recent, modern tracking are usually complicated systems. Generally, researchers concluded 2D tracking methods into two categories: generative methods and discriminative methods. The generative methods are assuming a process to generate lots of candidate regions and search for the most similar candidate as the target. The discriminative methods always train a classifier to separate targets from the backgrounds.

2.2.1. 2D visual tracking development

The early work of visual tracking started from Lucas and Lanade (1981), which adopted raw intensity holistic templates (LK method). However, this method did not consider the appearance variability, which results in it performed not well in tracking objects with significant changes. To overcome this limitation, the subspace-based tracking methods have been proposed to account the changes. For example, Black and Jepson (1998) have adopted the pre-trained Eigen basis representation in tracking. And Hager and Belhumer (1998) have updated the LK method with low-dimensional representations and gained better performance under varying conditions. Then, the sparse-representation-based methods (Mei and Ling; 2009) were proposed and have been improved which showed the high performance in tracking blurry objects. In this method, they used a dictionary of holistic intensity templates, which composed of target.

In order to handle occlusions, local sparse representations have been introduced for visual tracking (Jia et al. 2012; Bao et al. 2012). In the works of Bao, the accelerated proximal gradient approach was introduced to solve L1 minimization problems. Liu et al (2011) proposed a sparse representation method and adopted the mean-shift algorithm to locate objects. And this methods efficiently increased the tracking robustness. Then, a collaborative tracking algorithm that combined a discriminative classifier and a generative model, was proposed (Zhong et al. 2014) to enhance the tracking accuracy. Zhang et al (2012) converted visual tracking to a multi-task sparse representation learning problem.

As one of the most important information of images, color histograms have gained a lot of interests in visual tracking. Perez et al. (2002) embedded color histograms in a particle filter for visual tracking. And Coumaniciu et al. (2003) have proposed a color histogram-based mean-shift tracking method. Besides relying on pixel-wise statistics, the spatiograms was used to calculate both the statistical properties of pixels and their spatial relationships (Birchfield and Rangarajan, 2005). Considering the edge information, the histograms of oriented gradients (HOGs) have been adopted for visual tracking (Tang et al. 2007). To combine different types of features, covariance region descriptors (Tuzel et al. 2006) were introduced for object tracking. On the other hand, the local binary patterns (Ojala et al. 2002) and Haar-like features (Viola and Jones 2004) have also been explored to describe the object appearance for tracking.

For the discriminative tracking methods, various classifiers have been proposed, such as support vector machine (SVM) (Avidan 2004), structured output SVM (Hare et al. 2011), and ranking SVM (Bai and Tang, 2012). In specific, Avidan (2004) have integrated a trained SVM classifier in the optical flow framework. The multiple instance learning has been applied to tracking (Babenko et al. 2011), in which all positive and negative samples are put together to learn a discriminative model finally. Grabner et al. (2006) proposed an online boosting method to select proper features to separate objects from the background.

The tracking problem is now treated as an optimization framework and gradient descent methods can be used to locate the objects efficiently. For example, Fan et al. (2010) adopted a discriminative model to identify attentional regions with the gradient descent formulation to predict the objects. Sevilla-Lara and Learned-Miller (2012) proposed a tracking framework based on distribution fields. In this framework, it allows smoothing the objective function, and the object is located by searching for the local minimum. The dense sampling methods (Babenko et al. 2011) have also been adopted to solve the problem that visual tracking are usually nonlinear with local minima.

So far, using online update to describe the appearance variations plays an important role for successful object tracking. Matthews et al. (2004) proposed a new template update method, which updates the template with combining the fixed reference template extracted from the first frame and the result from the previous frame. Grabner et al. (2008) converted the update problem as a semi-supervised task and the classifier was updated with both labeled and unlabeled data. In order to exploit the potential of the unlabeled data, Kalal et al. (2010) have developed a tracking system where the semi-supervised learning method was used to select positive and negative samples for model update. In recent, the context information have been utilized to facilitate visual tracking because the context information can provide extra visual features from the surroundings of the objects. Dinh et al. (2011) exploited some supporters around the objects by using the method of sequential randomized forest. The context information is proved useful when the objects are fully occluded or out of the camera view.

Another line of research have been resorted to train deep networks with large scale of data, and then utilize the trained models to do visual tracking. In 2010, Fan et al proposed a present human tracking method which could learn a specific feature extractor with CNNs from a 2000 images dataset. Wang and Yeung (2013) have developed a deep learning tracking method which learns the generic features from a large scale of datasets (1 million images). Wang et al. (2015) used a two-layer CNN to learn hierarchical features from sequences. And this method takes into account complicated motion transformations in visual tracking. Zhang et al. (2016) presented a convolutional network based tracker which exploits the local structure and inner geometric layout information of the objects. Hong et al. (2015) put an additional layer of the online Support Vector Machine (SVM) on the top layer of CNN to learn the object appearance and discriminate it from the background. Nam and Han (2016) proposed a tracking method based on a CNN trained in a multi-domain learning framework.

2.2.2. Evaluation criteria in visual tracking

Considering the evaluation of tracking methods and comparison to the state-of-art, there is no standardized evaluation protocol right now. Current evaluation criteria focus on the robustness and accuracy, which have been widely adopted and proved useful. The accuracy criteria, such as center error (Adam et al. 2006) and region overlap score (Godec et al. 2013), demonstrate if the tracking correctly locates the target. The robustness criteria, such as tracking length (Kwon and

Lee, 2009) and failure rate (Kristan et al. 2010), describe the methods' ability of keep tracking the targets. The detailed explanations of each criteria are demonstrated as following.

Center error: The center error is one of the most popular ways of measuring tracking performance. It measures the distance between the center of predicted bounding box and the center of ground truth box. The advantage of center error comes from the minimal annotation requirement. The Equation 1 displays the calculation of center error CE. The Figure 2.3 is an illustration of the center error.

$$\text{Equation 1: } CE = \|x_t^T - x_t^G\|$$

t : represents a certain frame and $t\epsilon\{1, N\}$

x_t^T: the center loaction of tracked box at the t frame

x_t^G : the center loaction of ground truth box at the t frame

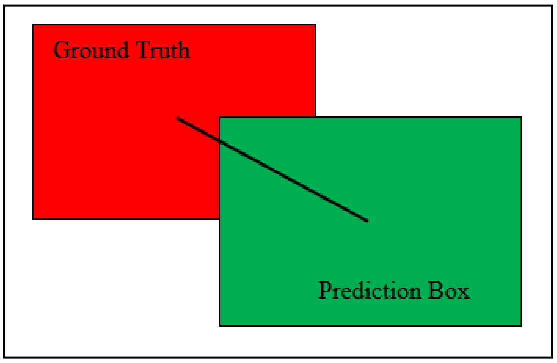


Figure 2.3: An illustration of center error

Region overlap score: The region overlap score is calculated as an overlap between predicted object region and the ground truth region. The Equation 2 displays the calculation of region overlap score OS and the Figure 2.4 is an illustration.

$$\text{Equation2: } OS = \left\|\frac{A_t^G \cap A_t^T}{A_t^G \cup A_t^T}\right\|$$

t : represents a certain frame and $t\epsilon\{1, N\}$

A_t^T: the tracked bounding area at the t frame

A_t^G : the ground truth bounding area at the t frame

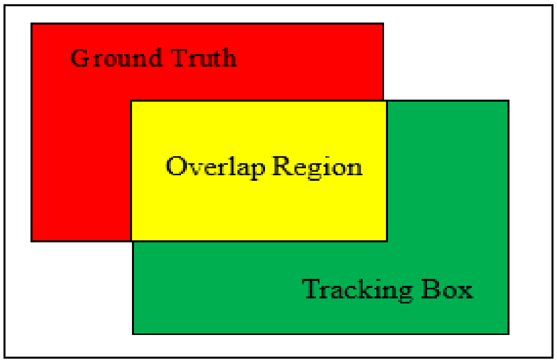


Figure 2.4: An illustration of region overlap score

Tracking length: The tracking length reports the number of frames from the first frame to its first failure. The failure criterion can be manual judged or inspected. But this criterion results in different results even by the same person. It is better to automate the failure criterion. And, it is possible to place a threshold on the center or overlap. The Figure 2.5 demonstrates the tracking length measure.

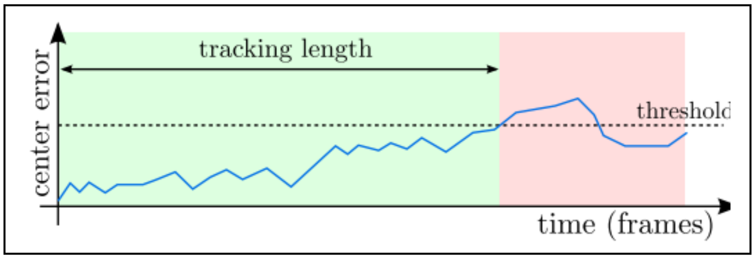


Figure 2.5: An illustration of the tracking length (Cehovin et al. 2014)

Failure rate: The tracking length could be used to evaluate the robustness performance of tracking methods. However, it is very sensitive at the beginning stage. If the beginning of sequence contains a difficult tracking situation, it results in a poor initialization and then this criteria would not have persuasion. The failure rate is created to solve this issue and it is working in a supervised system. It is calculated as the number of number of the tracking failure times over the number of the whole frames, which is demonstrated in the Figure 2.6.

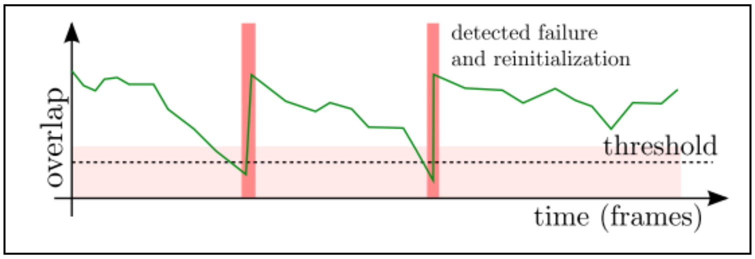


Figure 2.6: An illustration of the failure rate (Cehovin et al. 2014)

The evaluation criteria mentioned above are always combined to evaluate the performance of the tracking methods in a comprehensive manner. Wu et al. (2013) proposed the precison plot (Figure 2.7) and success plot (Figure 2.8) in order to do the quantitative analysis.

Precision plot: The precision plot shows the percentage of frames whose location errot is within the given threshold distance.

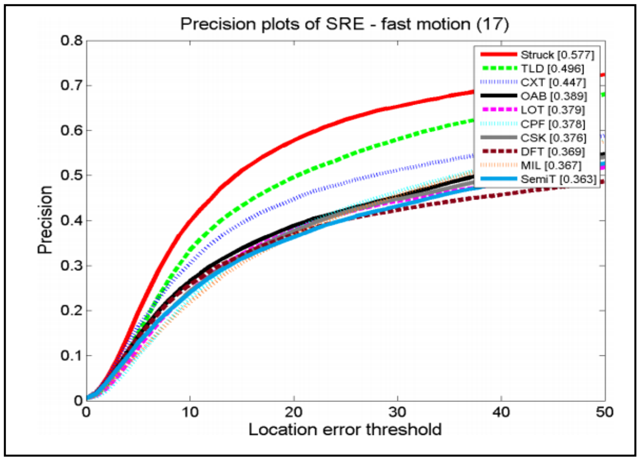


Figure 2.7: An illustration of the precision plot (Wu et al. 2013)

Success plot: The success plot shows the ratios of successful frames at the thresholds varied from 0 to 1, (e.g. to=0.5).

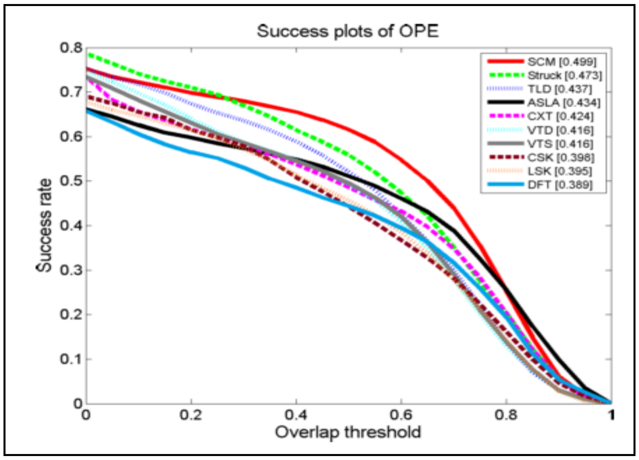


Figure 2.8: An illustration of the success plot (Wu et al. 2013)

Hybrid measure: Kristan et al. (2015) proposed a hybrid measure that puts the accuracy and robustness scores into one graph in order to decide which method has the better performance overall in terms of accuracy and robustness (Figure 2.9).

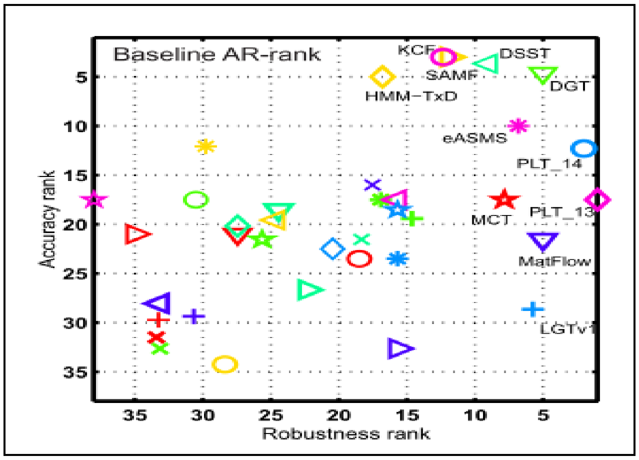


Figure 2.9: An illustration of the hybrid measure (Kristan et al. 2015)

CoTPS: Nawaz and Cavallaro (2013) proposed a threshold-independent and overlap-based measure called the Combined Tracking Performance Score (CoTPS). In the CoTPS, the video frames where the overlap scores are higher than a pre-defined threshold are defined as the successfully tracking frames. Then, the accuracy score is calculated as the number of the successfully tracking frames, while the robustness score is represented by the corresponding tracking length (Figure 2.10).

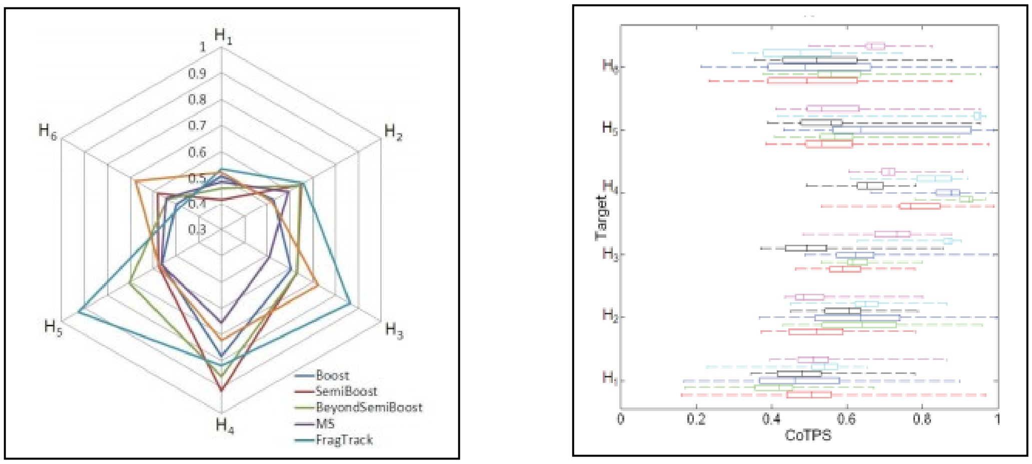


Figure 2.10: An illustration of the CoTPS measure (Nawaz and Cavallaro 2013)

Cehovin et al. (2014) concluded the state-of-art evaluation criteria and analyzed them in a correlation way. This analysis helps us to find the correlation between each two evaluation criteria and explore the relationship between different evaluation criteria (Figure 2.11).

