School of Design and the Built Environment

Automatic Scaffolding Productivity Measurement through Deep Learning

Wenzheng Ying

0000-0003-1759-8769

This thesis is presented for the Degree of

Doctor of Philosophy

of

Curtin University

August 2021

(left blank)

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Human Ethics (For projects involving human participants/tissue, etc) The research presented and reported in this thesis was conducted in accordance with the National Health and Medical Research Council National Statement on Ethical Conduct in Human Research (2007) – updated March 2014. The proposed research study received human research ethics approval from the Curtin University Human Research Ethics Committee (EC00262), Approval Number #HRE2019-0797

Signature:

Date:

(left blank)

Acknowledgements

I would like to express my sincere appreciation to my supervisors Prof. Xiangyu Wang and Prof. Changzhi Wu. They guided and supported me to walk through the whole procedure of my research work step by step. Especially, Prof. Xiangyu Wang not only encouraged me to explore the knowledge and techniques beyond my understanding at that time, but also organized basketball games among research team members so that I was able to face and overcome these research challenges with an excellent physical condition.

I would also like to thank my research colleagues: Dr. Junxiang Zhu, Dr. Yongze Song, Prof. Peng Wu, Dr. Tao Zhou, Weixiang Shi, Chengke Wu, Dr. Honglin Chi and others. They were always ready for help and gave me selfless assistance with the problems I came across.

The work was financially provided for by KAEFER Integrated Services and the Industry Doctoral Training Centre (IDTC) program of Australian Technology Network (ATN). Additionally, the Australasian Joint Research Centre for Building Information Modelling facilitated experiments and provided essential equipment and contacts for construction site visits.

Professional editor, John McAndrew, provided copyediting and proofreading services, according to the guidelines laid out in the university-endorsed national "Guidelines for Editing Research Theses".

Last but not the least, I would like to thank my family: my parents Bin Ying and Jie Xu, who are my backbones and support me spiritually everyday. My wife Guo Xing takes good care of me and you always tell me to be self confident and I can conquer the tough time.

Contents

Abstract

Scaffolding provides temporary support for personnel and materials to access high spaces for multiple construction purposes, such as installation, maintenance, and inspection. Scaffolding is regarded as a preliminary step before the start of the main construction tasks, is crucial to construction, and is closely linked with other construction trades. However, in Australia, scaffolding suffers from low productivity and high costs. One solution to this problem is timely and precise productivity reporting or feedback on labour and devices in onsite operations, which enables the management team to closely monitor ongoing conditions, and, thus, rapid response and correction become possible and expensive delays can be avoided.

This research develops two approaches to automatic measure scaffolding productivity by using video frames and static images respectively. This first approach employs onsite video cameras to capture scaffolder's activity and recognize and classify scaffolder's activity in real-time. Relevantly, activity analysis is defined as a continuous procedure of assessing and improving the amount of time that craft workers spend on specific construction trades, and, according to the Construction Industry Institute (CII), workface assessment is described as the first and fundamental step for the execution of activity analysis. Following on from this, a method of automatic scaffolding workface assessment is proposed by using 2D video camera to capture scaffolding activities and the deep learning model of key joints and skeleton extraction as well as machine learning classifiers were used for activity classification. Additionally, a case study was conducted and showed that the proposed method is a feasible and practical way for automatic scaffolding workface assessment. Consequently, a scaffolding indicator was derived from the results of scaffolding workface assessment. The results of the case study demonstrated the feasibility and robustness of the approach.

This second approach is proposed to detect and scaffolding structure from static images and extract scaffolding intersections (couplers and wedges) from scaffolding structures. Couplers and wedges are regarded as one of main scaffolding components and they present in a form of intersections in an image of scaffolding structure. By using the techniques of computer vision and deep learning, this approach extracted the scaffolding intersections and count the total number of the intersections in an image.

A mathematical model was built to connect the scaffolding intersections with the volume of a scaffolding project. Then the volume of scaffolding project and the productivity were derived by taking the scaffolding project design parameters and the total number of intersections into calculation. A case study of six scaffolding projects was conducted and the object detector YOLOv4 presented good performance and stability in object detection and productivity measurement.

This research not only demonstrated the feasibility and efficiency of vision-based approaches for the extraction of semantic information, but also enabled timely reflection of scaffolding productivity before the completion of a construction, and it released managers from manual observation and inspection in construction monitoring.

Keywords: activity recognition, construction monitoring, computer vision, deep learning, machine learning, scaffolding.

1 Background and motivation

1.1 Chapter introduction

Section 1.2 in this chapter provides a brief background on scaffolding work and its high importance in cost and time management in the construction industry. In Section 1.3, specific problems about productivity measurement and the status quo of scaffolding work are described in detail, highlighting the motivation for this research. Section 1.4 illustrates four objectives that this research plans to achieve as well as the scope of the research. Section 1.5 presents the thesis structure and offers an overall flowchart for this study.

1.2 Background

Construction corporations' profits have largely reduced due to increasing competition in the industry; however, project owners' demands are increasing both in the level of quality and efficiency (Hu, Chong, and Wang 2019).The smooth progress of the construction is an essential part of construction technologies improvement. When project delays or reworks occur due to unsound construction management, they can result in a reduction in profits and credibility loss. In the oil and gas industry, there is a huge demand for scaffolding work during the construction and maintenance stage of a liquefied natural gas (LNG) plant. In order to keep the LNG plants running smoothly and safely, regular scaffolding erection and dismantlement are conducted for periodical inspections and parts replacement (Moon et al. 2016). Currently, scaffolding work is often designed as a preliminary step before the start of main construction tasks, and it is also strongly connected with other construction trades (Hou et al. 2017). The delay of scaffolding projects would result in budget increase and project delay. Moreover, scaffolding suffers from low productivity and a high cost in Australia. Interviews conducted with 56 construction contractors show that scaffolding represents one of the 16 types of most wasteful components of indirect construction cost, and its expenditure accounts for about 12-15% of overall project cost (Kim and Teizer 2014, Hou et al. 2017). Thus, it can clearly be seen that timely and successful delivery of scaffolding work is a vital start to an entire construction project, and it is also a crucial element in progress and cost management.

Project monitoring and controlling (PMC) is a management process that monitors and controls an as-built project by periodically following and inspecting the project's indices, such as progress, cost, workload, productivity, and so on (Singh, Gu, and Wang 2011, Chong, Lee, and Wang 2017, Zhu et al. 2019, Zhu et al. 2020). PMC is regarded as a fundamental and key aspect in construction operations (Omar, Mahdjoubi, and Kheder 2018). Continuous and effective monitoring and assessment of productivity, progress, and quality are core tasks in PMC, and they significantly impact upon the project's ultimate success. PMC provides the opportunity to be aware of ongoing as-built status and identify as-built progress delay and it assists in launching amendments (Hu, Chong, and Wang 2019). Flawless PMC can ensure that projects are in accord with budget and schedule (Golparvar-Fard et al. 2011). However, the process of PMC has not been fully automated and there is a lack of widely accepted industry benchmarks for accurate measurement (Bosché et al. 2014). In the last decade, there has been a huge demand for accurate and fast automated monitoring methods in the industry of architectural engineering construction (AEC) (Li et al. 2018, Wu, Wang, and Wang 2016). Aiming to meet this demand, researchers have been exploring automated methods from the aspects of site resources tracking and as-built data modelling and interpretation.

1.2.1 Site resources tracking

A construction site is a multidimensional space that contains multiple resources: personnel, vehicles, materials, and other facilities, etc, and these resources are coordinated to ultimately complete the construction projects. Accordingly, a crucial element of PMC is identifying the status of these resources for the sake of safety and performance purposes. Accordingly, various sensing and tracking technologies are utilized for the localization and status identification of site resources in this research domain. Rebolj et al., pointed out that the visibility and status tracking of construction materials plays an important role in project delivery and introduced an automated subsystem for material tracking (Rebolj et al. 2008). Memarzadeh et al. investigated an automated approach for 2D detection of site personnel and equipment from video streams (Memarzadeh, Golparvar-Fard, and Niebles 2013). Gong and Caldas proposed a video interpretation model that can automatically extract productivity data and illustrated a case study on the tracking and analysis of the column pour process through

detecting and distinguishing concrete bucket status (Gong and Caldas 2010). Other positioning technologies, including Bluetooth, radio frequency identification (RFID), ultra wideband (UWB), and global positioning system (GPS), can also facilitate onsite resource tracking. Bluetooth is a technical standard of wireless telecommunication, which can achieve short-distance data exchange among fixed and mobile devices. It has the advantages of low cost, low energy consumption, and minimum infrastructure requirement. Park et al. tested the performance of Bluetooth technology applied as a proximity alert based on spatial distance for personnel and equipment in a work zone (Park et al. 2016). Fang et al. discussed that site resource localization is a central concern of applications of RFID in the construction domain (Fang et al. 2016). Cheng et al. evaluated the performance of commercial UWB devices for resource location tracking under a harsh construction environment (Cheng et al. 2011). A GPS system employs satellites for absolute localization, and it can provide the latitudinal and longitudinal information of objects installed with a GPS receiver, and GPS systems have been widely utilized in outdoor environments: vehicle navigation, underground and mine construction, and other industries. Generally, with the requirement of higher accuracy, the cost of the GPS sensor is higher. However, GPS weakens or ceases to function when the signal is blocked by high buildings or by terrain.