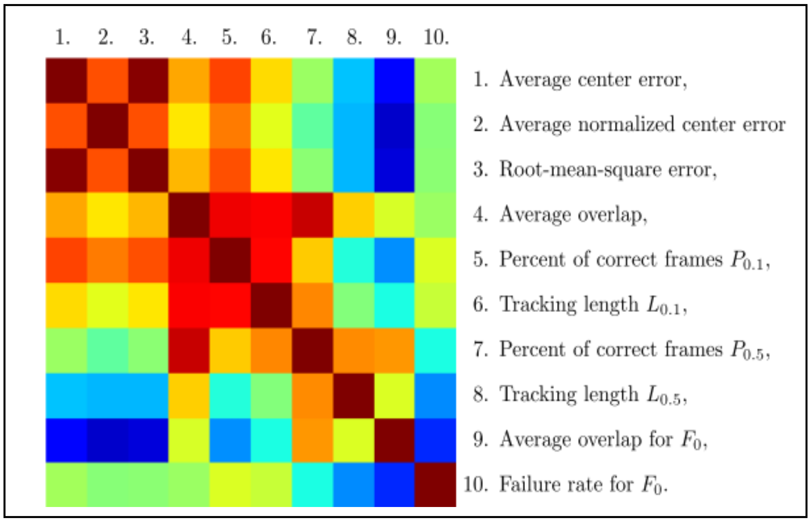


Figure 2.11: An illustration of the correlation analysis of evaluation criteria (Cehovin et al. 2014)

2.2.3. Visual tracking benchmarks

Until now, several benchmarks have been created to evaluate the performance of existing visual tracking methods. These benchmarks contained huge image data-sets and compared a large number of tracking methods based on different evaluation criteria. The detailed description of each benchmark has been summarized in Table 2.1.

Datasets	Published Year	Number of tracking methods	Number of videos
OTB1.0 (Wu Yi et al)	2013	29	50
ALOV (Smeulders et al)	2014	19	315
VOT2013 (Kristan Matej et al)	2013	27	356
VOT2015 (Kristan Matej et al)	2015	62	356
OTB2.0 (Wu Yi et al)	2015	31	100
NUS-PRO (Li Annan et al)	2016	20	365

Table 2.1: Summary of some popular benchmarks

OTB benchmark: Wu et al presented the OTB 1.0 benchmark in 2013, which tested 29 tracking methods on 50 different sequences. In 2015, Wu et al updated the work of OTB1.0 by OTB 2.0, which tested 31 tracking methods on 100 different sequences. The testing scenarios in OTB benchmark are general living scenarios, which is shown in the Figure 2.12.

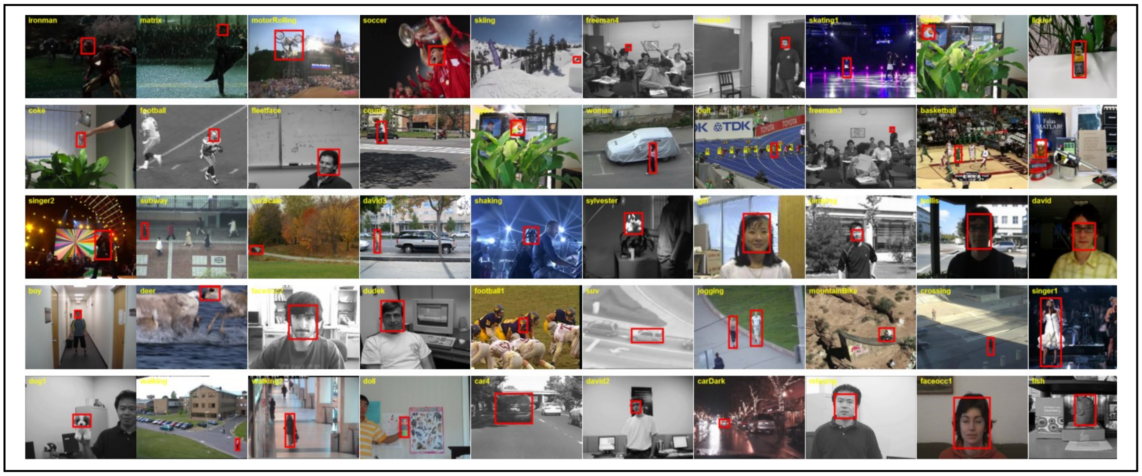
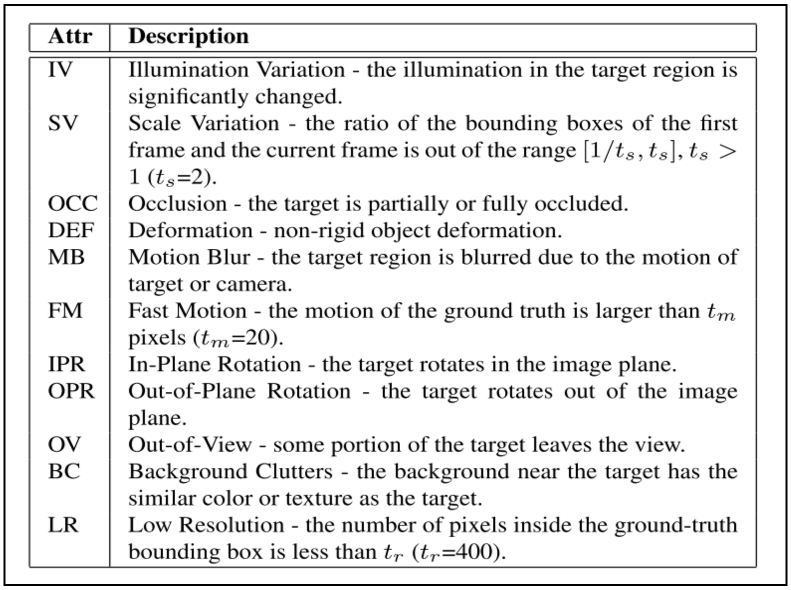


Figure 2.12: An illustration of sequences in OTB benchmarks (Wu et al. 2013)

In OTB datasets, they adopted success plot and precision plot to evaluate the accuracy performance. And then, they proposed two ways to analyze the robustness of a tracking method, which are perturbing the initialization temporally (i.e., start at different frames) and spatially (i.e., start by different bounding boxes). It is difficult to evaluating tracking method even there is proper criteria because many factors can affect the tracking performance. For better analysis of the strength and weakness of tracking methods, OTB benchmarks proposed an attribute-based methods, which annotate sequences with the 11 attributes shown in Figure 2.13.

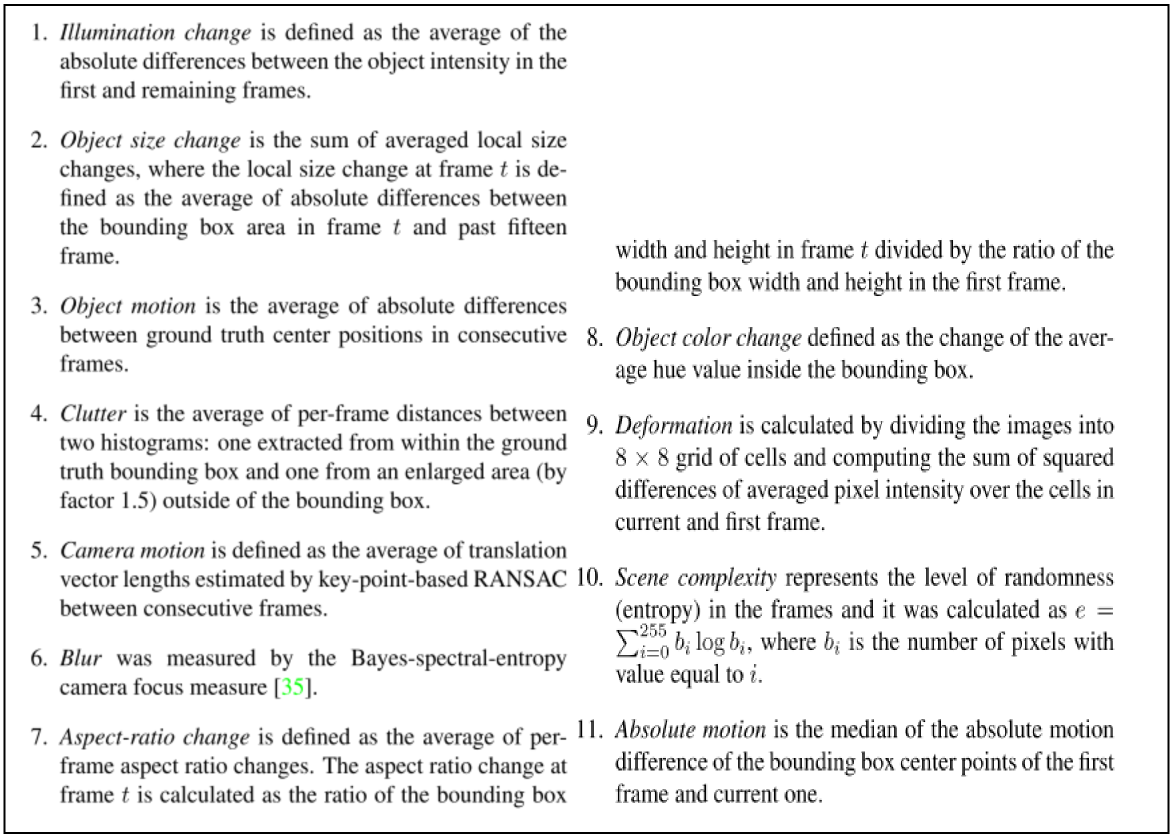


Attr	Description
IV	Illumination Variation - the illumination in the target region is significantly changed.
SV	Scale Variation - the ratio of the bounding boxes of the first frame and the current frame is out of the range $[1/t_s, t_s]$, $t_s > 1$ (t_s=2).
OCC	Occlusion - the target is partially or fully occluded.
DEF	Deformation - non-rigid object deformation.
MB	Motion Blur - the target region is blurred due to the motion of target or camera.
FM	Fast Motion - the motion of the ground truth is larger than t_m pixels (t_m=20).
IPR	In-Plane Rotation - the target rotates in the image plane.
OPR	Out-of-Plane Rotation - the target rotates out of the image plane.
OV	Out-of-View - some portion of the target leaves the view.
BC	Background Clutters - the background near the target has the similar color or texture as the target.
LR	Low Resolution - the number of pixels inside the ground-truth bounding box is less than t_r (t_r=400).

Figure 2.13: The list of attributes in OTB benchmarks (Wu et al. 2013)

The OTB benchmarks indicate that: 1) Background information is important for successful tracking. Using advanced learning techniques to utilize the background information in the discriminative models (e.g., Struck). 2) The local models are effective for tracking and local sparse representation methods have shown reliable performance compared with other tracking methods. 3) Motion models play a crucial for object tracking, especially when the motion of target is large. 4) The large-scale performance benchmarks facilitate better understanding of the state-of-the-art visual tracking methods.

VOT benchmark: The visual object tracking (VOT) challenge has been proposed in 2013 and updated in 2015. In the VOT 2015, performance of 62 tracking methods have been presented. All these 62 tracking methods have been tested on 356 sequences, and this is the largest datasets in computer vision community. Similar with the OTB benchmark, in the VOT benchmark researchers also annotated their benchmark with different attributes. The description of attributes adopted in VOT benchmark have been summarized in the Figure 2.14.



1. *Illumination change* is defined as the average of the absolute differences between the object intensity in the first and remaining frames.
2. *Object size change* is the sum of averaged local size changes, where the local size change at frame t is defined as the average of absolute differences between the bounding box area in frame t and past fifteen frame.
3. *Object motion* is the average of absolute differences between ground truth center positions in consecutive frames.
4. *Clutter* is the average of per-frame distances between two histograms: one extracted from within the ground truth bounding box and one from an enlarged area (by factor 1.5) outside of the bounding box.
5. *Camera motion* is defined as the average of translation vector lengths estimated by key-point-based RANSAC between consecutive frames.
6. *Blur* was measured by the Bayes-spectral-entropy camera focus measure [35].
7. *Aspect-ratio change* is defined as the average of per-frame aspect ratio changes. The aspect ratio change at frame t is calculated as the ratio of the bounding box width and height in frame t divided by the ratio of the bounding box width and height in the first frame.
8. *Object color change* defined as the change of the average hue value inside the bounding box.
9. *Deformation* is calculated by dividing the images into 8×8 grid of cells and computing the sum of squared differences of averaged pixel intensity over the cells in current and first frame.
10. *Scene complexity* represents the level of randomness (entropy) in the frames and it was calculated as $e = \sum_{i=0}^{255} b_i \log b_i$, where b_i is the number of pixels with value equal to i.
11. *Absolute motion* is the median of the absolute motion difference of the bounding box center points of the first frame and current one.

Figure 2.14: The list of attributes in VOT benchmarks (Matej et al. 2015)

In the VOT benchmark, there are also accuracy and robustness criteria to evaluate the performance. The VOT2015 benchmark indicates that the MDNet achieved the best performance in both accuracy and robustness, which had very few fails. This findings prove that the deep learning method has huge potentials, although this method requires a large number of computational resources.

ALOV benchmark: In 2014, Smeulders et al. proposed a comparative study of 19 visual tracking methods. The characteristic of ALOV benchmark is that it adopted the F-scores to combine the accuracy and robustness evaluation. The comparison results of ALOV indicate that: 1) the circumstance is very important for visual tracking, especially considering the occlusion and clutter situations. 2) TLD method performs remarkable on camera motion results from its well-designed detection and motion model. 3) The using of F-score permits comparison of tracking methods from a statistical view. 4) There still exist some difficulties in state-of-arts tracking methods, such as occlusions. 5) It could be seen that simple models with a low complexity perform better in this benchmark.

NUS-PRO benchmark: There were 315 video sequences downloaded and annotated from YouTube. All images in this benchmark have been converted to the same size, i.e., 1280×720 pixels. The description of sequences in NUS-PRO has been summarized in the Figure 2.15.

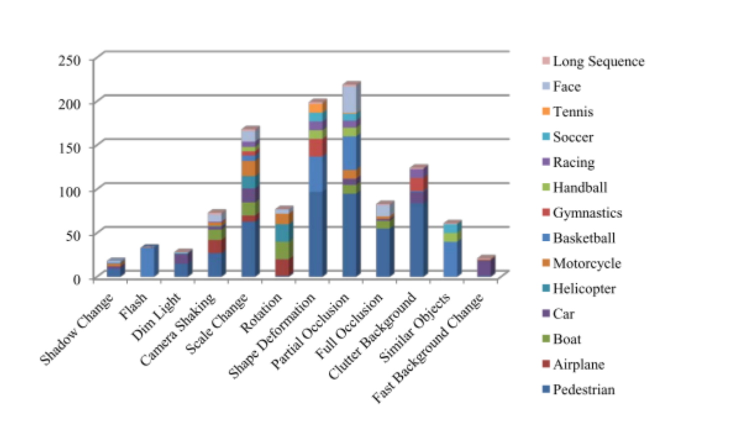


Figure 2.15: The description of sequences in NUS-PRO benchmark (Li Annan et al. 2016)

The tracking results of NUS-PRO benchmark shows that: 1) The ASLA, SCM, and LOT actually achieved the overall better performance. 2) The ASLA, SCM, and OAB have showed the better performance in long-term tracking. 3) Existing methods are not effective in handling full occlusions in this benchmark.