Except for resource tracking, the technologies described above have been studied in subdivision fields that reflect site resource status, for instance, site environment monitoring (Hughes, Yan, and Soga 2015), proximity alert (Marks and Teizer 2013), and supply chain management (Wang, Hu, and Zhou 2017).

1.2.2 As-built data modelling and interpretation

To accomplish PMC, another crucial task of modern construction management is to analyse as-built data from a macro perspective, especially targeting the project's progress control. Through scientific modelling and semantic interpretation, raw data captured from sites can be effectively transformed into construction progress indicators and researchers have devoted their efforts to this domain. Turkan et al. described a novel system combining visible data from site laser scans with as-planned 3D building information models into a project progress tracking system (Turkan et al. 2012). The potentials of 3D reconstruction from point cloud have also been explored and tested (Golparvar-Fard et al. 2011). Furthermore, researchers have been exploring

image-based automated analysis for progress monitoring. Image-based progress monitoring is designed to calculate or evaluate the quantity of finished construction components or the level of completion by analysing the still images captured on construction sites. To effectively and accurately gain quantitative results, this approach has three essential requirements for the image capture. First, the cameras should be located in relatively high positions where visual interruptions are minimized, and an entire view of the construction project is included. Second, it is necessary for cameras to possess the ability of mobility so that engineers or onsite managers can freely adjust the cameras' shooting angles and locations or zoom in and out to observe ongoing construction activities. Third, it must be feasible for the data transmission between cameras and the device of image processing for tracking and analysing purposes. Rapid and effective transmission can be achieved by wireless or wired Internet connections linked to a central server.

Image-based progress monitoring relies on effective feature extraction from site images. Effective feature extraction contributes to a large extent to the success of progress recognition. Kim et al. utilized many different image processing techniques, such as noise removal and colour filter, to extract a main construction structure for construction process tracking as a crucial aspect of the as-built modelling (Kim, Kim, and Kim 2013). Hung-Lin Chi et al. investigated scaffolding progress monitoring by integrating building information modelling (BIM) and image processing techniques, and they detected grid lines from site images and broke lines into segments to match scaffolding tubes. Further, they evaluated the quantity of scaffolding and then integrated the results from image processing with digital BIM model to track project progress. However, their image processing approach performed less effectively because it took only the visible external layer of scaffold into account and ignored the geometrical volume of scaffold (Chi et al. 2017).

Consequently, as-built models together with as-planned schedules or models are compared in this research to identify the discrepancy in actual construction progress, assisting project decision making and amendment.

1.2.3 Few studies on temporary structures

Temporary structures on construction sites are those structures that are established and utilized to assist with permanent construction projects. They function as a protective facility and offer a platform for access and support. When permanent structures are completed or consolidated, temporary structures need to be dismantled or removed. Common temporary structures include scaffolding, roadway decking, concrete platform, cranes, and others. The implementation of temporary structures presents a primary impact on the quality, efficiency, profitability, and safety of all ranges of construction projects (Ratay 2012). Despite a large number of studies in automated PMC, temporary structures have attracted limited attention in research, which is not accordance with their importance in construction.

Research interests were limited to the fields of safety, scheduling and planning for temporary structures in construction. Cranes were investigated in the topics of safety, scheduling and planning (Tam and Fung 2011, Kang, Chi, and Miranda 2009, Yang et al. 2014), because there are high risks of material falling and crane collisions. Some scholars took the scaffolding as the representative of temporary structures (Kim, Cho, and Zhang 2016, Kim and Cho 2015).

Scaffolding is one of the most common types of temporary structures, and it has been widely utilized among residential, commercial, and industrial constructions. Scaffolding provides personnel and materials with a temporal platform at a certain height for the purpose of installation, maintenance, inspection, and so on (Hou et al. 2017). In the oil and gas industry in Western Australia, there is a huge demand for scaffolding work during the construction and maintenance stage of a liquefied natural gas (LNG) plant, which is a type of mega-sophisticated gas processing facilities. In order to keep the LNG plants running smoothly and safely, regular scaffolding erection and dismantlement are conducted for periodical inspections and part replacement (Moon et al. 2016).

Research focuses of scaffolding were placed on the planning and design (Kim, Cho, and Zhang 2016, Kim and Teizer 2014, Hou et al. 2014), scheduling (Hou et al. 2017) and object recognition (Xu et al. 2018, Bangaru et al. 2021, Chi et al. 2017). Scaffolding productivity is a research area which has not been fully explored.

Productivity is a direct and frequently utilized performance indicator to reflect the efficiency and delivery speed of construction work, and it is also a popular research topic in the construction industry because of its crucial importance to the success of construction projects (Yi and Chan 2014) Timely and precise productivity reporting or feedbacks on labour and devices in onsite operations enable the management team to maintain a sharp awareness of ongoing conditions and subsequently respond quickly so that appropriate correction can be launched and expensive delays avoided.

1.3 Problem statement

Labour productivity is regarded as a key indicator in assessing the success of construction projects. The labour cost accounts for 30% to 50% of the total cost of a construction project worldwide. Thus, construction labour productivity directly influences a project's profitability. Effective monitoring and improvement on construction labour productivity becomes a crucial concern for construction companies control the cost and pursue profits. The systematic productivity monitoring and analysis of construction activities can help the management team to enhance working performance and quickly respond to project delays (Van Tam et al. 2021). The current methods of productivity and progress monitoring for scaffolding work include project level information systems, direct observation methods, and surveys or interviews; however, they mainly depend on manual data collection and observation. As a case study, in Darwin, Australia, the manager of a scaffolding company regularly spent a large number of man-hours in observing and recording the progress of dismantling scaffolds. Obvious drawbacks in manual monitoring exist: 1) the results of productivity measurement are subjective and susceptible to human errors and bias; 2) it requires project inspectors to qualify themselves with professional knowledge and experience; 3) it cannot be widely applicable as it lacks an objective standard for measurement and estimation; and 4) it is time-consuming and not cost effective due to the rising labour cost. To address this drawback of manual implementation, a large number of research in the past decade has explored the methods that can automate the process of productivity monitoring and measuring. These methods can be generally divided into non-vision-based approaches and vision based approaches. Non-vision based approaches reply on different types of sensors such as RFID, UWB and GPS to identify and track the workers' locations. However, interpreting the workers' activities

and productivities merely based on location information, without analysing their activities is imprecise and challenging. Vision based approaches capture and analyse the workers' activities through images or videos. Depth images and crowdsourcing from video streams have been explored for construction activity analysis (Liu and Golparvar-Fard 2015, Khosrowpour, Niebles, and Golparvar-Fard 2014). However, depth images perform well under indoor environments, but it is challenging to implement depth sensors at outdoor construction sites. Additionally, crowdsourcing would continuously generate extra cost and is only applicable at an experiment level.

Despite the critical role that scaffolding plays in construction operation, scaffolding is a research topic in construction management that has not been deeply explored. Previous studies on scaffolding have largely concentrated on project scheduling and logistic optimisation but have barely paid attention to scaffolding productivity measurement because of its sophisticated and dynamic features.

Motivated by the aforementioned practical challenges in actual construction operations, this study is trying to address the research gap on scaffolding productivity measurement and to develop an automatic approach with the help of ordinary cameras. This study is different from previous research, as ordinary RGB cameras are chosen to capture data that can provide both RGB images and videos in real-time. Compared to retrieving location-based information, ordinary cameras can provide rich content images and videos for activity analysis and productivity measurement. Furthermore, ordinary cameras have been widely used in construction industry due to their inexpensive nature and they are able to capture clear RGB data under outdoor construction environments. Deep learning and computer vision algorithms are implemented in the model developed in this research. Instead of manually collecting and measuring scaffolding productivity, this study provides construction managers with accurate numerical results of scaffolding productivity, and, hence, to release project inspectors or managers from repetitive and time-consuming inspections.

1.4 Scope and objectives

With the rapid growth of information technology, the development of artificial intelligence (AI), increased computing power and high-speed information

transmission, machine learning, and deep learning are gradually revealing their utility and power.