2.3. 2D Visual Tracking in Construction

The use of visual tracking in construction has been recently promoted to facilitate construction automation. Visual tracking has been widely utilized to track construction equipment and workers. For example, Yang et al. (2010) have developed a multiple tracking system to evaluate worker's performance. Gong and Caldas (2011) have measured the working cycles of a mini loader used the mean-shift tracking method. Park and Brilakis (2012) have investigated the tracking performance of scale-invariant feature transform and speeded up robust features for the tracking of construction workers. Yang et al. (2014) presented a single Gaussian background tracking method, which was used to track tower crane jibs for the purpose of identifying the working cycle times. Zhu et al. (2016) proposed a particle filtering method to track the workers and equipment in construction sites.

In 2012, Rezazadeh Azar and McCabe have adopted the visual tracking in order to recognize and calculate the dirt loading cycles during earthmoving works. A server-customer interactive tracking system has been developed to detect trucks being loaded and measure the loading time. Another important use of visual tracking is to monitor construction safety. It was used to protect workers on foot from potential collisions (Han and Lee 2013). In this study, there are four processes to achieve the safety monitoring for workers: Identification of Critical Unsafe Actions, Data Collection, Motion Capture, and Motion Recognition.

However, there are still limited comparative studies regarding the visual tracking methods in the construction scenarios. Park et al. (2011) once presented a comparative study of 2D visual tracking methods, where only kernel-based, contour-based, and point-based tracking methods were included for the comparison. The comparison results indicated that the kernel-based method performs better than point-based method in dealing with illumination variations and scale variations. Also, the kernel-based methods were considered as the most effective method for tracking construction site resources. This comparison work still remains some limitations in four aspects: 1) the novel tracking methods which adopted different schemes and achieved promising performance in computer vision were not considered. 2) the test scenarios in their study were mainly captured from the miniatures of construction jobsites with scaled equipment models. 3) in Park's study, there were only three types of attributes considered in construction sites (Illumination Variation, Occlusion, and Scale Variation). Actually, there exist more challenging factors in construction sites. 4) the evaluation criteria in that study was center location error, while the tracking methods should be evaluated by multiple criteria from. Therefore, the comparison results might not truly reflect the overall performance of the tracking methods on real construction sites.

CHAPTER 3: PROPOSED METHODOLOGY

In this chapter, the proposed methodology has been introduced in this research. In the first section, the methodology for visual tracking comparison will be presented in terms of methods selection, sequence selection, attributes annotation, and evaluation strategy. Then, the main steps of fusion of tracking and part-based tracking have also been illustrated.

3.1. Visual Tracking Comparison

The overall objective of this section is to evaluate and compare existing visual tracking methods in construction scenarios. The methods selected here for the evaluation and comparison have already shown the promising tracking performance in the computer vision community. The results and findings from this research are expected to help construction researchers and professionals select appropriate visual tracking methods that could meet their application demands, when dealing with complex and realistic construction conditions, such as occlusions and illumination changes. The overall steps adopted for the evaluation and comparison have been illustrated in Figure 3.1.

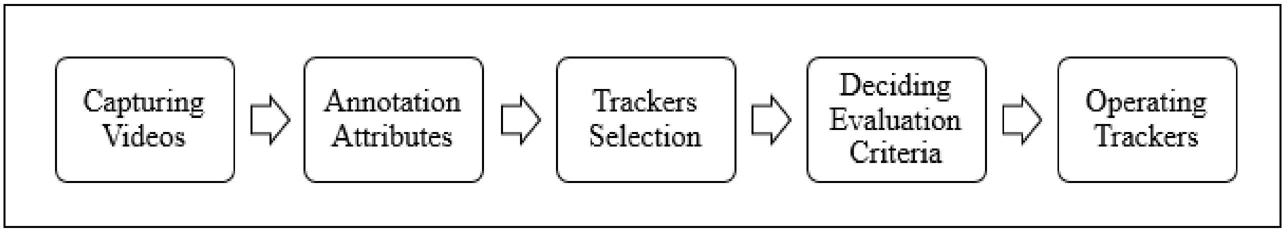


Figure 3.1: Flowchart of experimental methodology

3.1.1. Video sequence selection and attributes annotation

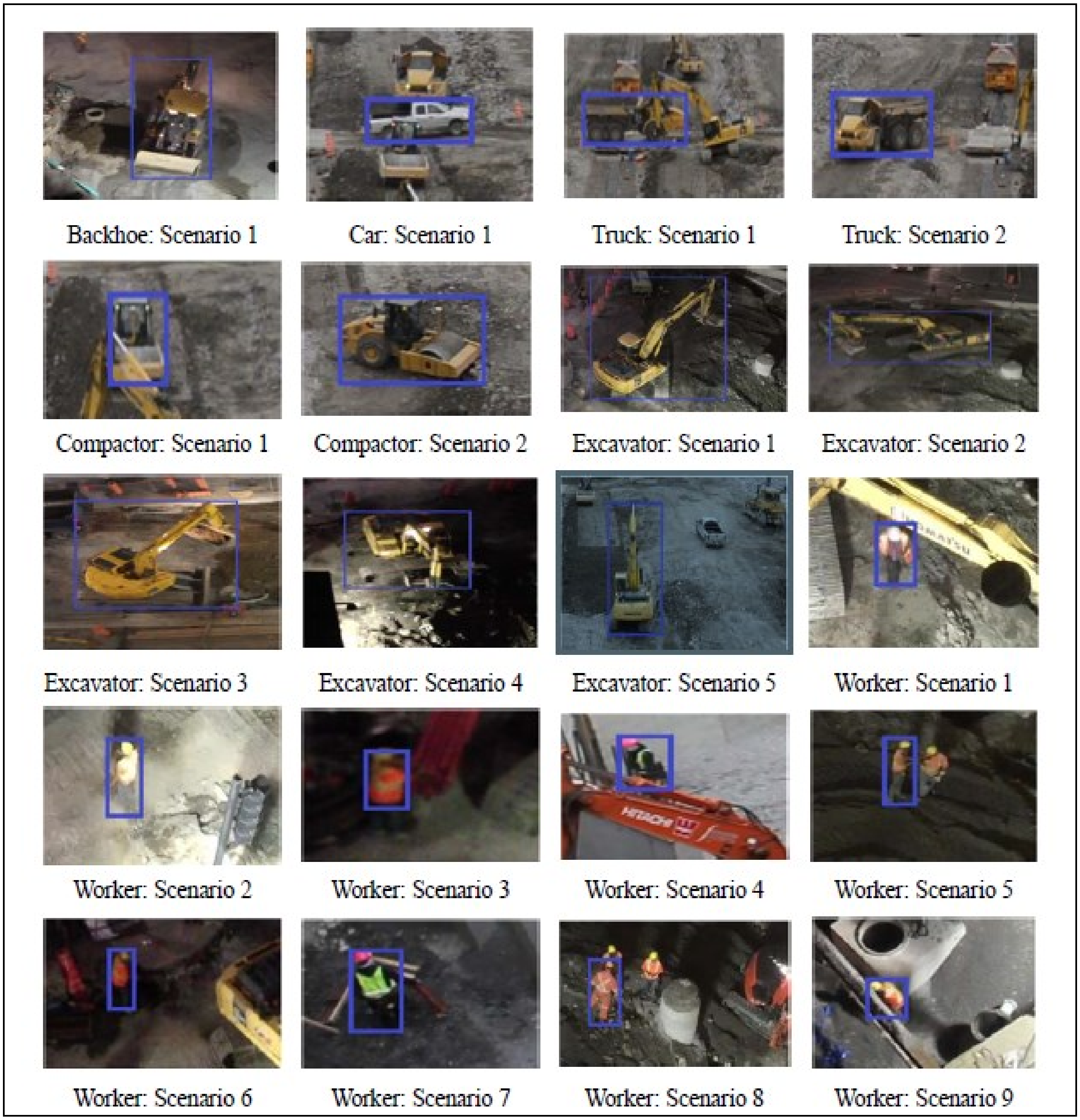
The construction video sequences were all captured from real construction jobsites. 20 videos that represent the construction of civil infrastructure, residential buildings, and municipal facilities were selected. These videos contain different types of common construction equipment such as excavators, backhoes, and compactors. In addition, construction workers are also included as one of the construction resources of interest.

Each video sequence was further annotated with the attributes to represent specific challenging factors which might affect the tracking results. In this study, five attributes, i.e. Occlusions (OCC), Illumination Variation (IV), Motion Blur (MB), Background Clutters (BC), and Scale Variation (SV), were defined, referring to the work of Wu et al. (2015). The definition of each attribute has been summarized in Table 3.1.

Attributes	Descriptions
Illumination Variation (IV)	Whether the target experiences significant illumination changes in the video sequence; e.g. an excavator is moving into or out from a shade.
Occlusions (OCC)	Whether the target is occluded in the video sequence, so that it could not be fully seen for a period.
Motion Blur (MB)	Whether the target is blurred in the video sequence due to its motions or because the camera is out of focus during the video capturing.
Background Clutters (BC)	Whether the target and the neighboring background objects in the video sequence have similar colors and/or textures.
Scale Variance (SV)	Whether the size of the target in the video sequence experiences significant changes; e.g. an excavator is moving close to or far away from the camera.

Table 3.1: Attributes annotated to test video sequences

Annotating the video sequences with the attributes improves the understanding of the tracking methods' strength and weakness. For example, there are 5 video sequences annotated with OCC. They could be put together to evaluate how well a tracking method is able to deal with occlusion conditions. The Figure 3.2 shows the examples of the construction videos with the attributes, while the parameters of the video sequences, such as the targets contained, the numbers of video frames, and the overall durations, have been listed in Table 3.2. It is worth noting that the durations of all the video sequences are less than 20 seconds in this study. The short durations are because only the video sequences with the challenging parts were selected for the tests. This idea was also supported in the computer vision community, where most of the test video sequences in existing benchmarks range from 5 seconds to 15 seconds (Wu et al. 2015; Kristan et al. 2015).



Backhoe: Scenario 1
Car: Scenario 1
Truck: Scenario 1
Truck: Scenario 2
Compactor: Scenario 1
Compactor: Scenario 2
Excavator: Scenario 1
Excavator: Scenario 2
Excavator: Scenario 3
Excavator: Scenario 4
Excavator: Scenario 5
Worker: Scenario 1
Worker: Scenario 2
Worker: Scenario 3
Worker: Scenario 4
Worker: Scenario 5
Worker: Scenario 6
Worker: Scenario 7
Worker: Scenario 8
Worker: Scenario 9

Figure 3.2: Video sequence descriptions

Target		Attributes	Number of frames	Duration (s)
Backhoe	Scenario 1	MB, BC	500	20
Car	Scenario 1	MB	350	14
Truck	Scenario 1	OCC	500	20
	Scenario 2	SV	400	16
Compactor	Scenario 1	OCC	500	20
	Scenario 2	SV	400	16
Excavator	Scenario 1	IV, BC	500	20
	Scenario 2	IV, BC	400	16
	Scenario 3	BC, SV	500	20
	Scenario 4	BC, SV	450	18
	Scenario 5	SV	500	20
Worker	Scenario 1	IV, MB, BC	170	7
	Scenario 2	IV, MB, BC	400	16
	Scenario 3	IV, MB, BC	300	12
	Scenario 4	OCC, SV	180	8
	Scenario 5	OCC, BC	270	11
	Scenario 6	IV, BC	300	12
	Scenario 7	BC	400	16
	Scenario 8	OCC, BC	300	12
	Scenario 9	IV, OCC, BC	200	8

Table 3.2: Summary of video sequences

3.1.2. Visual tracking methods selection

A total of fifteen visual tracking methods have been selected in this research study. The tracking methods selected here are not solely based on their published years. Also, their performances in existing object tracking benchmarks are considered. For example, the ASLA tracking method (Jia et al. 2012) was published in 2012, but its performance was better than many others published later according to the report of Wu et al. (2015). The specific selection process here is described as follows. First, 31 tracking methods in the OTB2.0 benchmark (Wu et al. 2015) were investigated. It was found that the SCM (Zhong et al. 2014), ASLA (Jia et al. 2012), CSK (Henriques et al. 2012), and L1APG (Bao et al. 2012) methods showed the top performance in the overall ranking (Wu et al. 2015). Also, Wu et al. (2015) reported that the SCM (Zhong et al. 2014), ASLA (Jia et al. 2012), CSK (Henriques et al. 2012), and DFT (Sevilla-Lara and Learned-Miller, 2012) methods were better than the others when handling occlusions and illumination variations;

the CSK (Henriques et al. 2012), TLD (Kalal et al. 2010), L1APG (Bao et al. 2012), and LOT (Oron et al. 2015) methods were good at overcoming motion blurs; the SCM (Zhong et al. 2014), ASLA (Jia et al. 2012), CSK (Henriques et al. 2012), and MTT (Zhang et al. 2012) methods were successful for addressing background clutters; and the SCM (Zhong et al. 2014), ASLA (Jia et al. 2012), CSK (Henriques et al. 2012), and L1APG (Bao et al. 2012) methods showed promising performance on handling scale variations. To summarize these findings, the following eight tracking methods, i.e. ASLA (Jia et al. 2012), CSK (Henriques et al. 2012), DFT (Sevilla-Lara and Learned-Miller, 2012), L1APG (Bao et al. 2012), LOT (Oron et al. 2015), MTT (Zhang et al. 2012), SCM (Zhong et al. 2014), and TLD (Kalal et al. 2010), were first selected.

In addition to the eight tracking methods mentioned above, another seven visual tracking methods were further selected, including STC (Zhang et al. 2014), DLT (Wang et al. 2013), CNT (Zhang et al. 2016), CF2 (Ma et al. 2015), KCF (Henriques et al. 2015), DCF (Henriques et al. 2015), and RPT (Li et al. 2015). The selection was mainly due to their recent developments and reliance on different machine learning or computer vision techniques, such as spatial context learning, deep learning, and kernelizing. Moreover, all of them have not been included and compared in existing benchmarks in the computer vision community. The detailed descriptions of all the tracking methods evaluated and compared in this research study have been summarized in Table 3.3.