This research aims to develop an automatic system to measure the productivity of scaffolding work through the vision-based approach using videos and images as inputs. Due to it being cost-effective and providing rich content, video surveillance is popular, and many construction sites have installed surveillance for safety and management purposes. Following this adoption, the digital data in this study is confined to RGB images and videos captured via ordinary RGB cameras or video surveillance on construction sites. Consequently, this research using RGB data as input, is designed to combine and adopt the techniques, including computer vision, machine learning and deep learning for automatic recognition, and interpretation and productivity measurement. The four specific objectives of the research are as follows:

1) To establish a vision-based database of scaffolding erection, which involves onsite data collection on different sites to support the training and testing steps of machine learning.

2) To develop a vision-based system by integrating open-source computer vision structures and different kinds of computer vision algorithms for reliable recognition and classification.

3) To build a theoretical connection between the targets recognised by the vision-based system and the productivity of scaffolding work through literature reviews and experiment.

4) To form an indicator that reflects the productivity of scaffolding work by implementing algorithms and calculations and to test its validity with real results.

1.5 Thesis structure

The thesis consists of five sections: introduction, literature review, methodology, experiment and validation, and conclusion and discussion, as shown in Figure 1.1.

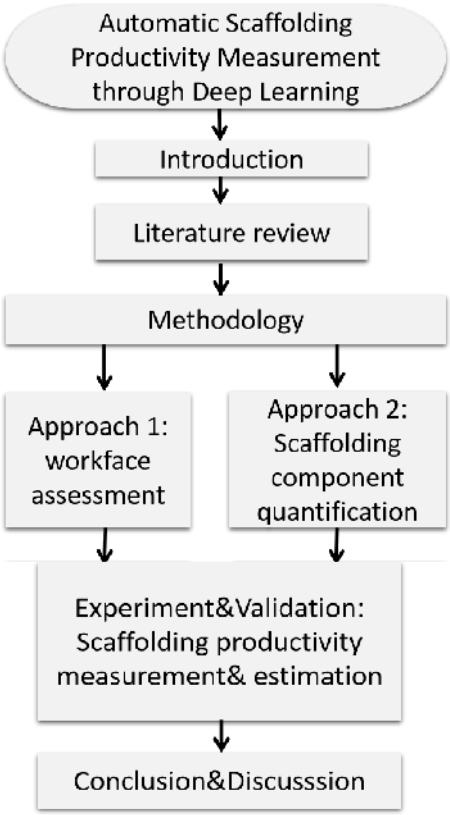


Figure 1.1. Thesis structure

First, in Chapter 1, the introduction has illustrated the current status of scaffolding work in the construction industry and the problems that have motivated this thesis as well as presenting the scope and objectives of this research. Secondly, the literature review presents in Chapter 2: 1) the definition of productivity and the relevant research on construction labour productivity; 2) the state of art of workface assessment and activity analysis; 3) the state of art of image-based construction progress monitoring; and 4) the review of vision-based research areas in construction workers' activity. Chapter 3 briefly introduces the research framework of the thesis. Chapter 4 introduces the first approach of automatic workface assessment and its process of experiment and validation for scaffolding productivity measurement. Chapter 5 illustrates the second approach of scaffolding component quantification as well as the experiment and validation of scaffolding productivity estimation. Chapter 6 provides the conclusion and discussion of this thesis.

2 Literature review

2.1 Chapter introduction

In this chapter, Section 2.2 first introduces different definitions and calculation methods of productivity and then reviews the studies on construction labour productivity (CLP) at activity level. In addition, workface assessment as an initial stage for activity analysis is introduced and reviewed. Section 2.3 introduces the workface assessment in construction activity and its state of the art. Section 2.4 reviews the three main research interests in construction workers' activity: ergonomic posture, safety management, and productivity improvement. In section 2.5, the studies of vision-based construction component recognition are reviewed.

2.2 Construction labour productivity

2.2.1 The definition of productivity

Productivity is a term deriving from economics that denotes the efficiency of production. The *Oxford Advanced Learner's Dictionary* defines productivity from three key points: first, the ability of being productive; second, the rate at which goods or service are produced; and third, the rate indicates how well resources such as materials, time, labour, or money are spent (Dictionary 2000). Also, the term “productivity” is used to describe the conversion rate from the aggregate inputs to the associated output. Following this general idea, construction productivity can be regarded as a quantitative measure of the conversion from a combination of inputs to outputs. It can be expressed and analysed in two forms: 1) total factor productivity, which measures the relationship between outputs and all kinds of inputs involved; and 2) partial factor productivity, which only focuses on the relation of one single input to the outputs. Since the construction industry involves a large number of various construction trades and the workforce is one of the most important factors in construction operations, construction productivity relies on labour performance and productivity to a large extent. Consequently, construction labour productivity (CLP) is viewed as an excellent index for the management team to evaluate project performance (Jarkas 2010).

There are many different definitions of CLP that evaluate the construction efficiency from diverse perspectives. Firstly, outputs in a period of time, usually hourly or daily, are commonly used as a form of CLP in construction research; a labour hour is utilized as the input and the physical quantity of construction work can be selected as the outputs according to the construction trade (Sonmez and Rowings 1998, Hanna et al. 2008). For example, scaffolding erection uses the labour hour as the input and the cubic meters of scaffold as the output. Additionally, concrete placement uses the same input unit and the cubic meters or yards of concrete as the output (Eastman and Sacks 2005). As illustrated in Equations (2.1) and (2.2), CLP can be expressed in either way

$$\mathrm{CLP} = \frac{\text{Output}}{\text{input}} = \frac{\text{Finished workload}}{\text{Unit time}} \tag{2.1}$$

$$\mathrm{CLP} = \frac{\text{Input}}{\text{Output}} = \frac{\text{Work time}}{\text{Installed quantity}} \tag{2.2}$$

Equations (2.1) and (2.2) are the reciprocal of each other. For the convenience of comparison, or when the numerator is relatively small, the calculation of CLP is switched between (2.1) and (2.2) or, instead, the denominator is increased.

There are two challenges in productivity measurement at the construction project level. First, different construction trades use their associated units for measurement. For example, welding is measured in the unit of inch/hour and steel placement is calculated in meter/hour, while scaffolding erection is measured in cubic meter/hour. Thus, no comparability exists between various construction activities. The other challenge occurs when variances exist between job characteristics in the same construction activity. For example, for concrete placement, the average productivity level of pouring walls is higher than the one for pouring columns (Yi and Chan 2014). To tackle these challenges, construction professionals, including the American Association of Cost Engineers, introduced another definition of productivity which is based on the comparison between the real productivity and an absolute standard, reflecting a relative productivity level (Allmon et al. 2000). However, absolute norm or benchmark of productivity is prone to becoming outdated due to new emerging

techniques. Thus, performance ratio (PR) has been introduced by management teams as a ratio of actual productivity compared to expected productivity; the equation is illustrated in Equation 2.3.

$$\text{Performance Ratio(PR)}_{\text{i,t}} = \frac{\text{Actual Productivity}_{\text{i,t}}}{\text{Expected Productivity}_{\text{i,t}}} \tag{2.3}$$

where i denotes the activity in a construction project; t denotes workday under consideration; an expected productivity is determined by taking an activity standard and workload, such as workdays and finished quantities, into consideration (Thomas and Yiakoumis 1987).

The expected productivity can be adjusted due to bad weather conditions or disruption. PR is a unitless ratio determined by dividing the actual productivity by a benchmark productivity, and it allows comparison among different construction activities. If the PR value is greater than 1, it means that the actual activity outperforms the expected productivity; if the PR value is less than 1, it shows that the actual productivity is inferior to the expectation. One of the advantages of the PR method is that work time is not an essential factor to be measured, so that the progress level and performance can be determined by the completed work quantity.

Peddi et al. classified the instances of one specific construction activity into 3 categories: effective work, contributory work and ineffective work. Then the real-time productivity of a work was computed as (see Equation (2.4)) below (Peddi et al. 2009).

$$P = \frac{T_{effective} + T_{contributive}}{T_{effective} + T_{contributive} + T_{ineffective}} \tag{2.4}$$

where $T_{effective}$: Time of effective activities in scaffolding

$T_{constributive}$: Time of contributory activities in scaffolding

$T_{ineffective}$: Time of ineffective activities in scaffolding

At industry or company level, economists calculate CLP from a macroeconomic perspective (Park, Thomas, and Tucker 2005). The U.S Bureau of Labor Statistics (BLS) utilized Equation (2.5) to calculate the labour productivity in different industries.