Tracking Methods	Sparse Representation		Searching Scheme			Deep learning	Generative(G) or Discriminative(D)	Characteristics & Selection Reasons
	Local	Holistic	Particle Filter	Local Optimum	Dense Sampling			
CNT (Zhang et al. 2016)	√		√			√	D	Reliance on convolutional neural networks
CF2 (Ma et al. 2015)			√			√	D	Reliance on convolutional neural networks
DCF (Henriques et al. 2015)			√				D	Reliance on linear and Gaussian correlation filters; Fast; Integration with HOG features
KCF (Henriques et al. 2015)			√				D	Reliance on kernelized correlation filters
LOT (Oron et al. 2015)			√				G	Using partial-appearance representation; Top performance in handling BC in OTB 2.0
RPT (Li et al. 2015)			√				D	Reliance on exploiting reliable backgrounds information; Kernerlizing
SCM (Zhong et al. 2014)	√		√				G&D	Using sparse collaborative appearance model; Top overall performance in OTB 2.0
STC (Zhang et al. 2014)					√		G	Learning dense spatio-temporal relationship from context; Fast
DLT (Wang et al. 2013)			√			√	D	Reliance on convolutional neural networks
ASLA (Jia et al. 2012)	√		√				G	Using structural local sparse appearance model; Top overall performance in OTB2.0
CSK (Henriques et al. 2012)					√		D	Utilizing the circulant structure; Top overall performance in OTB2.0
DFT (Sevilla-Lara & Learned-Miller, 2012)				√			G	Using distribution representation; Top performance in handling OCC in OTB 2.0
L1APG (Bao et al. 2012)		√	√				G	Using accelerated proximal gradient; Top performance in handling SV in OTB 2.0
MTT (Zhang et al. 2012)		√	√				G	Online learning multiple instance; Top performance in handling BC in OTB 2.0
TLD (Kalal et al. 2010)					√		D	Using positive and negative labels to train an online classifier; First detection tracking

Table 3.3: Description of tracking methods in this study

3.1.3. Evaluation criteria and strategies

In this study, both accuracy and robustness were utilized to evaluate the selected tracking methods. For the accuracy aspect, the overlap score (OS) and center location error (CE) are adopted following Eq. 1 and 2. The tracking length (TL) is calculated with Eq. 3 to evaluate tracking methods in terms of robustness for each video sequence.

The strategy of using the evaluation criteria to test one tracking method is described as follows. In the experiments, each test video sequence is first manually annotated to locate the construction target of interest as the ground truth. Then, the ground truth in the first video frame is directly extracted to help the method initialize the tracking process. The tracking process continues, until it is found that the method fails to locate the target any more (i.e. its OS is less than 0.5). The tracking method is not reinitialized after its first failure. Instead, its tracking performance in the next ten video frames after the failure are considered, when calculating its average sequence overlap score (AOS) and center location error ratio (CER) (Eq. 4 and 5).

$$OS = \left\| \frac{A_t^G \cap A_t^T}{A_t^G \cup A_t^T} \right\| \qquad \text{(Eq. 1)}$$

$$CE = \| x_t^T - x_t^G \| \qquad \text{(Eq. 2)}$$

$$TL = \frac{n}{N} \qquad \text{(Eq. 3)}$$

$$AOS = \frac{1}{n+10} \sum_{t=1}^{n+10} \left\| \frac{A_{t+10}^G \cap A_{t+10}^T}{A_{t+10}^G \cup A_{t+10}^T} \right\| \qquad \text{(Eq. 4)}$$

$$CER = \frac{1}{n+10} \sum_{t=1}^{n+10} \left\| \frac{x_{t+10}^G - x_{t+10}^T}{size(A_{t+10}^G)} \right\| \qquad \text{(Eq. 5)}$$

Where N is the number of total vides frames for a test; n is the video frame where the method fails; t represents a certain frame and $t \epsilon \{1, N\}$; A_t^T is the tracked bounding area at the t frame; A_t^G is the ground truth bounding area at the t frame; x_t^G is the central location of ground truth box; x_t^T represents the central location of tracked box.

In this study, CER is adopted to evaluate the accuracy of the tracking methods instead of CE. This is because CE completely ignores the size of the target under tracking, which results in its sensitivity to the ground truth through manual annotations. In addition, it is found that the tracking methods behaved differently after their first tracking failures (i.e. OS < 0.5). They might lose the targets shortly with the significant OS decrease; or still be able to follow the targets and have their OS slightly below 0.5 in a short period. That is why the tracking performances on additional ten video frames after the first failure are included in Eq. 4 and 5, to differentiate the tracking methods in a more comprehensive manner.

3.2. Fusion of Tracking

It is common to find that general better methods perform worse in specific sequences and some tracking methods are extremely effective in some sequences. In this research, the goal is to fuse the tracking results of sixteen tracking methods, which only requires the positions of bounding boxes as input, and achieve better tracking results in construction scenarios. The main steps of the fusion method has been illustrated in the Figure 3.3. For each frame, these are fifteen bounding boxes generated from sixteen tracking methods, which are the inputs in the fusion processing. First of all, these fifteen bounding boxes were imported by the attraction function, which is used for calculating the relationship between each bounding box and others. Then, five bounding boxes have been removed, which did not perform well in the attraction function. Moreover, these ten bounding boxes have been compared with previous bounding box in last frame by the similarity function. The similarity function was used to generate weights of ten bounding boxes, and each weight represent how similar it is with previous bounding box. Finally, these ten bounding boxes and weights have been fusion to one bounding box by adopting the non-maximum suppression method. This fusion method have been employed in all tested sequences in comparison works.

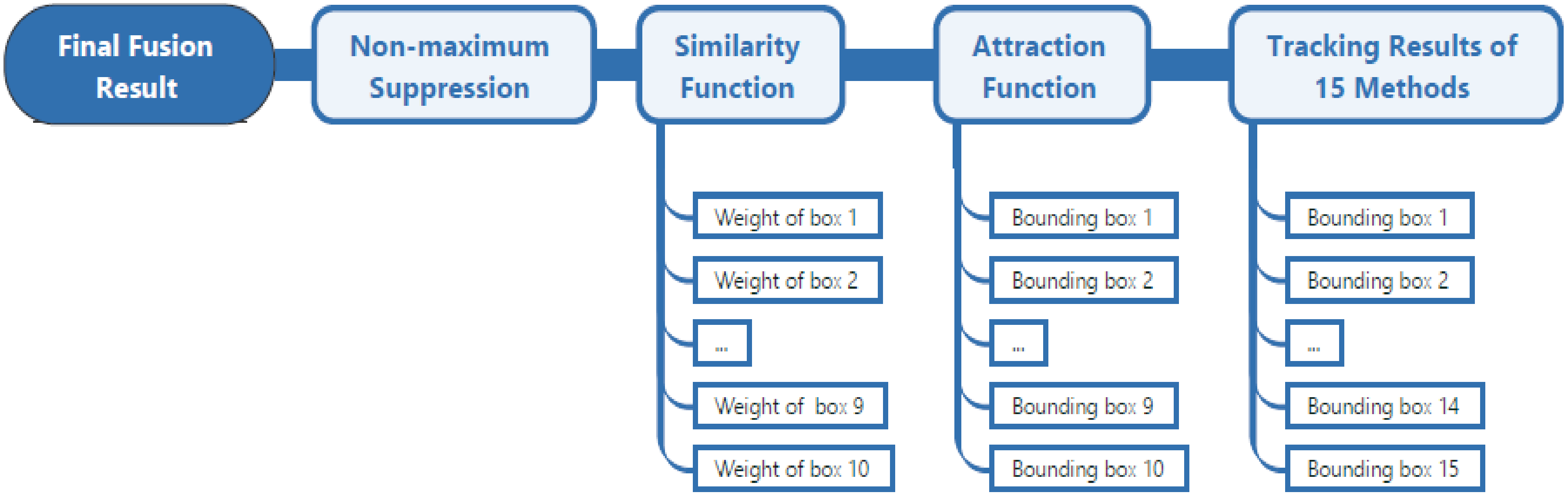


Figure 3.3: Flowchart of tracking fusion

3.2.1. Attraction function

The attraction function is based on the idea that the stronger candidate will attract other candidates and be closer with others. In this study, the attraction function considered not only the distance but also the area between each two bounding boxes. The attraction function is calculated as the Eq. 6 and 7.

$$d(b_j, c) = \frac{OS(b_j,c)}{Area_L/Area_S} \qquad \text{(Eq.6)}$$

$$a(c) = \sum_{j \in M} d(b_j, c) \qquad \text{(Eq. 7)}$$

In the equation, the b_j represents the bounding box from a tracking method except the c. The $OS(b_j, c)$ means the Overlap Score of bounding box c and the bounding box b_j . The $Area_L$ is the larger Area of these two bounding boxes, while the $Area_S$ is the smaller Area of two bounding boxes. M represents the tracking methods adopted in this study and $a(c)$ is the attraction of bounding box c. For each bounding box in each frame, the attraction has been calculated according to the Eq. 6 and Eq. 7, and the larger attraction has been assumed the better performance. After the calculation of attraction, five bounding boxes that did not perform well have been removed and the rest of ten bounding boxes were imported into the similarity function.

3.2.2. Similarity function

The similarity function is used to calculate the similarity of each candidate and the tracked results in the previous frame. In this study, the structural similarity index method (SSIM) (Wang et al. 2004) has been adopted. The SSIM is based on three terms comparison between two images, which are the luminance term, the contrast term, and the structural term. The main function of SSIM has been demonstrated in the Eq. 8-Eq. 11.

$$SSIM(x, y) = [l(x, y)]^{\alpha}. [c(x, y)]^{\beta}. [s(x, y)]^{\gamma} \qquad \text{(Eq. 8)}$$

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \qquad \text{(Eq. 9)}$$

$$c(x,y) = \frac{2\delta_x\delta_y + C_2}{\delta_x^2 + \delta_y^2 + C_2} \qquad \text{(Eq. 10)}$$

$$s(x,y) = \frac{2\delta_{xy} + C_3}{\delta_x\delta_y + C_3} \qquad \text{(Eq. 10)}$$

The $C_1, C_2, and\ C_3$ are default parameters, where $\mu_x, \mu_y, \delta_x, \delta_y, and\ \delta_{xy}$ are the local means, standard deviations, and cross-covariance for images x, y. After calculating the similarity of each candidate and the tracked result in the last frame, the SSIM outputs a value from zero to one. This value is considered as weight and used in the non-maximum suppression process.

3.2.3. Non-maximum suppression

The non-maximum suppression (NMS) is formulated as local maximum search, which means to find a local maximum is greater than all its neighbors (Neubeck and Luc 2006). The NMS has been widely used in objection detection in order to get the proper bounding box when there are multiple bounding boxes. The main idea of NMS is to sort all bounding boxes according to their area. The bounding box with largest area has been selected and compared with others. If the overlap region of two bounding boxes is smaller than a threshold (0.9), the smaller bounding box will be suppressed. Otherwise, the overlap region will be the new bounding box and compare with next bounding box to repeat this process. However, this method has some limitation if it is used in this study directly. It is because the NMS is starting from the largest bounding box and it is easily to make mistakes if the beginning box cannot represent the object. In this study, the weights generated from similarity process have been implemented in the NMS. It means the NMS starts the fusion from the abounding box with largest weight or possibility.

3.3. Part-based Tracking

The tracking of articulated construction equipment is a challenging in the construction scenarios. It was found that that most visual tracking methods could track the articulated equipment with the similar performance of tracking the un-articulated one, when the articulated equipment was just moving from one place to another on the construction site, such as Excavator1, Excavator2, and Excavator5. However, when an excavator is excavating, a lot of self-occlusions are produced due to its boom movements. These self-occlusions significantly impact the tracking performances (Excavator 3 and Excavator 4) and results in non-effective tracking. The Figure 1.2 shows an example tracking performance of an excavator.

Generally, an excavator includes four mainly tracking components: boom, dipper, bucket and “house” (driving cab). The single-objects tracking algorithms usually focus on the house of the excavators because this component has biggest area and moves slowly, when the buckets move fast and hard to be predicted. Therefore, there are two initial tracking boxes adopted in this study, which is showed in the Figure 3.4. The first part is the “house” and grab rails, and the second part is bucket and dipper. And we find the two tracking boxes can always reflect the tracking box of the whole excavator. Then, two bounding boxes have been combined together as the final tracking result.

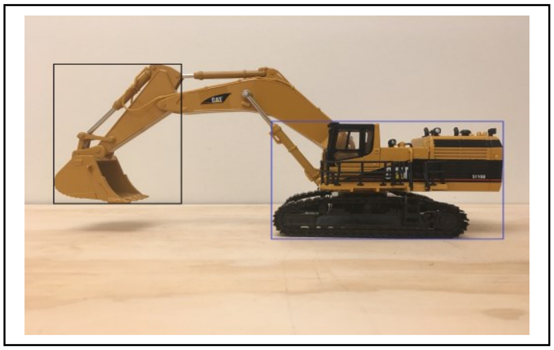


Figure 3.4: Example of initial positions of tracking boxes

The first part (cab and rails) is relatively easy to track, while the second part (bucket and dipper) is harder to track due to the fast moving and self-occlusions in earth moving works. In this study, one better accuracy tracking method in the comparison work has been used to track the cab and trails part. Meanwhile, top five tracking methods have been adopted to track the bucket part for seeking the better performance and avoiding the invalid tracking of articulated equipment. The five part-based tracking methods have been tested on the Excavator sequences and adopted the same evaluation strategy as the comparison work.

CHAPTER 4: RESULTS AND DISCUSSION

The tracking methods for the comparison in this research study were all implemented and tested in the Matlab R2014b platform under the 64-bit operating system, Microsoft Windows 7 Enterprise. The main hardware configuration includes an Intel® i7-4720HQ CPU (central processing Unit) @2.60 GHz, a 16 gigabytes memory, and an NVIDIA® GeForce® GTX 965M with 2GB GDDR5 Graphic Processing Unit. In this chapter, there are four sections. The results of comparison will be introduced at the first. Then, the results of tracking fusion and part-based tracking will be reported. In the final section, the discussion based on the previous results will be illustrated.

4.1. Visual Tracking Comparison Results

The detailed tracking results are summarized in Table 4.1. In all fifteen tracking methods tested and compared in this research study, the ASLA (Jia et al. 2012), SCM (Zhong et al. 2014), MTT (Zhang et al. 2012), L1APG (Bao et al. 2012), CSK (Henriques et al. 2012) and DFT (Sevilla-Lara & Learned-Miller 2012) tracking methods were noted to achieve the overall better performance in both accuracy and robustness than the others. In order to statistically validate this finding, the paired two samples t-tests (Simonoff 2002) were further conducted to calculate the corresponding confidence levels. The t-test results were summarized in Table 4.2. For example, it could be seen that the confidence levels to accept that the ASLA (Jia et al. 2012) method was overall better than the STC (Zhang et al. 2014), KCF (Henriques et al. 2015), DCF (Henriques et al. 2015), TLD (Kalal et al. 2010), LOT (Oron et al. 2015), RPT (Li et al. 2015), CNT (Zhang et al. 2016), CF2 (Ma et al. 2015), and DLT (Wang et al. 2013) methods reached 95% in this study.