This approach only takes account of labour hours consumed, excluding holidays and sick leave but including paid or unpaid overtime. Wen and Albert described a similar approach which adopted gross product originating (GPO) as the numerator (see Equation (2.6)). GPO is an economic concept referring to overall construction output at industry level, and it is commonly calculated in chained dollars in order to eliminate the inflation effect (Yi and Chan 2014).

$$\text{Labour Productivity} = \frac{Q_t}{Q_0} \div \frac{T_t}{T_0} \tag{2.5}$$

where $Q = \text{quantity of output}$

$T = \text{total labour hours}$

$t = \text{the current year}$

$0 = \text{the base year}$

$$CLP = \frac{GPO}{\sum_1^{12} E_i H_i} \tag{2.6}$$

Where GPO = gross product originating of construction industry is calculated in chained dollars; E_i= the average number of labours in month; H_i= average work hours in month i.

Different definitions of CLP share the common factor that they are all designed to either measure the quantitative amount of the efficiency of one type of construction activity or increase comparability at the project level or industry level among various activities or industries. Since scaffolding belongs to a particular construction trade, the CLP of scaffolding involved in this study will concentrate on its definition at the activity level to quantitatively reflect the performance and production rate of the scaffolding activity.

2.2.2 CLP analysis and evaluation at the activity level

Many studies have been devoted to CLP analysis and evaluation at the activity level. Before the step of CLP modelling and evaluation, it is critical to investigate the factors influencing CLP. A better understanding of the factors influencing CLP at construction

activity level will assist project planners and managers to distribute project resources more efficiently and will enhance the budget control, project scheduling, resource management, and many other aspects. AbouRizk, Knowles and Hermann analysed 27 projects and 39 pipe installation activities and investigated 33 factors influencing CLP, including crew size, installed quantity, material type, and location (AbouRizk, Knowles, and Hermann 2001). Herbsman and Ellis divided CLP influence factors into two main groups: technological factors and organizational factors. Technological factors include specification factors, design factors, location factors, and material factors, while organizational factors contain production factors, labour factors and social factors (Herbsman and Ellis 1990). Different construction activities, including rigging pipes, welding, steel erecting and fixing, formwork, and concrete pouring were taken as research subjects (Ezeldin and Sharara 2006, Fayek and Oduba 2005).

As discussed, CLP is influenced by a variety of factors. Moreover, the effect of these factors on different construction activities varies from one to another. For example, bad weather conditions have a greater impact on outdoor construction activities, such as concrete pouring. Thus, for CLP estimation purpose, researchers have attempted to develop several different models using various modelling techniques to quantitively study the relationship between the relevant factors and CLP. Generally, four main model techniques, consisting of statistical analysis, regression analysis, expert systems, and artificial neural networks (ANN), have been utilized for CLP analysis and evaluation.

Using regression analysis, Srinavin and Mohamed studied the impact of a thermal environment on the CLP of three construction activities: painting, brick laying, and excavation, representing light, moderate, and heavy construction task types, respectively (Srinavin and Mohamed 2003). Thomas and Yiakoumis collected and combined CLP data from masonry, formwork, and structural steel and implemented statistical analysis to discover the relationship between temperature and CLP (Thomas and Yiakoumis 1987). These studies mentioned above only investigated one single factor affecting CLP, while there are also some cases discussing the impacts of multiple factors on CLP. For example, Thomas and Sakarcan established a factor model taking three factors into account to forecast the CLP. In addition, in terms of

statistical and regression analysis, a significantly large size of datasets are usually required for model establishment (Thomas and Sakarcan 1994).

Expert systems function like decision trees, which are able to classify or predict numerical results, but expert systems can also process linguistic descriptions from construction experts or professionals. For example, Christian and Hachey created an expert system for concrete pouring (Christian and Hachey 1995). Also, Fayek and Oduba developed a fuzzy expert system to forecast the CLP of welding and pipe rigging, but the system can only produce qualitative results for CLP prediction. (Fayek and Oduba 2005).

An artificial neural network was a type of algorithm first introduced in the 1990s and was inspired by the mechanism and structure of the biological neural networks in a human brain, where the signals are received, transmitted, and emitted back and forth through multiple layers of neurons. Ezeldin and Sharara estimated the CLP of form assembly, concrete pouring, and steel fixing by using ANN, and they collected over 70 datasets for each construction activity, including the factors crew size, work duration, and work quantity (Ezeldin and Sharara 2006). Abourizk et al. proposed a two-stage ANN to predict the CLP of welding and pipe installation (AbouRizk, Knowles, and Hermann 2001). An ANN model is a more powerful tool with considerable potential to deal with the quantitative prediction of the effects of multiple factors. Compared to statistical and regression analysis, ANN has more parameters for parameter tuning to optimize its prediction performance, whereas statistical and regression analysis provides a more parsimonious and straightforward model, and regression models enable the user to choose the linear or quadratic relationships to be used in modelling. CLP modelling at the activity level faces a challenge in that it requires a large quantity of historical statistical data for training purposes whatever modelling technique is used, and it is usually difficult to collect historical statistical datasets for different construction activities. Additionally, with the revolution of new construction technologies, the performance and efficiency of construction activities keeps updating, which becomes another challenge, and the outdated statistical datasets are not able to precisely reflect or estimate current CLP.

2.3 Workface assessment

Since scaffolding is a particular construction activity with its own particular features, the CLP of scaffolding at the activity level is chosen to be investigated. Activity analysis has proven to be a feasible and practical approach for monitoring onsite operation and for analysing the conditions causing delays or productivity decline (Gouett et al. 2011). Activity analysis is defined as a continuous procedure of assessing and improving the amount of time that craft workers spend on one specific construction trade. This amount of time is referred to as direct work time. Direct work activity is the activity that construction workers directly place physical effort towards. In 2010, the Construction Industry Institute (CII) proposed a detailed guideline for activity analysis. According to CII, workface assessment is described as the first and fundamental step for the execution of activity analysis (CII 2010). Figure 1. shows an example of workface assessment, which indicates the distribution of activity categories where a period of time is spent. Workface assessment aims to timely reflect construction productivity before the release of cost and progress reports. It is a practical procedure for measuring the activity rates of construction workers throughout a long period of time, which relies on a professional onsite supervisor as an observer who determines and classifies the activities executed by construction workers. An example of workface assessment is shown in Fig 2.1. Unfortunately, the current approach to workface assessment remains that of manual inspection and it faces many challenges: first, manual data collection takes extra labour to observe, record, and analyse the operation, which is not cost effective due to high rising labour cost; second, close manual observation may result in an abnormal reflection of workers' performance caused by the Hawthorne effect, whereby the workers adjust their behaviours or productivity as a result of the awareness of being observed; third, manual interpretation takes repetitive and random observations and a long period of time, and the results heavily rely on the supervisor's own experience (Gouett et al. 2011, Khosrowpour, Niebles, and Golparvar-Fard 2014).

ACTIVITY ANALYSIS TABLE						
Name of Observer	XXX					
Date	23/01/2018					
Project Name						
Trades Observed	SCAFFOLDING					
Observation Records	Round 1		Round 2		Round 3	
Time Range	9AM	10AM	11AM	12AM	2PM	3PM
Direct Work	~~IIII~~ II		~~IIII~~		~~IIII~~ III	
Prep Work			III		III	
Tools and Equipment					I	
Travelling	III		IIII		II	
Waiting/Idling	II		II		IIII	
Comments						

Figure 2.1. An example of a work assessment form

As discussed, then, current workface assessment takes time-consuming and labour-intensive manual observation. To address these limitations, efforts, including different machine learning algorithms to automize this process, have been made by scholars in the past few years. Cheng et al. utilized both location information and worker's body posture for automated activity analysis at the task level. Their method combined the data obtained from UWB for location tracking and the body posture data from a wearable 3-axial accelerometer (Cheng et al. 2013). However, this approach merely used one single location and body posture as the training sample to infer each activity category. Also, activity recognition between two activities that include intra-class variation and inter-class similarity at the same location would be challenging. Khosrowpour, Niebles and Golparvar_Fard proposed a method based on depth cameras for activity recognition and applied a bag-of-poses histogram and hidden Markov model algorithm to distinguish each activity (Khosrowpour, Niebles, and Golparvar-Fard 2014). However, depth cameras are sensitive to sunlight and a large amount of construction activities are located outdoors. Thus, their method is more feasible for interior construction operations, but it may not be applicable to the external construction environment, such as scaffolding.

Human activity recognition (HAR) performs a core and fundamental procedure in this research as it allows automation of the workface assessment process. Human activity recognition employs machines to comprehend and categorize a series of human activities from different data sources. Based on the type of data source, HAR can be

divided into either a sensor-based approach or a vision-based approach. The former exploits data from wearable sensors and other sensors. The sensor is attached to human limbs and other body parts, or the surroundings close to human activity, to continuously collect the target's activity. Sensor-based HAR produces one-dimensional data, such as time series signals. Accelerometer, gyroscope, and magnetometer are three common wearable sensors, which detect and recognize user's activities by analysing the signal deviation when different activities are performed. Other sensors such as radio frequency identification (RFID), global positioning system (GPS) and ultra-wideband (UWB) also can be integrated into smart phones, wristbands, helmets, or safety vests and provide a user's trajectory, absolute, or relative locations to infer the user's activity. However, sensor-based HAR suffers from either high cost or sensitivity to external environment. For instance, a gyroscope is relatively expensive and there is a need to attach at least two on key joints for HAR; also, an accelerometer is sensitive to temperature. In addition, most of these sensors need to be powered by batteries.