		ALSA (Jia et al. 2012)	CSK (Henriques et al. 2012)	CNT (Zhang et al. 2016)	CF2 (Ma et al. 2015)	DFT (Sevilla-Lara & Learned-Miller, 2012)	DCF (Henriques et al. 2015)	DLT (Wang et al. 2013)	KCF (Henriques et al. 2015)	L1APG (Bao et al. 2012)	LOT (Oron et al. 2015)	MTT (Zhang et al. 2012)	RPT (Li et al. 2015)	SCM (Zhong et al. 2014)	STC (Zhang et al. 2014)	TLD (Kalal et al. 2010)
Backhoe	AOS	0.82	0.79	0.80	0.76	0.78	0.74	0.85	0.72	0.81	0.77	0.82	0.74	0.80	0.76	0.69
	CER	0.03	0.02	0.02	0.06	0.04	0.04	0.03	0.06	0.03	0.22	0.03	0.06	0.03	0.03	0.14
	TL	0.35	0.40	0.41	0.41	0.41	0.51	0.25	0.51	0.39	1.00	0.36	0.41	0.38	0.26	1.00
Car	AOS	0.90	0.86	0.92	0.65	0.88	0.73	0.72	0.73	0.89	0.86	0.90	0.85	0.92	0.72	0.72
	CER	0.06	0.07	0.05	0.26	0.05	0.08	0.04	0.08	0.06	0.08	0.05	0.08	0.05	0.04	0.12
	TL	1.00	1.00	1.00	0.13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.71
Truck	AOS	0.85	0.81	0.81	0.75	0.82	0.65	0.73	0.65	0.81	0.86	0.82	0.81	0.85	0.76	0.77
	CER	0.10	0.15	0.11	0.25	0.13	0.10	0.09	0.11	0.11	0.14	0.13	0.18	0.10	0.11	0.21
	TL	0.87	0.57	0.85	0.33	0.64	0.36	0.25	0.45	0.77	0.80	0.72	0.48	0.93	0.43	0.36
Compactor	AOS	0.84	0.82	0.82	0.79	0.83	0.69	0.83	0.71	0.80	0.86	0.81	0.75	0.84	0.76	0.76
	CER	0.06	0.05	0.07	0.09	0.05	0.07	0.07	0.06	0.07	0.14	0.06	0.12	0.06	0.05	0.15
	TL	0.58	0.54	0.53	0.63	0.54	0.55	0.61	0.64	0.56	0.36	0.57	0.38	0.67	0.58	0.36
Excavator	AOS	0.82	0.85	0.81	0.86	0.80	0.81	0.80	0.81	0.83	0.81	0.83	0.76	0.83	0.72	0.77
	CER	0.10	0.08	0.08	0.12	0.17	0.16	0.05	0.14	0.10	0.11	0.09	0.19	0.09	0.09	0.18
	TL	0.74	0.86	0.84	0.71	0.63	0.86	0.47	0.71	0.74	0.72	0.74	0.71	0.70	0.51	0.58
Worker	AOS	0.85	0.83	0.74	0.80	0.82	0.72	0.55	0.73	0.81	0.78	0.82	0.78	0.80	0.72	0.74
	CER	0.13	0.10	0.17	0.14	0.10	0.11	0.20	0.12	0.10	0.12	0.10	0.12	0.12	0.09	0.17
	TL	0.72	0.57	0.60	0.76	0.65	0.58	0.14	0.55	0.64	0.57	0.70	0.56	0.62	0.60	0.30
Average	AOS	**0.84**	**0.83**	0.79	0.80	0.82	0.73	0.68	0.74	**0.82**	0.81	**0.82**	0.78	**0.83**	0.73	0.75
	CER	0.10	**0.09**	0.12	0.14	0.11	0.11	0.12	0.11	**0.09**	0.12	**0.09**	0.14	**0.10**	**0.08**	0.17
	TL	**0.72**	0.65	**0.69**	**0.64**	0.65	0.64	0.33	0.61	0.67	0.65	**0.70**	0.58	**0.68**	0.56	0.44

Table 4.1: Overall tracking performance (top five were identified with bold)

	STC (Zhang et al. 2014)	KCF (Henriques et al. 2015)	DCF (Henriques et al. 2015)	TLD (Kalal et al. 2010)	LOT (Oron et al. 2015)	RPT (Li et al. 2015)	CNT (Zhang et al. 2016)	CF2 (Ma et al. 2015)	DLT (Wang et al. 2013)
ASLA (Jia et al. 2012)	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
SCM (Zhong et al. 2014)	0.95	0.95	0.95	0.95	0.85	0.95	0.90	0.90	0.95
MTT (Zhang et al. 2012)	0.95	0.95	0.95	0.95	0.95	0.95	0.90	0.90	0.95
L1APG (Bao et al. 2012)	0.95	0.95	0.95	0.95	0.90	0.95	0.90	0.90	0.95
CSK (Henriques et al. 2012)	0.95	0.95	0.95	0.95	0.90	0.95	0.90	0.90	0.95
DFT (Sevilla-Lara & Learned-Miller, 2012)	0.95	0.95	0.95	0.95	0.85	0.95	0.90	0.90	0.95

Table 4.2: Confidence level for the overall performance comparison

Attributes-based Performance: In addition to the overall performance, the performance of the tracking methods in each attribute was also considered. OCC is one of the most common challenges in construction sites. In the test video sequences, there are three OCC levels: heavy (Worker 8 and 9), moderate (Worker 4), and slight (Truck 1 and Compactor 1). Figure 4.1 shows the examples of the targets under three OCC levels and the tracking results of OCC have been showed in the Figure 4.2.

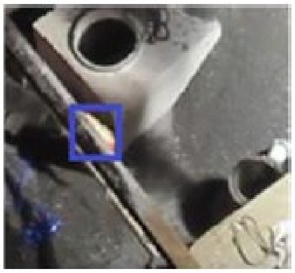
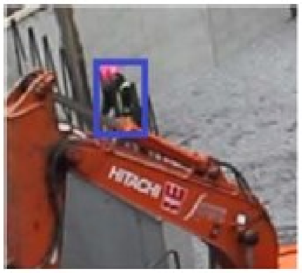
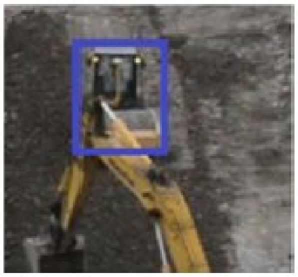
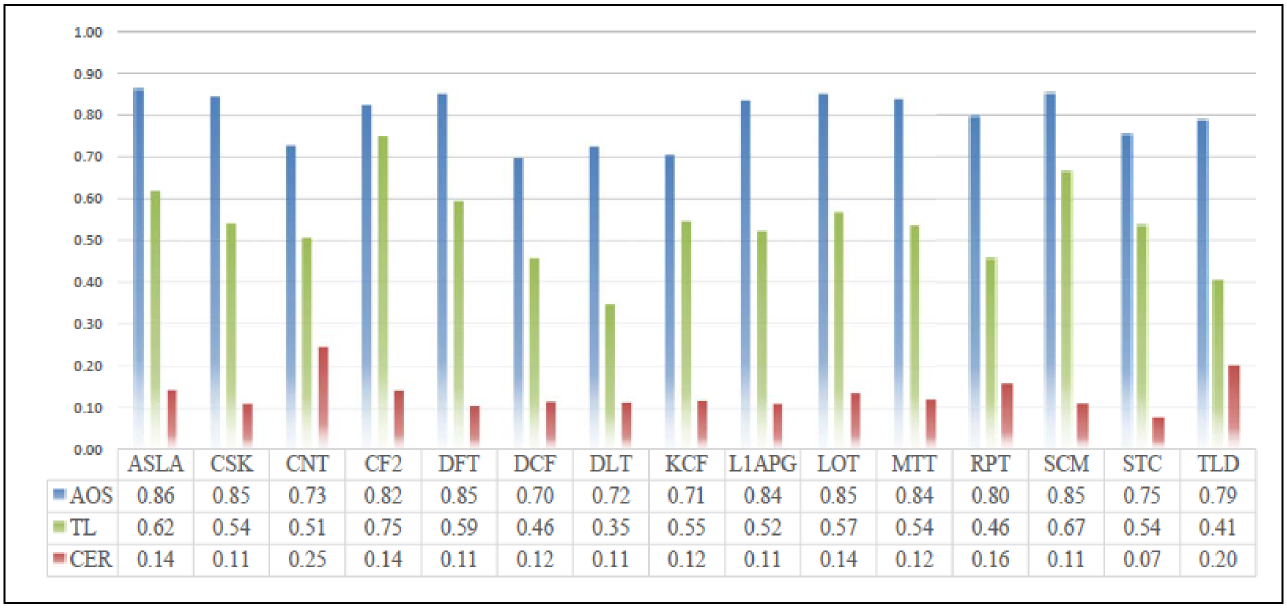
	Heavy Occlusions	Moderate Occlusions	Slight Occlusions
Occlusion Ratio	≥80%	30%-80%	≤30%
Occlusion Frame			

Figure 4.1. Description of different occlusion categorizes



	ASLA	CSK	CNT	CF2	DFT	DCF	DLT	KCF	L1APG	LOT	MTT	RPT	SCM	STC	TLD
🟦 AOS	0.86	0.85	0.73	0.82	0.85	0.70	0.72	0.71	0.84	0.85	0.84	0.80	0.85	0.75	0.79
🟩 TL	0.62	0.54	0.51	0.75	0.59	0.46	0.35	0.55	0.52	0.57	0.54	0.46	0.67	0.54	0.41
🟥 CER	0.14	0.11	0.25	0.14	0.11	0.12	0.11	0.12	0.11	0.14	0.12	0.16	0.11	0.07	0.20

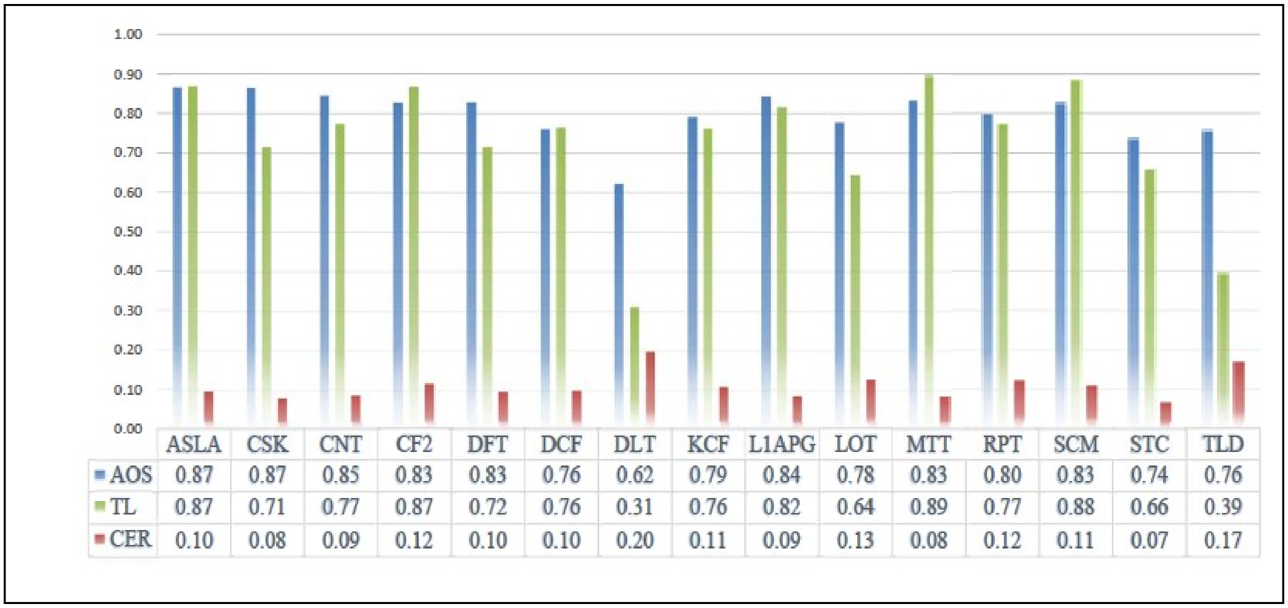
Figure 4.2. Tracking performance of OCC

In this study, the ASLA (Jia et al. 2012), DFT (Sevilla-Lara and Learned-Miler, 2012), SCM (Zhong et al. 2014), CSK (Henriques et al. 2012), CF2 (Ma et al. 2015), L1APG (Bao et al. 2012) and LOT (Oron et al. 2015) methods showed the relatively robust and accurate tracking performance under OCC conditions. Similar to the overall performance comparison, the paired two samples t-tests (Simonoff, 2002) were also conducted and the results were summarized in Table 4.3. It could be seen that the 95% confidence level were achieved to accept that the ASLA (Jia et al. 2012), DFT (Sevilla-Lara and Learned-Miler, 2012), SCM (Zhong et al. 2014), CSK (Henriques et al. 2012), CF2 (Ma et al. 2015), L1APG (Bao et al. 2012) and LOT (Oron et al. 2015) methods were better than the STC (Zhang et al. 2014), KCF (Henriques et al. 2015), DCF (Henriques et al. 2015), TLD (Kalal et al. 2010), MTT (Zhang et al. 2012), RPT (Li et al. 2015), CNT (Zhang et al. 2016), and DLT (Wang et al. 2013) methods in handling occlusions. In contrast, the confident levels to accept that these methods are better than MTT (Zhang et al. 2012) method under occlusion conditions only range from 60% to 90%.

	STC (Zhang et al. 2014)	KCF (Henriques et al. 2015)	DCF (Henriques et al. 2015)	TLD (Kalal et al. 2010)	MTT (Zhang et al. 2012)	RPT (Li et al. 2015)	CNT (Zhang et al. 2016)	DLT (Wang et al. 2013)
ASLA (Jia et al. 2012)	0.95	0.95	0.95	0.95	0.80	0.95	0.95	0.95
DFT (Sevilla-Lara & Learned-Miller, 2012)	0.95	0.95	0.95	0.95	0.90	0.95	0.95	0.95
SCM (Zhong et al. 2014)	0.95	0.95	0.95	0.95	0.85	0.95	0.95	0.95
CSK (Henriques et al. 2012)	0.95	0.95	0.95	0.95	0.85	0.95	0.95	0.95
CF2 (Ma et al. 2015)	0.95	0.95	0.95	0.95	0.60	0.95	0.95	0.95
L1APG (Bao et al. 2012)	0.95	0.95	0.95	0.95	0.60	0.95	0.95	0.95
LOT (Oron et al. 2015)	0.95	0.95	0.95	0.95	0.90	0.95	0.95	0.95

Table 4.3. Confidence level for the OCC performance comparison

As for handling IV, the ASLA (Jia et al. 2012) and CSK (Henriques et al. 2012) were better than the others. The tracking results of IV have been summarized in the Figure 4.3. The results from the paired two samples t-tests (Simonoff, 2002) in Table 4.4 showed 90% ~ 95% confident levels to accept that the ASLA (Jia et al. 2012) and CSK (Henriques e t al. 2012) methods were better, when being compared with the STC (Zhang et al. 2014), KCF (Henriques et al. 2015), DCF (Henriques et al. 2015), TLD (Kalal et al. 2010), DFT (Sevilla-Lara and Learned0Miller, 3012), MTT (Zhang et al. 2012), SCM (Zhang et al. 2014), LOT (Oron et al. 2015), RPT (Li et al. 2015), CF2 (Ma et al. 2015), and DLT (Wang et al. 2013) methods. The confident levels were 80% when the ASLA (Jia et al. 2012) and CSK (Henriques e t al. 2012) methods were better than the L1APG (Bao et al. 2012) and CNT (Zhang et al. 2016) methods under IV conditions.