The vision-based approach collects human activities with the help of visual devices, such as video surveillance or camera, and generates multi-dimensional data, including 2D images, 3D images, or video frames. Although this approach releases participants from wearing sensors and the image processing technique develops, it relies heavily on visual data quality (Assadzadeh et al. 2021). Elements include the image resolution and illumination condition, which to some extent affect the robustness of recognition. However, vision-based devices have been widely used in construction and other industries for safety management purposes. The cost of surveillance cameras is relatively low compared to other sensors.

2.4 Vision-based research interests in construction workers' activity

The integration of high-resolution cameras, increased capability of data storage, and the availability of high-speed telecommunications over the past decade has allowed rich content from onsite construction operations to be observed. Today, cameras have been broadly applied for contractors and owners to monitor ongoing construction activities. With the help of computer vision algorithms, scholars have been able to research areas in construction workers' activities, mainly in the areas of ergonomics

posture, safety management, and productivity improvement. This section reviews the state-of-the-art techniques for activity recognition and the application domains in the construction industry which researchers have mainly focussed upon.

Thanks to the rapid development of sensor technologies in recent years, automatic HAR has received a lot of attention and has been widely explored in many domains. As discussed above, the technologies for HAR can generally be divided into sensor-based approach and vision-based approach.

Based on sensor modality the sensor-based approach can be classified into three types: body-worn sensors, environment sensors, and object sensors. Body-worn sensors are often attached to the human body to detect the movement of the body by continuously recording the signal data. Accelerometers, gyroscopes, and magnetometers are the three most frequently used body-worn sensors. Their applications focus on the activities of daily living (ADL) and sports. Environment sensors measure the interaction between a human and the environment. For instance, pressure sensors, temperature sensors, and sound sensors can measure their corresponding environmental parameters. Environment sensors can be used to sense hand gestures and user's daily activities, and they have been used in smart home development. Object sensors are installed on objects which are close to human movements in order to infer human activity from the detection of object movement. For example, drinking activity can be detected by installing an accelerometer on a cup. GPS or RFID modules are attached to worker's helmets to monitor the user's location so as to infer a worker's activity status (Sarafianos et al. 2016, Dang et al. 2020).

Ryu et al. proposed to recognize worker's activities by using wristband IMU and conducted a case study of masonry work (Ryu et al. 2019). Bangaru, Wang and Aghazadeh assessed the reliability of wearable IMU and EMG for HAR in construction (Bangaru, Wang, and Aghazadeh 2020). Bangaru et al. established an ANN based model for the classification of scaffolder’s work by using data retrieved from IMG and EMG sensors (Bangaru et al. 2021). However, these two studies faced technical challenges as the models were trained with at-rest activities and as a result only standardized and repeated movements could be effectively identified and recognized and transitional signals or actions were excluded or could not be correctly

distinguished. Our approach takes image sequences of each activity as the input data, which includes both typical actions and transitional actions to address this limitation.

According to the data type, the vision-based approach can generally be divided into RGB data and RGB-D data. RGB data refers to the images consisting of red, green and blue colour bands in the spectrum, and RGB data can be collected through ordinary cameras or video surveillance. RGB-D data is produced by RGB-D cameras which can not only capture the original RGB data but also collect depth information. Dang et al. pointed out that RGB data has the advantages that it is extensively available as well as affordable and allows the capture of rich content for the subjects (Dang et al. 2020). Compared to RGB data, RGB-D data provides depth information, which can enhance the performance of HAR. The disadvantages of RGB-D data include computation complexity as well as high costs. Scholars have explored and developed several models for both RGB data and RGB-D data and tested these models with public or benchmark datasets (Sarafianos et al. 2016, Dang et al. 2020).

2.4.1 Ergonomics posture

Ergonomic issues have long been a concern in manual construction work. Since manual work in construction often requires human labour to perform repetitive activities with a heavy workload, working in awkward posture can result in fatigue, injuries, or severe accidents. Researchers have been exploring automatic methods of recognizing ergonomic posture and have generated many assessment guidelines which focus on work-related musculoskeletal disorders (WMSDs). For example, the Rapid Upper Limb Assessment (RULA) and Ovako Working Posture Analysing System (OWAS) are followed by most researchers for the definition and classification of manual activities and postures (Fığlalı et al. 2015, Rahman 2014). Ray and Teizer investigated a method of real-time posture analysis for ergonomics training by using depth cameras (Ray and Teizer 2012). Yan et al. explored the potential of ergonomic posture recognition via ordinary 2D cameras (Yan, Li, Wang, et al. 2017). Zhang, Yan and Li analysed the joint angles from a 3D skeleton generated by multi-stage convolutional neural networks (CNNs) in order to recognize movements of body parts (Zhang, Yan, and Li 2018). Yan et al. collected and processed motion data to warn of hazardous movement patterns related to the head, neck, and trunk with the help of wearable inertial measurement units (IMUs), a type of physical sensor attached to body

limbs or the trunk (Yan, Li, Li, et al. 2017). Seo et al. combined depth sensors and an own-designed system for biomechanical analysis (Seo et al. 2015).

2.4.2 Safety management

Additionally, researchers have made efforts to address safety concern about worker's construction activities. Violation of operational rules about safety management might cause serious equipment damage or even fatal accidents. Exploring an automatic system for warning of unsafe activities relieves the burden of onsite supervisors and offers timely detection of potential hazards. Yu et al. selected three unsafe behaviours: leaning on handrails, dumping from height, and ladder climbing and then measured the skeleton angles generated from a depth sensor to identify unsafe behaviours in an experimental environment (Yu et al. 2017). Alwasel et al. presented a framework to identify the productive and safe poses of masons by implementing video cameras and IMUs attached to two control groups and a machine learning algorithm support vector machine was employed as a classifier (Alwasel et al. 2017). Han, Lee and Peña-Mora focused on ladder climbing, a specific construction activity with considerable risk, and managed to map the body joints onto a 3D space by capturing and extracting motion data from a depth sensor system (Han, Lee, and Peña-Mora 2013).

2.4.3 Productivity improvement

A few studies have aimed to explore construction workers' activity analysis so as to achieve productivity improvement. Luo, et al. integrated RGB image streams, optical flow streams, and grey image streams and trained them separately with CNNs to achieve workforce activity recognition (Luo et al. 2018). Peddi et al. proposed a system for measuring construction worker's productivity through analysing worker poses and classifying worker activity into three domains: effective, inefffective, and contributory work (Peddi et al. 2009). Khosrowpour et al. employed depth cameras for activity analysis and developed a bag of poses activity classifier and a hidden Markov model (HMM) for classification (Khosrowpour, Niebles, and Golparvar-Fard 2014). Calderon, Roberts and Golparvar-Fard synthesized pose sequences for the vision-based activity analysis of an excavator (Torres Calderon, Roberts, and Golparvar-Fard 2021). Activity analysis of a construction device is conducted by inferring the device status from images. Regarding activity analysis of construction workers, it is more

sophisticated and challenging to extract useful features from images. Apart from the vision-based approach, the mechanical approach, which basically attaches different mechanical sensors to the human body to capture the signals of body and limb movement, can also be regarded as a useful tool for motion capture (Guraliuc et al. 2011). Passive RFID and IMUs, including gyroscopes and magnetometers, have been studied as practical tools (Amendola, Bianchi, and Marrocco 2015). Joshua and Varghese studied activity recognition in construction by using accelerometers (Joshua and Varghese 2011). Although magnetic sensors have also been adopted, they are susceptible to metal surroundings, so are not suitable for scaffolding work (Aloui, Villien, and Lesecq 2015). To increase accuracy, Alwasel et al. combined IMUs with video cameras to study masons' productivity (Alwasel et al. 2017). The mechanical method has proven feasible in a lab environment; however, it may face difficulties being widely applied to construction sites due to the inconveniences caused by the attachment of sensors to the body (e.g., when wearing and washing clothing).

2.5 Vision-based Construction component recognition

Automated construction component recognition mainly relies on vision-based computer technologies to detect and recognize structural components. It facilitates construction progress and productivity monitoring, which plays an important role in PMC. Timely recognition of discrepancies between as-built data and as-planned information allows project participants to make amendments and adjustments to minimize both financial and time loss. Current practices of progress and productivity monitoring largely depend on periodic manual inspection and daily reports of onsite supervisors, which is costly, time consuming, and error-prone (Kopsida, Brilakis, and Vela 2015, Ekanayake et al. 2021). To automate the inspection process, construction component recognition is regarded as an indispensable step, and there is strong relevant research in this field.

According to the methods of retrieving data, one popular method in construction component recognition is to capture as-is data or as-built data using laser scanning and a stereo camera. The collected data provides point cloud in a 3D coordinate system where every cloud point is located in x, y and z coordinates. Son and Kim introduced a method of 3D structural component recognition and modelling that collected 3D data

via a stereo camera (Son and Kim 2010). Wang and Kim concluded that the data of a 3D point cloud are mainly utilized in two applications: 3D reconstruction and geometry quality inspection (Wang and Kim 2019). The targets for 3D reconstruction include construction facilities and vehicles, indoor components, earthwork surfaces, and the whole building. Xu et al. investigated the reconstruction of scaffold from the point cloud data collected on construction sites (Xu et al. 2018). Chae et al. explored the modelling of earth surfaces by using a 3D laser scanner (Chae et al. 2011). Although laser scanning can precisely capture every point of a target construction component, this method requires a large amount of time and computational power for the formation of construction components. For instance, it took up to 7 hours to generate a single column (Golparvar-Fard et al. 2011). Furthermore, the method of laser scanning is relatively expensive and its technical difficulties lie on the accurate separation of the point cloud, because these points are disorderly scattered and contain limited object-related information (Almukhtar et al. 2021).

Another popular approach to collecting data for the recognition is to utilize 2D images and videos. Compared to laser scanning, cameras and video surveillance are inexpensive and convenient to use. However, retrieving data from construction site images, which can be incomplete and noisy, is a difficult problem (Trucco and Kaka 2004). A simple approach that uses computer vision methods, is to compare a sequence of images from a fixed camera and find the differences in the construction process to estimate the progress (Lukins and Trucco 2007, Ibrahim et al. 2009). However, these methods have a limited success rate, and they are not fully automated. Automated detection and identification of building elements according to shape and materials have been proposed using image processing techniques (Zhu and Brilakis 2010a, b, Brilakis, Soibelman, and Shinagawa 2005). Texture, colour, and shape information has been used to classify construction materials, such as concrete and steel (Zhu and Brilakis 2010a, Brilakis and Soibelman 2008) and to detect and count the number of bricks on a façade (Hui, Park, and Brilakis 2014). Window detection (Yang and Tian 2010, Stoeter, Le Mauff, and Papanikolopoulos 2000) algorithms have also been developed. Multiple views geometry for retrieving the 3D reconstruction of building structures has also been presented (Roh et al. 2009, Son and Kim 2010).

Regarding the construction components, research objects that were set as targets and captured ranged from construction terrain, including geological and earthwork models, building elements, such as columns, doors and windows, and infrastructure components, including roads, bridges, and pipelines (Son, Kim, and Kim 2015). Trucco and Kaka developed a framework for the component recognition of doors and windows by comparing the Hausdorff distance between the regional grey levels in two 2D images (Trucco and Kaka 2004). Lukins and Trucco presented a prototype model for detecting the deviation of concrete columns in 2D images (Lukins and Trucco 2007). Wu and Kim proposed an approach for detecting concrete columns by implementing the algorithms of computer vision (Wu and Kim 2004). Narazaki et al. developed a multi-scale CNN method to automatically detect a bridge's structural components (Narazaki et al. 2020). Hamledari, McCabe and Davari established a computer vision-based model for the automatic detection of indoor partitions, including studs, insulation, electrical outlets, as well as the status of drywall sheets (Narazaki et al. 2020). Zhu and Brilakis explored the potential of automatically identifying concrete regions by integrating computer vision and machine learning technologies (Zhu and Brilakis 2010b). In addition, Lee and Park developed a machine learning based approach to effectively extract and count the number of reinforcing bars from an image (Lee and Park 2019). Also, some studies utilized a support vector machine and other classifiers to segment the reinforcing bar area on a conveyor belt (Liu, Li, and Liu 2015, Nie, Hung, and Huang 2016).

2.6 Chapter summary

In conclusion, from the perspective of activity analysis, current research focused on sensor-based approaches and vision-based approaches. Sensor-based approaches have their limitations including: 1) can only providing location-based information to infer ongoing construction activities; 2) many sensors need to be attached on human body parts. Vision-based approaches extract construction activities from RGB data or RGB-D data. The capture of RGB-D data is vulnerable to sunshine, so RGB-D data is feasible to indoor construction environment, but it is not applicable to outdoor construction sites. From the direction of construction component recognition, current practices through point cloud are time consuming and require huge computational power. Object recognition from 2D images still focused on simple construction components such as windows and drywall sheet. This study proposed to develop a model by using RGB images and videos to automate productivity measurement of scaffolding activity.

3 Research framework

In this chapter, an overall research framework for automatic scaffolding productivity measurement is introduced, which can be applied to the scaffolding inspection or scaffolding progress monitoring.

As described in our research scope above, this research mainly focuses on the data analysis and interpretation of vision data captured in scaffolding procedure. Scaffolding is regarded as an important preliminary step in a construction project. As shown in Figure 3.1, a conventional scaffolding project starts from the design and scheduling stage drafted by professional planners and managers who take construction's needs, safety, and related regulations into consideration. Having been approved, a scaffolding project moves on to its erection stage conducted by scaffolders. Our proposed process can be integrated into the current inspection process, and it assists site inspectors to conduct workface assessment and scaffolding component quantification for scaffolding productivity and progress monitoring. Scaffolders continue their work if the feedback from our proposed method meet the as-planned criteria, such as productivity and schedule; if not, the urgent problem is forwarded to a periodic progress meeting to discuss and evaluate whether delays can be addressed

at the work task level. If work task amendment can be made, for instance, by reallocating available site resources, actions and decisions are made by the site manager and then sent to scaffolders for execution, so, assisting timely progress. If a major deviation of progress is the issue, the project manager is required to redesign and reschedule the project at the project-level with the planners.

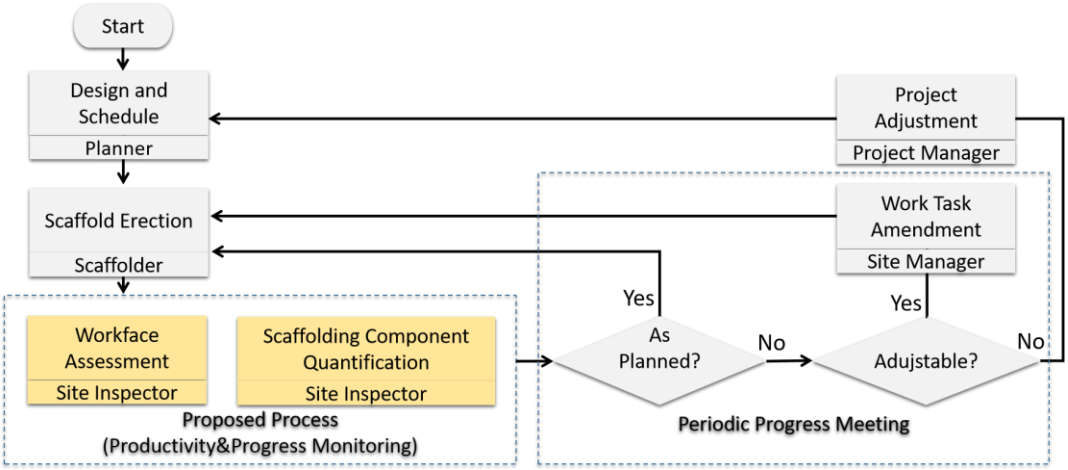


Figure 3.1. Roles of the proposed productivity monitoring in the existing scaffolding workflow

This research proposes to intake either static photos or dynamic video data captured in scaffolding procedure as input data. The overall workflow of data collection and processing is displayed in Figure 3.2. Video surveillance is placed onsite to collect scaffolders' activities in real-time, and through semantic interpretation this first workflow enables managers and inspectors to obtain scaffolders' work statuses and to conduct workface assessment. Static site images are collected and updated daily or weekly, and then, with the assistance of computer vision and deep learning, automatically detecting, estimating, and quantifying the main scaffolding components are proposed as a further step. Site inspector can select either approach based on onsite circumstances and conditions. If an onsite camera is close enough to clearly capture a scaffolder's movement, automatic workface assessment is more applicable. If an onsite camera is fixed and captures scaffolding project from more than 7 meters, scaffolding component quantification is more suitable to use. Thus, these static and dynamic information flows generate as-built information of scaffolding, which facilitates onsite productivity and progress monitoring.