	ASLA	CSK	CNT	CF2	DFT	DCF	DLT	KCF	L1APG	LOT	MTT	RPT	SCM	STC	TLD
AOS	0.87	0.87	0.85	0.83	0.83	0.76	0.62	0.79	0.84	0.78	0.83	0.80	0.83	0.74	0.76
TL	0.87	0.71	0.77	0.87	0.72	0.76	0.31	0.76	0.82	0.64	0.89	0.77	0.88	0.66	0.39
CER	0.10	0.08	0.09	0.12	0.10	0.10	0.20	0.11	0.09	0.13	0.08	0.12	0.11	0.07	0.17

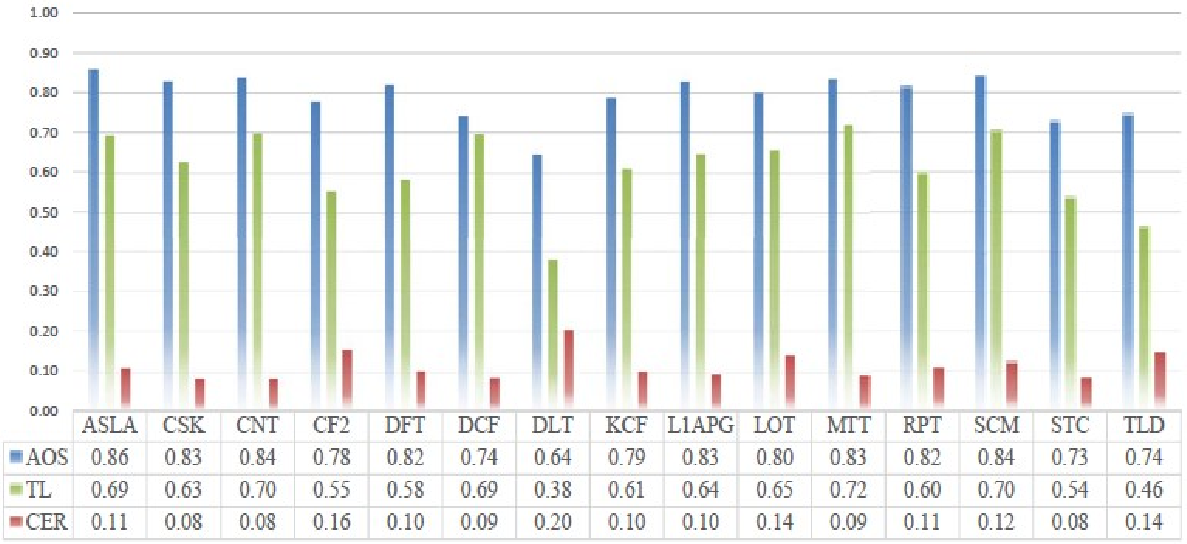
Figure 4.3: Tracking performance of IV

	STC (Zhang et al. 2014)	KCF (Henriques et al. 2015)	DCF (Henriques et al. 2015)	TLD (Kalal et al. 2010)	DFT (Sevilla-Lara & Learned-Miller, 2012)	L1APG (Bao et al. 2012)	MTT (Zhang et al. 2012)	SCM (Zhong et al. 2014)	LOT (Oron et al. 2015)	RPT (Li et al. 2015)	CNT (Zhang et al. 2016)	CF2 (Ma et al. 2015)	DLT (Wang et al. 2013)
ASLA (Jia et al. 2012)	0.95	0.95	0.95	0.95	0.95	0.80	0.95	0.90	0.95	0.95	0.80	0.95	0.95
CSK (Henriques et al. 2012)	0.95	0.95	0.95	0.95	0.95	0.80	0.90	0.90	0.95	0.95	0.80	0.95	0.95

Table 4.4: Confidence level for the IV performance comparison

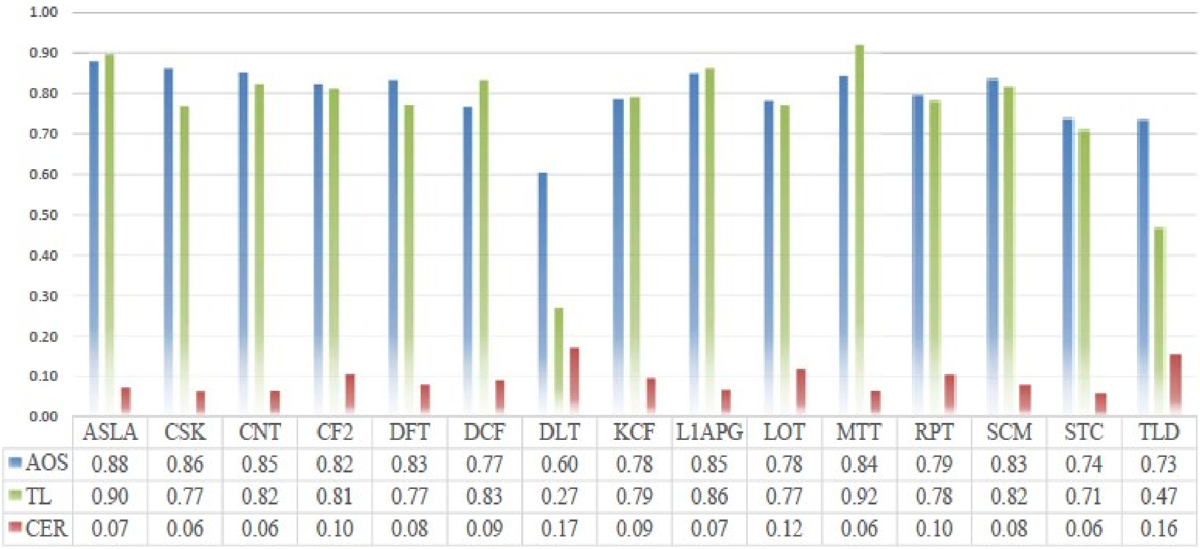
The results under the BC and MB conditions were similar. It was found that the ASLA (Jia et al. 2012), L1APG (Bao et al. 2012), CSK (Henriques et al. 2012), and MTT (Zhang et al. 2012) methods outperformed the others. In the SV conditions, the better ones were the ASLA (Jia et al. 2012), CSK (Henriques et al. 2012), and SCM (Zhang et al. 2014) methods. The confidence levels from the corresponding paired two samples t-tests (Simonoff, 2002) were shown in Table 4.5 and

Table 4.6. And the tracking results of MB, BC, and SV have been summarized in the Figure 4.4, Figure 4.5, and Figure 4.6 respectively.



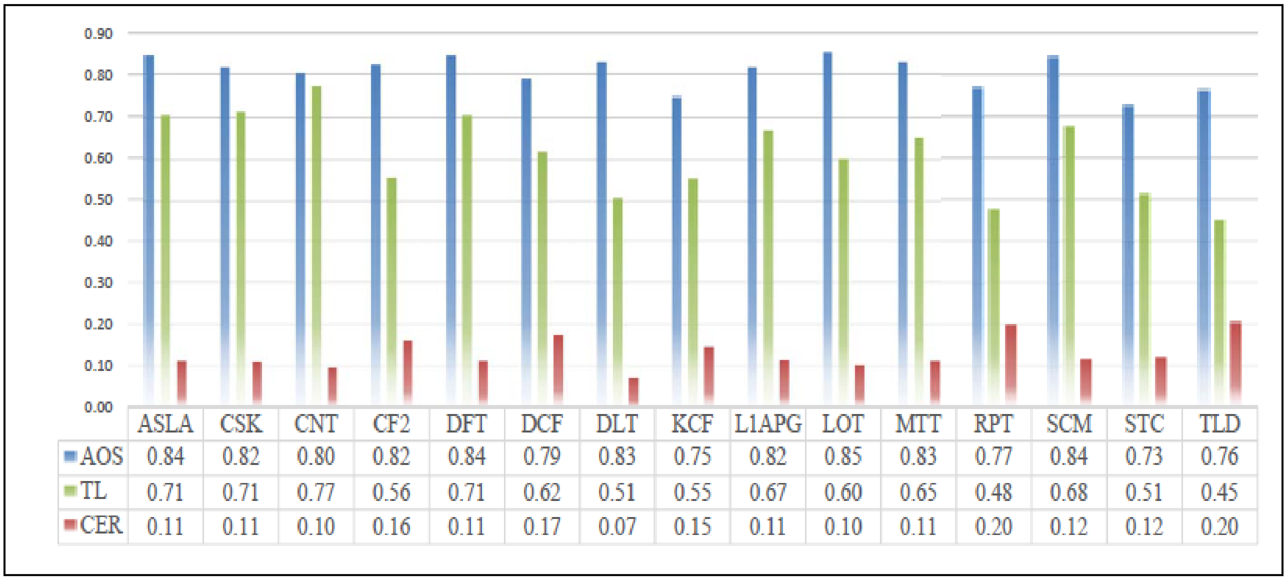
	ASLA	CSK	CNT	CF2	DFT	DCF	DLT	KCF	L1APG	LOT	MTT	RPT	SCM	STC	TLD
🟦AOS	0.86	0.83	0.84	0.78	0.82	0.74	0.64	0.79	0.83	0.80	0.83	0.82	0.84	0.73	0.74
🟩TL	0.69	0.63	0.70	0.55	0.58	0.69	0.38	0.61	0.64	0.65	0.72	0.60	0.70	0.54	0.46
🟥CER	0.11	0.08	0.08	0.16	0.10	0.09	0.20	0.10	0.10	0.14	0.09	0.11	0.12	0.08	0.14

Figure 4.4: Tracking performance of MB



	ASLA	CSK	CNT	CF2	DFT	DCF	DLT	KCF	L1APG	LOT	MTT	RPT	SCM	STC	TLD
🟦AOS	0.88	0.86	0.85	0.82	0.83	0.77	0.60	0.78	0.85	0.78	0.84	0.79	0.83	0.74	0.73
🟩TL	0.90	0.77	0.82	0.81	0.77	0.83	0.27	0.79	0.86	0.77	0.92	0.78	0.82	0.71	0.47
🟥CER	0.07	0.06	0.06	0.10	0.08	0.09	0.17	0.09	0.07	0.12	0.06	0.10	0.08	0.06	0.16

Figure 4.5: Tracking performance of BC



	ASLA	CSK	CNT	CF2	DFT	DCF	DLT	KCF	L1APG	LOT	MTT	RPT	SCM	STC	TLD
AOS	0.84	0.82	0.80	0.82	0.84	0.79	0.83	0.75	0.82	0.85	0.83	0.77	0.84	0.73	0.76
TL	0.71	0.71	0.77	0.56	0.71	0.62	0.51	0.55	0.67	0.60	0.65	0.48	0.68	0.51	0.45
CER	0.11	0.11	0.10	0.16	0.11	0.17	0.07	0.15	0.11	0.10	0.11	0.20	0.12	0.12	0.20

Figure 4.6: Tracking performance of SV

	STC (Zhang et al. 2014)	KCF (Henriques et al. 2015)	DCF (Henriques et al. 2015)	TLD (Kalal et al. 2010)	DFT (Sevilla-Lara & Learned-Miller, 2012)	SCM (Zhong et al. 2014)	LOT (Oron et al. 2015)	RPT (Li et al. 2015)	CF2 (Ma et al. 2015)	DLT (Wang et al. 2013)
ASLA (Jia et al. 2012)	0.95	0.95	0.95	0.95	0.95	0.90	0.95	0.95	0.90	0.95
CSK (Henriques et al. 2012)	0.95	0.95	0.95	0.95	0.95	0.90	0.95	0.95	0.85	0.95
CNT (Zhang et al. 2016)	0.95	0.95	0.95	0.95	0.75	0.80	0.95	0.95	0.80	0.95
L1APG (Bao et al. 2012)	0.95	0.95	0.95	0.95	0.80	0.80	0.95	0.95	0.80	0.95
MTT (Zhang et al. 2012)	0.95	0.95	0.95	0.95	0.70	0.70	0.95	0.95	0.75	0.95

Table 4.5: Confidence level for the BC performance comparison

	STC (Zhang et al. 2014)	KCF (Henriques et al. 2015)	DCF (Henriques et al. 2015)	TLD (Kalal et al. 2010)	DFT (Sevilla-Lara & Learned-Miller, 2012)	L1APG (Bao et al. 2012)	MTT (Zhang et al. 2012)	LOT (Oron et al. 2015)	RPT (Li et al. 2015)	CNT (Zhang et al. 2016)	CF2 (Ma et al. 2015)	DLT (Wang et al. 2013)
ASLA (Jia et al. 2012)	0.95	0.95	0.95	0.95	0.95	0.95	0.80	0.95	0.95	0.95	0.95	0.85
CSK (Henriques et al. 2012)	0.95	0.95	0.95	0.95	0.95	0.95	0.80	0.95	0.95	0.95	0.95	0.80
SCM (Zhong et al. 2014)	0.95	0.95	0.95	0.95	0.95	0.95	0.80	0.95	0.95	0.95	0.95	0.80

Table 4.6: Confidence level for the SV performance comparison

4.2. Fusion of Tracking Results

The detailed tracking results of fusion are summarized in Table 4.7. Comparing with other fifteen tracking methods tested in this research study, the fusion method has shown better performance than top tracking methods, such as ASLA (Jia et al. 2012), SCM (Zhong et al. 2014), MTT (Zhang et al. 2012), L1APG (Bao et al. 2012), CSK (Henriques et al. 2012) and DFT (Sevilla-Lara & Learned-Miller 2012) in accuracy. It is noticed that the fusion methods has increased the tracking performance over 10% in AOS than the better tracking method in each sequence. On the other hand, the fusion method has not shown better performance in the robustness evaluation criteria. In most sequences, the fusion method ranks the top five performance in TL, while it ranks lower in some sequences

		ALSA (Jia et al. 2012)	CSK (Henriques et al. 2012)	CNT (Zhang et al. 2016)	CF2 (Ma et al. 2015)	DFT (Sevilla-Lara & Learned-Miller 2012)	DCF (Henriques et al. 2015)	DLT (Wang et al. 2013)	KCF (Henriques et al. 2015)	L1APG (Bao et al. 2012)	LOT (Oron et al. 2015)	MTT (Zhang et al. 2012)	RPT (Li et al. 2015)	SCM (Zhong et al. 2014)	STC (Zhang et al. 2014)	TLD (Kalal et al. 2010)	Fusion Method
Backhoe	AOS	0.82	0.79	0.80	0.76	0.78	0.74	0.85	0.72	0.81	0.77	0.82	0.74	0.80	0.76	0.69	**0.90**
	CER	0.03	0.02	0.02	0.06	0.04	0.04	0.03	0.06	0.03	0.22	0.03	0.06	0.03	0.03	0.14	**0.02**
	TL	0.35	0.40	0.41	0.41	0.41	0.51	0.25	0.51	0.39	1.00	0.36	0.41	0.38	0.26	1.00	**0.41**
Car	AOS	0.90	0.86	0.92	0.65	0.88	0.73	0.72	0.73	0.89	0.86	0.90	0.85	0.92	0.72	0.72	**0.95**
	CER	0.06	0.07	0.05	0.26	0.05	0.08	0.04	0.08	0.06	0.08	0.05	0.08	0.05	0.04	0.12	**0.03**
	TL	1.00	1.00	1.00	0.13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.71	**1.00**
Truck	AOS	0.85	0.81	0.81	0.75	0.82	0.65	0.73	0.65	0.81	0.86	0.82	0.81	0.85	0.76	0.77	**0.93**
	CER	0.10	0.15	0.11	0.25	0.13	0.10	0.09	0.11	0.11	0.14	0.13	0.18	0.10	0.11	0.21	**0.05**
	TL	0.87	0.57	0.85	0.33	0.64	0.36	0.25	0.45	0.77	0.80	0.72	0.48	0.93	0.43	0.36	**0.68**
Compactor	AOS	0.84	0.82	0.82	0.79	0.83	0.69	0.83	0.71	0.80	0.86	0.81	0.75	0.84	0.76	0.76	**0.92**
	CER	0.06	0.05	0.07	0.09	0.05	0.07	0.07	0.06	0.07	0.14	0.06	0.12	0.06	0.05	0.15	**0.03**
	TL	0.58	0.54	0.53	0.63	0.54	0.55	0.61	0.64	0.56	0.36	0.57	0.38	0.67	0.58	0.36	**0.61**
Excavator	AOS	0.82	0.85	0.81	0.86	0.80	0.81	0.80	0.81	0.83	0.81	0.83	0.76	0.83	0.72	0.77	**0.85**
	CER	0.10	0.08	0.08	0.12	0.17	0.16	0.05	0.14	0.10	0.11	0.09	0.19	0.09	0.09	0.18	**0.06**
	TL	0.74	0.86	0.84	0.71	0.63	0.86	0.47	0.71	0.74	0.72	0.74	0.71	0.70	0.51	0.58	**0.81**
Worker	AOS	0.85	0.83	0.74	0.80	0.82	0.72	0.55	0.73	0.81	0.78	0.82	0.78	0.80	0.72	0.74	**0.88**
	CER	0.13	0.10	0.17	0.14	0.10	0.11	0.20	0.12	0.10	0.12	0.10	0.12	0.12	0.09	0.17	**0.05**
	TL	0.72	0.57	0.60	0.76	0.65	0.58	0.14	0.55	0.64	0.57	0.70	0.56	0.62	0.60	0.30	**0.68**
Average	AOS	0.84	0.83	0.79	0.80	0.82	0.73	0.68	0.74	0.82	0.81	0.82	0.78	0.83	0.73	0.75	**0.91**
	CER	0.10	0.09	0.12	0.14	0.11	0.11	0.12	0.11	0.09	0.12	0.09	0.14	0.10	0.08	0.17	**0.06**
	TL	0.72	0.65	0.69	0.64	0.65	0.64	0.33	0.61	0.67	0.65	0.70	0.58	0.68	0.56	0.44	**0.68**