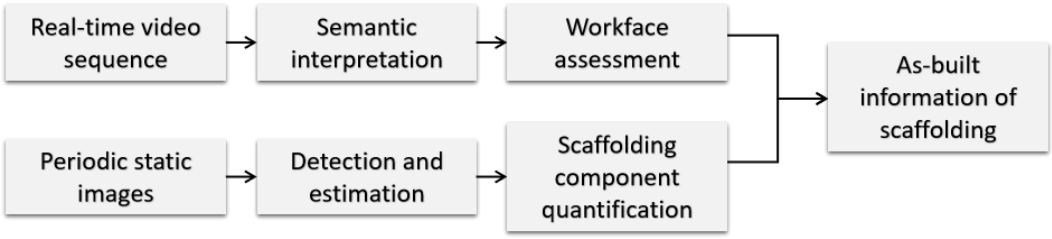


Figure 3.2. Overall workflow of data collection and processing

In the following chapters, these two technical workflows would be separately illustrated in detail.

4 Automated Workface Assessment of Scaffolding

4.1 Chapter introduction

This section illustrates the research design main procedures employed, as shown in Figure 4.1. To recognize scaffolding activities, it is essential to define these activities in advance. According to the principles of activity analysis, scaffolding activities are divided into three categories. For the data collection, an ordinary 2D camera was adopted as our video capture device. In the feature extraction step, the key joints of the human body were used as the feature to be extracted from every video frame. To decrease the impact of intra-class variation and inter-class similarity, a model for 3D pose estimation was proposed. Then, discriminative classifiers were trained with our annotated data to detect and recognize scaffolding activities. The part of the case study, and validation, is described in the next section.

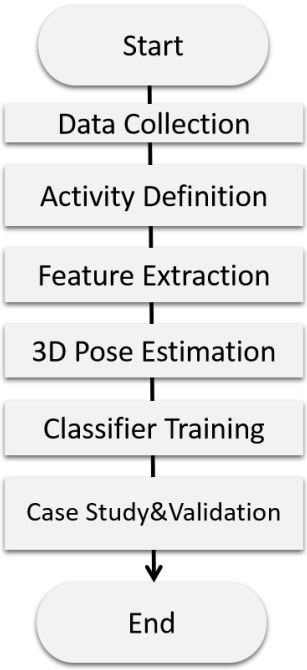


Figure 4.1. Research design of automatic workface assessment

4.2 2D ordinary video camera for data collection

Using ordinary video camera for human motion capture has many advantages over other data capture approaches. Video surveillance employs ordinary video cameras and provides 2D videos to monitor and record activities, and it has been broadly installed for management purposes in a large variety of industries. Compared to IMUs, an ordinary camera is non-intrusive: there is no requisite for devices to be attached to a worker's body. Also, a depth sensor works properly under indoor conditions, but it is sensitive to an outdoor environment in that it produces a lot of noise when it get exposed to outdoor radiation (Zhang, Yan, and Li 2018). Compared to a depth sensor, an ordinary video camera is more cost efficient and performs reliably on outdoor construction sites (Zhang, Yan, and Li 2018). Furthermore, video contains rich and intuitive information, which can not only be applied for monitoring, recording, and safety management, but also can be broadly used for teaching purpose and accident analysis. To this end, an ordinary video camera was employed in this research as the human motion capture device for workface assessment in construction (Zhou et al. 2019).

Investigating human action estimation from video frames captured by 2D ordinary camera, scholars have explored several principle approaches, such as appearance-based (Ali and Shah 2008), trajectories-based (Sheikh, Sheikh, and Shah 2005), volume-based (Jiang, Drew, and Li 2006), and interest-point-based (Liu, Ali, and Shah 2008). As human body joint points contain rich meaningful information for action estimation, the approach based on the extraction of key joint points has been widely studied in the field of computer vision and has become the state of art method for human activity recognition. Hence, the extraction of human key joint points and body skeleton estimation were adopted in our research for workface assessment in construction.

This study proposed a vision-based method using 3D skeleton point estimation and supervised machine learning for workface assessment from video sequences. For activity definition, in accordance with the guideline of workface assessment,the scaffolding operation is divided into three categories: direct work, essential contributory work, and ineffective work. For model validation, video data from actual scaffolding operations is collected to train and test our model.

4.2.1 Intra-class variation

Ordinary cameras only provide a 2D video stream without depth information so that one major challenge in human activity recognition from 2D cameras is view invariance, which includes intra-class variation and inter-class similarity (Zhang, Yan, and Li 2018). As regards intra-class variation, homogeneous activities might be identified as different activities as this class of activities are recorded and viewed from different directions. As regards inter-class similarity, heterogeneous activities may share similarities from certain viewpoints. For instance, in Figure 2(a)(b) inter-class similarity occurs where the walking posture presents similar skeleton appearance in 2D with the transporting posture. In Figure 2(b)(c), intra-class variation occurs where the same activity transporting is performed but filming from different viewpoints makes the variant skeletons. A robust human activity recognition method should have the capability to not only distinguish activities of different classes but also tolerate the intra-class variations in one homogeneous activity.

(a) (b) (c)

Figure 4.2. Inter-class similarity and intra-class variation. 2(a) and 2(b) represent walking and transporting two different activities, while their skeleton appearances show similarity from the same viewpoint. 2(b) and 2(c) display the same activity, while their skeleton appearances present variation from different viewpoints.

Compared with a 2D model, a 3D human skeleton or joint positions can provide depth data from one more dimension to effectively decrease view variance and assist in the

process of discrimination and classification. More detailed information can be provided by 3D features, which enables the process of action recognition more reliable and feasible. Thus, to enhance accuracy and robustness, this study utilized the extraction of 3D human skeleton from 2D video frames captured by ordinary camera for workface assessment in construction.

4.3 Activity definition

Scaffolding operation is sophisticated and dynamic and scaffolding involves a series of activities and various body postures. However, it can be analysed and simplified into repetitive sequences of individual activities in accordance with the process of workface assessment. For example, Khosrowpour et al. took interior drywall operation as a case study and divided the operations into seven categories (Khosrowpour et al. 2014). In line with the principles of activity analysis and with the assistance of interviews from onsite scaffolders and scaffolding supervisors, the scaffolding operation was analysed and categorized into three sections below, shown in Figure 4.3.

1. Direct work: the real process of contributing to a unit being constructed (Gouett et al. 2011, Thomas and Daily 1983). Additionally, from the perspective of lean construction, direct work is the process that adds value to a construction work, which employers are willing to pay for (Shou et al. 2017). For scaffolding operations, the part of scaffold erecting is direct work.

2. Essential contributory work: not the direct set up but the activities necessary to establish the construction unit. This category involves transporting materials and tools, receiving instructions, necessary communication between co-workers, and so on. One typical essential contributory work for scaffolding work is scaffold transporting (Peddi et al. 2009).

3. Ineffective work: the activities that contribute nothing to production probably due to inefficient material or labour supply and poor communication. These representative activities include idling or waiting (Peddi et al. 2009).

These three categories are applied in this study to investigate scaffolding operations.

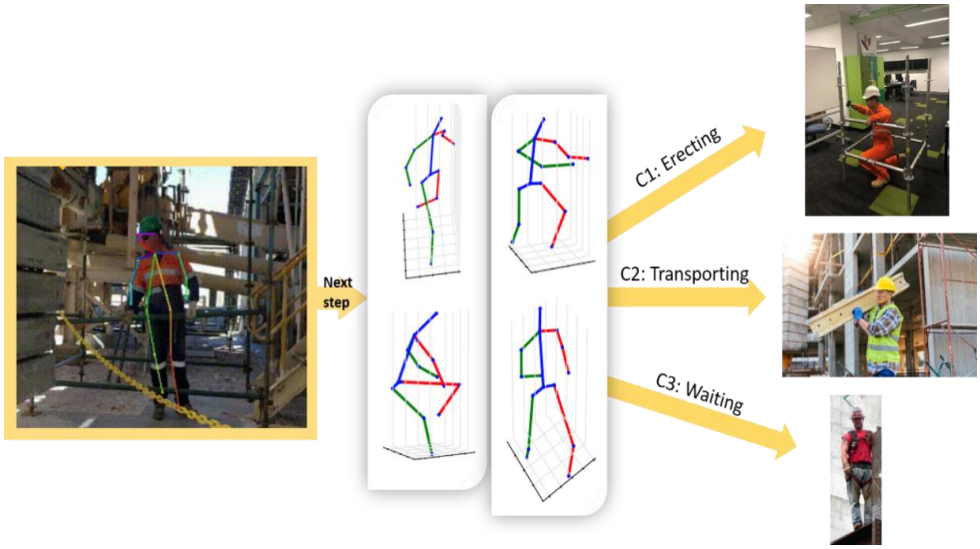


Figure 4.3. Scaffolding activity categories. Scaffolding activities are classified into erecting (direct work), transporting (essential contributory work) and waiting or idling (ineffective work).