Table 4.7: Overall tracking performance with fusion method (the results of fusion method were identified with bold)

4.3. Part-based Tracking Results

In this study, the two parts based tracking was applied. The overall accuracy tracking method ASLA (Jia et al. 2012) was selected to track the part of cab and trails. Meanwhile, five tracking methods, ASLA (Jia et al. 2012), CNT (Zhang et al. 2016), CF2 (Ma et al. 2015), MTT (Zhang et al. 2012), and SCM (Zhong et al. 2014), which perform well in overall robustness criteria, have been selected to track the part of bucket. Then, the two parts tracking results have been combined together and named AS-ASLA, AS-CNT, AS-CF2, AS-MTT, and AS-SCM. These five part-based tracking methods have been applied in the Excavator sequences and compared with six tracking methods, which perform overall better in comparison parts. The comparison results have been indicated in the Table 4.8.

		ALSA (Jia et al. 2012)	CSK (Henriques et al. 2012)	DFT (Sevilla-Lara & Learned-	L1APG (Bao et al. 2012)	MTT (Zhang et al. 2012)	SCM (Zhong et al. 2014)	AS-ASLA	AS-CNT	AS-CF2	AS-MTT	AS-SCM
Excavator1	AOS	0.75	0.88	0.79	0.88	0.77	0.75	0.91	0.88	0.90	0.85	0.92
	CER	0.08	0.09	0.15	0.09	0.12	0.08	0.05	0.07	0.09	0.10	0.04
	TL	1.00	1.00	0.61	1.00	1.00	0.94	1.00	1.00	1.00	1.00	1.00
Excavator2	AOS	0.84	0.91	0.90	0.85	0.88	0.91	0.94	0.90	0.91	0.88	0.90
	CER	0.04	0.03	0.04	0.15	0.04	0.04	0.03	0.05	0.05	0.07	0.05
	TL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Excavator3	AOS	0.77	0.67	0.68	0.70	0.72	0.67	0.86	0.86	0.88	0.85	0.88
	CER	0.14	0.11	0.15	0.13	0.12	0.15	0.04	0.06	0.06	0.08	0.05
	TL	0.3	0.28	0.27	0.29	0.28	0.27	1.00	1.00	1.00	1.00	1.00
Excavator4	AOS	0.78	0.78	0.87	0.79	0.80	0.88	0.91	0.88	0.92	0.87	0.88
	CER	0.19	0.12	0.19	0.19	0.17	0.19	0.08	0.09	0.12	0.12	0.08
	TL	0.41	1.00	0.27	0.40	0.40	0.26	1.00	1.00	1.00	1.00	1.00
Excavator5	AOS	0.94	0.99	0.75	0.95	0.97	0.96	0.98	0.95	0.94	0.94	0.95
	CER	0.04	0.05	0.31	0.05	0.02	0.02	0.02	0.02	0.03	0.05	0.02
	TL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average	AOS	0.82	0.85	0.80	0.83	0.83	0.83	0.92	0.89	0.91	0.88	0.91
	CER	0.10	0.08	0.17	0.12	0.09	0.10	0.04	0.06	0.07	0.09	0.05
	TL	0.74	0.86	0.63	0.74	0.74	0.69	1.00	1.00	1.00	1.00	1.00

Table 4.8: Tracking performance of part-based methods

It was noticed that the single object tracking methods have shown reliable performance in tracking Excavator1, Excavator2, and Excavator5. In these three sequences, the excavators were moving from one place to another. In test video sequences of Excavator 3 and Excavator 4, the equipment were excavating and the performance of single object tracking methods have decreased a lot, especially in tracking length. The proposed part-based tracking methods have gained much better performance than single tracking methods in both robustness and accuracy. In fact, the AS-ASLA methods have achieved the better performance than others in tracking excavators in this study.

4.4. Discussions

In the visual tracking comparison section, those six methods with better overall tracking performances, four of them relied on the sparse representations, either local or holistic. On the other hand, the overall tracking performance for most of the methods without adopting the sparse representation strategy was not promising. Therefore, it indicated that the methods built upon the sparse representations might be more effective when tracking the targets in the construction scenarios. This may be because the sparse representation followed a natural, biological manner to simulate how human eyes capture objects; and the dictionary learning adopted in the sparse representation helped to build a low-rank, simplified representation model to describe a complex object in the scene (Zhang et al. 2015). As a result, most of the methods built upon the spare representations in this research study could effectively handle multiple challenging factors, such as illumination changes, occlusions, etc., when being tested to track construction targets.

Compare with the finding of computer vision community, it found that tracking methods that perform better in computer vision always perform well in construction sites. But there still has some difference. The ASLA showed the better performance under each attribute in construction sites, while this method did not perform so well in Motion Blur in the OTB benchmark. Also, the TLD method has shown good performance in OTB benchmark, and it performed as one of the worst methods in this study.

Also, it was found that most tracking methods based on discriminative classifiers did not perform well in the construction scenarios, although their effectiveness has been proved in the computer vision community. The reason behind is partly because the construction scenarios are always severely cluttered with materials, equipment, tools, and workers. They are more complex than the test scenarios of existing benchmarks in the computer vision community. Discriminative tracking methods mainly conducted the target tracking by differentiating it from the test video sequence background. The complexity in the construction scenarios affected the differentiative effectiveness. The SCM method adopted a hybrid framework of using generative and discriminative classifiers. As a result, its overall tracking performance was ranked 2nd in AOS, 5th in CER, and 4th in TL. In addition, the tracking methods that adopted deep learning did not perform well in this study; however, they still have huge potentials, especially when being used to track construction targets under severe occlusions. For example, it was found that the CF2 method could keep tracking the targets even when they were temporarily out of view for a short period of time.

In this study, a new tracking fusion method has been proposed, which is aiming to combine the tracking results of multiple tracking methods to produce the better results. This method is based on the attraction function, similarity function and non-maximum suppression. Moreover, it does not rely on any pre-processes and would be suitable for fusion any kinds of tracking methods. It could be found that this fusion methods have increased the accuracy performance over 10% for all testing sequences. Meanwhile, this fusion method have not shown a huge improvement in robustness. This may because that the fusion method is tend to close the majority tracking area and it is easy to fail once most of tracking methods fail.

The tracking of articulated construction equipment is still challenging in the construction scenarios. The video sequences adopted for the tests include excavators and backhoes. It was found that that most visual tracking methods under the tests could track the articulated equipment with the similar performance of tracking the un-articulated one, when the articulated equipment was just moving from one place to another on the construction site. For example, the excavators were moving without occlusions in the test video sequences of Excavator 1, Excavator 2, and Excavator 5. However, when an excavator is excavating, a lot of self-occlusions are produced due to its boom movements. These self-occlusions significantly impact the tracking performances, as shown under the evaluation criteria of tracking length. Take the test video sequences of Excavator 3 as an example. As for the video sequence of Excavator 3, there were no tracking methods that completed the tracking of the excavator in the whole video sequence. The highest TL was 0.3, achieved by the ALSA method.

The proposed part-based methods have significantly enhanced the excavator tracking performance in both robustness and accuracy in this study. In fact, this concept also could be used in tracking other equipment. The two parts methods can be changed to three, four or more parts in order to track more complex equipment and activities in construction. On the other hand, the single- object methods used in this study can be replaced with other better performed trackers and it is supposed to receive better results. However, there exist certain limitations for part-based tracking. When the target is divided into some parts, it is easier to lose the quickly moving part and results in the decrease of robustness. And the part-based method may not make breakthroughs in tracking occlusions because it cannot exceed the ability of original tracking method

CHAPTER 5: CONCLUSIONS AND FUTUER WORKS

Visual tracking technologies are becoming more and more important in the modern construction sites and could be used in productivity analysis, material tracking, and safety monitoring. However, few efforts have been put on evaluating the accuracy and robustness of these tracking methods in the construction scenarios. Also, there still remains problems that current tracking methods have not shown reliable performance in tracking articulated equipment, such as excavators, backhoes, and dozers etc.

This study proposed a comparative study of 2D visual tracking methods in the construction scenarios. A total of fifteen visual tracking methods from the computer vison community were selected for the comparison purpose. These methods were tested with twenty video sequences, which contain various construction resources of interest, including excavators, workers, backhoes, and compactors. All the video sequences were annotated manually to build the ground truths and characterized by the attributes (i.e. OCC, IV, BC, MB, and SV) to evaluate the tracking methods in quantitative and detailed manner. The paired two samples t-tests were also conducted to statistically validate the results. According to the quantitative comparison of tracking methods, two improvements were further conducted. One is to fuse the tracking results of individual tracking methods through the framework, which is based on the attraction function, similarity function and non-maximum suppression. The other improvement is to track the articulated equipment (excavators as an example) by the idea of part-based tracking.

The comparison results in this research study showed that the ASLA, SCM, MTT, L1APG, CSK, and DFT methods overall performed better than the others. Most of them relied on the sparse representations and generative classifiers to implement the visual tracking. The testing results of the fusion of tracking have indicated that the fusion method is a strong framework in enhancing the tracking accuracy and could be used to fuse the results of any tracking methods. In addition, the proposed part-based tracking methods have improved the tracking performance in both accuracy and robustness, when being used to track the articulated equipment.

Our future work will focus on extending the construction datasets and tracking methods in order to test more novel methods on more complex and challenging scenarios. Meanwhile, the part-based methods need to be tested on more videos which recoded from different views. On the other hand, it is noticed that the tracking systems are becoming diversified. And the difference between tracking and other vision-based technologies has been decrease. It should also be thought that how to combine other vision-based technologies in future tracking systems.

REFERENCE

Adam, A., Rivlin, E., & Shimshoni, I. (2006). Robust fragments-based tracking using the integral histogram. In Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on (Vol. 1, pp. 798-805). IEEE.

Auvinet, E., Reveret, L., St-Arnaud, A., Rousseau, J., & Meunier, J. (2008). Fall detection using multiple cameras. In 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 2554-2557). IEEE.

Amit, Y., & Trouvé, A. (2007). Pop: Patchwork of parts models for object recognition. International Journal of Computer Vision, 75(2), 267.

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. ACM Computing Surveys (CSUR), 43(3), 16.

Akatsuka, H., & Imai, S. (1987). Road signposts recognition system (No. 870239). SAE Technical P Avidan, S. (2004).

Support vector tracking. IEEE transactions on pattern analysis and machine intelligence, 26(8), 1064-1072.aper.

Bao, C., Wu, Y., Ling, H., & Ji, H. (2012). Real time robust l1 tracker using accelerated proximal gradient approach. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 1830-1837).

Bai, Y., & Tang, M. (2012, June). Robust tracking via weakly supervised ranking svm. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 1854-1861). IEEE.

Babenko, B., Yang, M. H., & Belongie, S. (2009, June). Visual tracking with online multiple instance learning. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 983-990). IEEE.

Babenko, B., Yang, M. H., & Belongie, S. (2011). Robust object tracking with online multiple instance learning. IEEE transactions on pattern analysis and machine intelligence, 33(8), 1619-1632.

Burl, M., Weber, M., & Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. Computer Vision—ECCV'98, 628-641.

Belshaw, M., Taati, B., Snoek, J., & Mihailidis, A. (2011, August). Towards a single sensor passive solution for automated fall detection. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 1773-1776). IEEE.

Black, M. J., & Jepson, A. D. (1998). Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. International Journal of Computer Vision, 26(1), 63-84.

Birchfield, S. T., & Rangarajan, S. (2005, June). Spatiograms versus histograms for region-based tracking. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 2, pp. 1158-1163). IEEE.

Cannons, K. (2008). A review of visual tracking. Dept. Comput. Sci. Eng., York Univ., Toronto, Canada, Tech. Rep. CSE-2008-07.

Carneiro, G., & Jepson, A. D. (2005, June). The distinctiveness, detectability, and robustness of local image features. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 2, pp. 296-301). IEEE.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. IEEE Transactions on pattern analysis and machine intelligence, 23(6), 681-685.

Coughlan, J., Yuille, A., English, C., & Snow, D. (2000). Efficient deformable template detection and localization without user initialization. Computer Vision and Image Understanding, 78(3), 303-319.

Crandall, D., Felzenszwalb, P., & Huttenlocher, D. (2005, June). Spatial priors for part-based recognition using statistical models. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 10-17). IEEE.

Comaniciu, D., Ramesh, V., & Meer, P. (2003). Kernel-based object tracking. IEEE Transactions on pattern analysis and machine intelligence, 25(5), 564-577.

Cehovin, L., Kristan, M., & Leonardis, A. (2014, March). Is my new tracker really better than yours?. In Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on (pp. 540-547). IEEE.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 1, pp. 886-893).

Daugman, J. G. (1980). Two-dimensional spectral analysis of cortical receptive field profiles. Vision research, 20(10), 847-856.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence, 34(4), 743-761.

D'Apuzzo, N. (2003). Surface measurement and tracking of human body parts from multi station video sequences. Institut für Geodäsie und Photogrammetrie an der Eidgenössischen Technischen Hochschule ETH Zürich.

Dick, A. R., Torr, P. H., Ruffle, S. J., & Cipolla, R. (2001). Combining single view recognition and multiple view stereo for architectural scenes. In Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on (Vol. 1, pp. 268-274). IEEE.

Dinh, T. B., Vo, N., & Medioni, G. (2011, June). Context tracker: Exploring supporters and distracters in unconstrained environments. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 1177-1184). IEEE.

El-Hakim, S., Beraldin, J. A., & Picard, M. (2002, September). Detailed 3D reconstruction of monuments using multiple techniques. In ISPRS/CIPA International Workshop on Scanning for Cultural Heritage Recording, Corfu, Greece (pp. 58-64).

Florack, L. M., ter Haar Romeny, B. M., Koenderink, J. J., & Viergever, M. A. (1994). General intensity transformations and differential invariants. Journal of Mathematical Imaging and Vision, 4(2), 171-187.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32(9), 1627-1645.

Fergus, R., Perona, P., & Zisserman, A. (2003, June). Object class recognition by unsupervised scale-invariant learning. In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on (Vol. 2, pp. II-II). IEEE.

Fu, M. Y., & Huang, Y. S. (2010, July). A survey of traffic sign recognition. In Wavelet Analysis and Pattern Recognition (ICWAPR), 2010 International Conference on (pp. 119-124). IEEE.

Fan, J., Xu, W., Wu, Y., & Gong, Y. (2010). Human tracking using convolutional neural networks. IEEE Transactions on Neural Networks, 21(10), 1610-1623.

Fitzgibbon, A., & Zisserman, A. (1998). Automatic camera recovery for closed or open image sequences. Computer Vision—ECCV'98, 311-326.

Ferrari, V., Tuytelaars, T., & Gool, L. V. (2003, June). Wide-baseline multiple-view correspondences. In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on (Vol. 1, pp. I-I). IEEE.

Gao, J., Ling, H., Hu, W., & Xing, J. (2014, September). Transfer learning based visual tracking with gaussian processes regression. In European Conference on Computer Vision (pp. 188-203). Springer International Publishing.

Guarnieri, A., Vettore, A., El-Hakim, S., & Gonzo, L. (2004). Digital photogrammetry and laser scanning in cultural heritage survey. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 35, B5.

Gosai, S. J., Kwak, J. H., Luke, C. J., Long, O. S., King, D. E., Kovatch, K. J., ... & Silverman, G. A. (2010). Automated high-content live animal drug screening using C. elegans expressing the aggregation prone serpin α1-antitrypsin Z. PloS one, 5(11), e15460.

Grabner, H., & Bischof, H. (2006, June). On-line boosting and vision. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on (Vol. 1, pp. 260-267). IEEE.

Grabner, H., Leistner, C., & Bischof, H. (2008). Semi-supervised on-line boosting for robust tracking. Computer Vision–ECCV 2008, 234-247.

Godec, M., Roth, P. M., & Bischof, H. (2013). Hough-based tracking of non-rigid objects. Computer Vision and Image Understanding, 117(10), 1245-1256.

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

Heckenberg, D. (2006, June). Performance evaluation of vision-based high DOF human movement tracking: a survey and human computer interaction perspective. In Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on (pp. 156-156). IEEE. Chicago

Han, S., & Lee, S. (2013). A vision-based motion capture and recognition framework for behavior-based safety management. Automation in Construction, 35, 131-141.

Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M. M., Hicks, S. L., & Torr, P. H. (2016). Struck: Structured output tracking with kernels. IEEE transactions on pattern analysis and machine intelligence, 38(10), 2096-2109.

Hoiem, D., Efros, A. A., & Hebert, M. (2008). Putting objects in perspective. International Journal of Computer Vision, 80(1), 3-15.

Hager, G. D., & Belhumeur, P. N. (1998). Efficient region tracking with parametric models of geometry and illumination. IEEE transactions on pattern analysis and machine intelligence, 20(10), 1025-1039.

Hong, Seunghoon, Tackgeun You, Suha Kwak, and Bohyung Han. "Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network." In ICML, pp. 597-606. 2015.

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

Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2012). Exploiting the circulant structure of tracking-by-detection with kernels. In European conference on computer vision (pp. 702-715). Springer Berlin Heidelberg.

Jia, X., Lu, H., & Yang, M. H. (2012). Visual tracking via adaptive structural local sparse appearance model. In Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on (pp. 1822-1829). IEEE.

Jahne, B. (Ed.). (2000). Computer vision and applications: a guide for students and practitioners. Academic Press.

Juang, C. F., & Chang, C. M. (2007). Human body posture classification by a neural fuzzy network and home care system application. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37(6), 984-994.

Kalal, Z., Matas, J., & Mikolajczyk, K. (2010). Pn learning: Bootstrapping binary classifiers by structural constraints. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 49-56). IEEE.

Koch, C., Jog, G. M., & Brilakis, I. (2012). Automated pothole distress assessment using asphalt pavement video data. Journal of Computing in Civil Engineering, 27(4), 370-378.

Koenderink, J. J., & van Doorn, A. J. (1987). Representation of local geometry in the visual system. Biological cybernetics, 55(6), 367-375.

Ke, Y., & Sukthankar, R. (2004, June). PCA-SIFT: A more distinctive representation for local image descriptors. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on (Vol. 2, pp. II-II). IEEE.

Kumar, B. R., Joseph, D. K., & Sreenivas, T. V. (2002). Teager energy based blood cell segmentation. In Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on (Vol. 2, pp. 619-622). IEEE.

Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Cehovin, L., Fernandez, G., & Pflugfelder, R. (2015). The visual object tracking vot2015 challenge results. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 1-23).

Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Cehovin, L., ... & Khajenezhad, A. (2013). The visual object tracking vot2013 challenge results. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 98-111).

Kristan, M., Kovacic, S., Leonardis, A., & Pers, J. (2010). A two-stage dynamic model for visual tracking. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 40(6), 1505-1520.

Kwon, J., & Lee, K. M. (2009). Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 1208-1215).

Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In IJCAI (Vol. 81, No. 1, pp. 674-679).

Leung, T., & Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. International journal of computer vision, 43(1), 29-44.

Lazebnik, S., Schmid, C., & Ponce, J. (2005). A sparse texture representation using local affine regions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1265-1278.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

Liebowitz, D., Criminisi, A., & Zisserman, A. (1999, September). Creating architectural models from images. In Computer Graphics Forum (Vol. 18, No. 3, pp. 39-50). Blackwell Publishers Ltd.

Lezoray, O., Elmoataz, A., Cardot, H., & Revenu, M. (1999). Arctic: An automatic cellular sorting system using image analysis. In Proceedings of vision interface (Vol. 99).

Lanigan, P. E., Paulos, A. M., Williams, A. W., & Narasimhan, P. (2006). Trinetra: assistive technologies for the blind. CyLab, 51.

Liu, C. L., Lee, C. H., & Lin, P. M. (2010). A fall detection system using k-nearest neighbor classifier. Expert systems with applications, 37(10), 7174-7181.

Li, Y., Zhu, J., & Hoi, S. C. (2015). Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 353-361).

Li, A., Lin, M., Wu, Y., Yang, M. H., & Yan, S. (2016). Nus-pro: A new visual tracking challenge. IEEE transactions on pattern analysis and machine intelligence, 38(2), 335-349.

Liu, S., Zhang, T., Cao, X., & Xu, C. (2016). Structural correlation filter for robust visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4312-4320).

Ma, C., Huang, J. B., Yang, X., & Yang, M. H. (2015). Hierarchical convolutional features for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision (CVPR) (pp. 3074-3082).

Mei, X., & Ling, H. (2009). Robust visual tracking using L 1 minimization. In 2009 IEEE 12th International Conference on Computer Vision (pp. 1436-1443).

Marĉelja, S. (1980). Mathematical description of the responses of simple cortical cells. JOSA, 70(11), 1297-1300.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. IEEE transactions on pattern analysis and machine intelligence, 27(10), 1615-1630.

Matthews, L., Ishikawa, T., & Baker, S. (2004). The template update problem. IEEE transactions on pattern analysis and machine intelligence, 26(6), 810-815.

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research, 31(13), 3812-3814.

Nawaz, T. and Cavallaro, A. (2013). A protocol for evaluating video trackers under real-world conditions. IEEE Transactions on Image Processing, 22(4), 1354-1361.

Nam, H., and Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4293-4302).

Oron, S., Bar-Hillel, A., Levi, D., & Avidan, S. (2015). Locally orderless tracking. International Journal of Computer Vision, 111(2), 213-228.

Ongun, G., Halici, U., Leblebicioglu, K., Atalay, V., Beksaç, M., & Beksaç, S. (2001). Feature extraction and classification of blood cells for an automated differential blood count system. In Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on (Vol. 4, pp. 2461-2466). IEEE.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, 24(7), 971-987.

Plantinga, W. H., & Dyer, C. R. (1986, October). An algorithm for constructing the aspect graph. In Foundations of Computer Science, 1986., 27th Annual Symposium On (pp. 123-131). IEEE.

Power, P. W., & Schoonees, J. A. (2002, November). Understanding background mixture models for foreground segmentation. In Proceedings image and vision computing New Zealand (Vol. 2002, pp. 10-11).

Pérez, P., Hue, C., Vermaak, J., & Gangnet, M. (2002). Color-based probabilistic tracking. Computer vision—ECCV 2002, 661-675.

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

Park, M. W., and Brilakis, I. (2012). "Construction worker detection in video frames for initializing vision trackers." Autom. Constr., 28, 15–25.

Rempel, D. M., Harrison, R. J., & Barnhart, S. (1992). Work-related cumulative trauma disorders of the upper extremity. Jama, 267(6), 838-842.

Rezazadeh Azar, E., & McCabe, B. (2012). Vision-based recognition of dirt loading cycles in construction sites. In Construction Research Congress 2012: Construction Challenges in a Flat World (pp. 1042-1051).

Rougier, C., St-Arnaud, A., Rousseau, J., & Meunier, J. (2011). Video surveillance for fall detection. INTECH Open Access Publisher.

Ramot, D., Johnson, B. E., Berry Jr, T. L., Carnell, L., & Goodman, M. B. (2008). The Parallel Worm Tracker: a platform for measuring average speed and drug-induced paralysis in nematodes. PloS one, 3(5), e2208.

Remondino, F., & El-Hakim, S. (2006). Image-based 3D modelling: a review. The Photogrammetric Record, 21(115), 269-291.

Ramoser, H., Laurain, V., Bischof, H., & Ecker, R. (2006, January). Leukocyte segmentation and classification in blood-smear images. In Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the (pp. 3371-3374). IEEE.

Schmid, C., & Mohr, R. (1997). Local grayvalue invariants for image retrieval. IEEE transactions on pattern analysis and machine intelligence, 19(5), 530-535.

Schneiderman, H., & Kanade, T. (2000). A statistical method for 3D object detection applied to faces and cars. In Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on (Vol. 1, pp. 746-751). IEEE.

Stenger, B., Mendonça, P. R., & Cipolla, R. (2001). Model-based 3D tracking of an articulated hand. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 2, pp. II-II). IEEE.

Sinha, N., & Ramakrishnan, A. G. (2003, October). Automation of differential blood count. In TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region (Vol. 2, pp. 547-551). IEEE.

Schörnich, S., Wallmeier, L., Gessele, N., Nagy, A., Schranner, M., Kish, D., & Wiegrebe, L. (2013). Psychophysics of human echolocation. In Basic Aspects of Hearing (pp. 311-319). Springer New York.

Sevilla-Lara, L., & Learned-Miller, E. (2012). Distribution fields for tracking. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 1910-1917).

Simonoff, J. (2008). Statistical analysis using Microsoft Excel. <http://people.stern.nyu.edu/jsimonof/classes/1305/pdf/excelreg.pdf> (April. 4, 2017)

Smeulders, A. W., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A., & Shah, M. (2014). Visual tracking: An experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7), 1442-1468.

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

Song, S., and Xiao, J. (2013). Tracking revisited using rgbd camera: Unified benchmark and baselines. In Proceedings of the IEEE international conference on computer vision (pp. 233-240).

Tuzel, O., Porikli, F., & Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In European conference on computer vision (pp. 589-600). Springer Berlin Heidelberg.

Torralba, A. (2003). Contextual priming for object detection. International journal of computer vision, 53(2), 169-191.

Tang, F., Brennan, S., Zhao, Q., & Tao, H. (2007, October). Co-tracking using semi-supervised support vector machines. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on (pp. 1-8). IEEE.

Tu, Z., & Bai, X. (2010). Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(10), 1744-1757.

Tekin, E., & Coughlan, J. M. (2009, December). An algorithm enabling blind users to find and read barcodes. In Applications of Computer Vision (WACV), 2009 Workshop on (pp. 1-8). IEEE.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. International journal of computer vision, 57(2), 137-154.

Van De Weijer, J., & Schmid, C. (2006). Coloring local feature extraction. Computer Vision–ECCV 2006, 334-348.

Van den Akker, J. M., Van Hoesel, C. P. M., & Savelsbergh, M. W. (1999). A polyhedral approach to single-machine scheduling problems. Mathematical Programming, 85(3), 541-572.

Varma, M., & Zisserman, A. (2002). Classifying images of materials: Achieving viewpoint and illumination independence. Computer Vision—ECCV 2002, 255-271.

Wang, N., & Yeung, D. Y. (2013). Learning a deep compact image representation for visual tracking. In Advances in neural information processing systems (pp. 809-817).

Wren, C. R., Azarbayejani, A., Darrell, T., & Pentland, A. P. (1997). Pfinder: Real-time tracking of the human body. IEEE Transactions on pattern analysis and machine intelligence, 19(7), 780-785.

Winlock, T., Christiansen, E., & Belongie, S. (2010, June). Toward real-time grocery detection for the visually impaired. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 49-56). IEEE.

Wang, N., Shi, J., Yeung, D. Y., & Jia, J. (2015). Understanding and diagnosing visual tracking systems. In Proceedings of the IEEE International Conference on Computer Vision (CVPR) (pp. 3101-3109).

Wang, N., Li, S., Gupta, A., & Yeung, D. Y. (2015). Transferring rich feature hierarchies for robust visual tracking. arXiv preprint arXiv:1501.04587.

Weerasinghe, I. T., & Ruwanpura, J. Y. (2009). Automated data acquisition system to assess construction worker performance. In Construction Research Congress (pp. 61-70).

Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 2411-2418).

Wu, Y., Lim, J., & Yang, M. H. (2015). Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9), 1834-1848.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., & Gall, J. (2013). A survey on human motion analysis from depth data. In Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications (pp. 149-187). Springer Berlin Heidelberg.

Yin, Y., Zhang, X., Williams, R., Wu, X., Anderson, D. D., & Sonka, M. (2010). LOGISMOS—layered optimal graph image segmentation of multiple objects and surfaces: cartilage segmentation in the knee joint. IEEE transactions on medical imaging, 29(12), 2023-2037.

Yang, J., Arif, O., Vela, P. A., Teizer, J., and Shi, Z. (2010). Tracking multiple workers on construction sites using video cameras. Adv.Eng. Inf., 24(4), 428–434.

Yang, J., Vela, P., Teizer, J., and Shi, Z. (2014). “Vision-based tower crane tracking for understanding construction activity.” Journal of Computing in Civil Engineering, 10.1061 / (ASCE) CP.1943-5487.0000242, 103–112.

Zhang, T., Ghanem, B., Liu, S., & Ahuja, N. (2012). Robust visual tracking via multi-task sparse learning. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 2042-2049).

Zhang, K., Liu, Q., Wu, Y., & Yang, M. H. (2016). Robust Visual Tracking via Convolutional Networks Without Training. IEEE Transactions on Image Processing, 25(4), 1779-1792.

Zhang, K., Zhang, L., Liu, Q., Zhang, D., & Yang, M. H. (2014). Fast visual tracking via dense spatio-temporal context learning. In European Conference on Computer Vision (pp. 127-141). Springer International Publishing.

Zhong, W., Lu, H., & Yang, M. H. (2014). Robust object tracking via sparse collaborative appearance model. IEEE Transactions on Image Processing, 23(5), 2356-2368.

Zhu, Z., Ren, X., & Chen, Z. (2016). Visual Tracking of Construction Jobsite Workforce and Equipment with Particle Filtering. Journal of Computing in Civil Engineering, 30(6), 04016023.