4.4 Key joints extraction

This section illustrates the whole structure of 3D key joints extraction and posture estimation. In order to automatically recognize a scaffolder's posture under working conditions, the OpenPose system is integrated with the 3D joint estimation model to extract 3D skeletons and key joints from 2D video frames. OpenPose is a multi-stage CNN system and represents the state of art real-time model extracting human body key points from 2D video frames (Cao et al. 2019).

As shown in Figure 4.4, a human skeleton consisting of 18 joints was extracted by the implementation of the OpenPose system: a two-branch convolutional neural network (CNN). Part confidence maps were produced by the first branch throughout Stage 1 and Stage t. Part affinity fields (PAFs) were generated by the second branch for limb association from Stage 1 and Stage t. Then the combination of part confidence maps and PAFs was parsed by the greedy algorithm to predict the 2D key points of human body in the image.

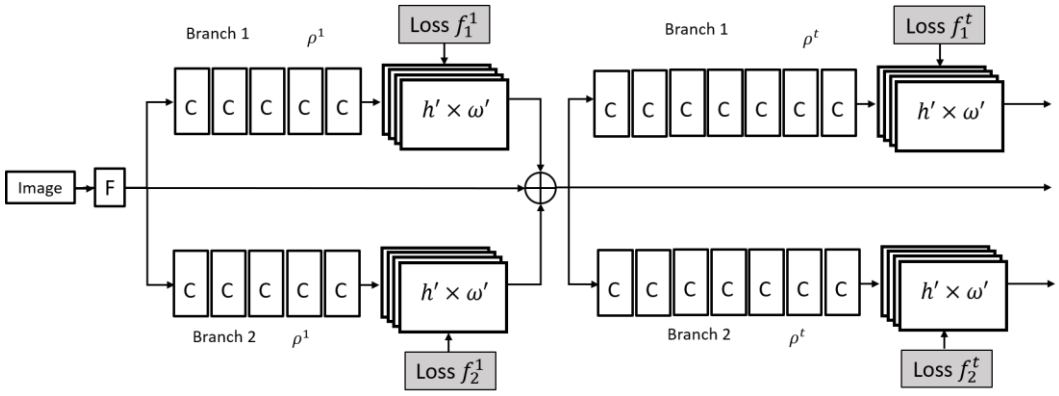


Figure 4.4. The structure of the OpenPose model.

The image was initially processed by a 10-layer convolutional network, where a set of feature maps F was formed (Simonyan and Zisserman 2014). As the input of stage 1, feature maps F was separately passed through two branches of CNN: branch 1 generated a set of parts confidence maps S1, which predicted human body parts in the image. Branch 2 formed a set of part affinity fields L1, which inferred the association of body parts. Parts confidence maps S1 and part affinity fields L1 together with feature maps F were concatenated for the next stage. At each following stage N, the prediction maps, including parts confidence maps S and part affinity fields L, were transmitted through two branches and concatenated with feature maps F for refining prediction. To iteratively train the multistage CNN, two loss functions were added at the end of each stage for each branch. The loss functions were designed to eliminate the difference between the prediction maps and the ground truth, which was labelled manually and reflected correct body parts and part associations.

At every stage N, for each joint P this convolutional network generated belief maps for every pixel, indicating the confidence level that a joint point appears in any pixel (u, v) of one single image. At stage 1, the weights for each existing layer of convolutional networks were initialized by applying the weights of the convolutional pose machine and those layers at the rest of stage N (N > 1) were randomly initialized. The architecture was trained through back propagation by using the Human 3.6M dataset, which contains 3.6 million human poses and corresponding 3D pose information (Ionescu et al. 2013). For the conversion from pixel belief maps into body joint localization, the pixel with the most confidence level was chosen as the location of each joint.

4.5 3D pose estimation

Under the feature extraction section, 2D key joints were firstly extracted using the OpenPose model from every image frame and, subsequently, were lifted 2D key joints into 3D. Then the raw coordinates of the 3D joints were converted into relative coordinates by subtracting the x, y, and z value of central point under raw coordinates. Each joint point was separated into x, y, and z; three features and 42 features were used for machine learning classification. As shown in Figure 4.5, the 3D pose estimation was formed on the basis of 2D human joint prediction. The belief maps produced by the OpenPose system were used as one of the inputs at each stage. Each stage of the structure of the 3D pose estimation merged two elements: (1) the belief maps from the 2D joint predictor (OpenPose system) and (2) the projected belief maps generated by the 3D pose projection as the inputs. The process underwent six stages. The 3D projection was designed, first, to lift the 2D coordinates into 3D and project 2D point locations into 3D, and then the fusion layer combined the 2D joint belief maps and 3D projected belief maps and propagated them into a set of 2D point landmarks in order to iteratively reinforce the 2D joint prediction as well as the 3D projection. At Stage 6, the fused belief map became the final output, and it was eventually projected to 3D models. The architecture of 3D estimation uses back propagation and trains the model end to end (Zhang, Yan, and Li 2018).

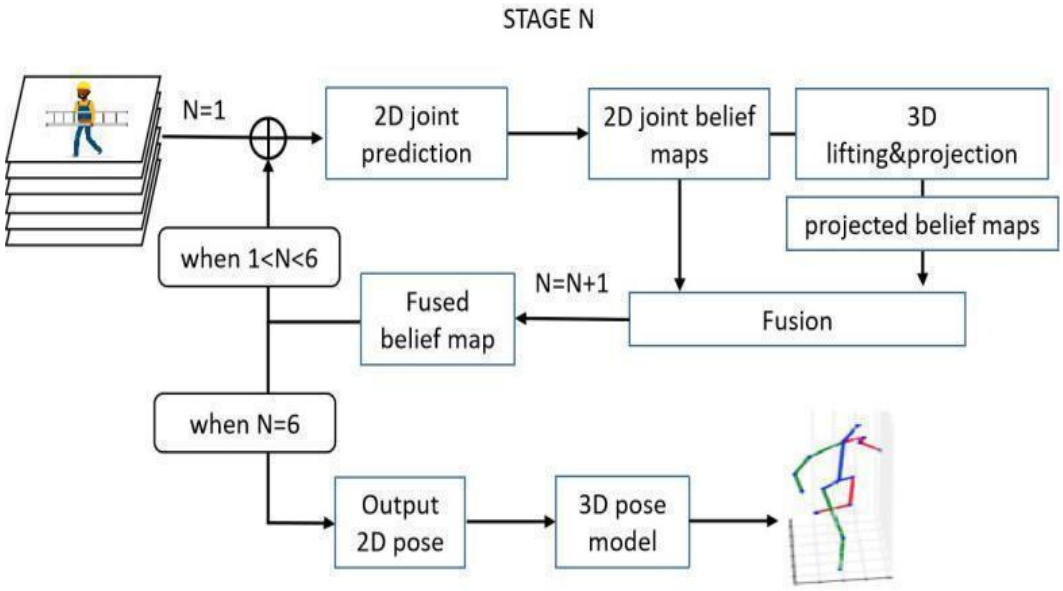


Figure 4.5. The structure of 3D pose estimation model

Inspired both by the approach (Pitelis, Russell, and Agapito 2013) that represents the space of human poses as a mixture of principal components analysis (PCA) and the concept that identifies poses as an interpolation between pose categories, the probabilistic 3D model consisted of a mixture of probabilistic PCA models with a number of clusters, and the PCA models were trained by the expectation maximization (EM) algorithm.

First, poses were subdivided into several pose categories P and the Euclidean distance d between pairs are computed. The objective was to look for a set of samples S that keeps the distance between joint points and their nearest sample minimized (see Equation (4.1)). This searching process was iterative and applied greedy selection that retains the previous S until another s that additionally minimizes the Euclidean distance was found. The process stopped when the chosen candidate became close enough to the existing candidate with little discrimination and the aligned points of certain pose category were assigned to its closest candidate s. The OpenPose system returns the belief maps $b_p[u, v]$ of landmark locations at every pixel and the most confident pixels Y_p are selected as the location of each landmark (see Equation (4.2)) so as to convert the belief maps into 2D locations for 3D estimation in the next step.

$$\underset{s}{argmin} \sum_{\mathrm{p} \in P} \underset{s \in S}{mind}(s, \mathrm{P}) \tag{4.1}$$

$$Y_p = argmax\, b_p\,[u, v] \tag{4.2}$$

A unimodal Gaussian 3D pose model was implemented to estimate the 3D human pose of a single frame. This model in practice took 20 samples and optimized the 3D pose reconstruction for each single rotation to achieve rotational invariance on the ground. A non-linear least squares solver was applied to find the basis coefficients of the best-found solution, and this approach provided results close to the global optima with the same average accuracy but less computational cost. The output of the belief maps from OpenPose model is taken as the input to an extra new layer that utilises pretrained 3D human pose model for 3D lifting from 2D poses.

Next, the generated 3D joint key points were projected onto a new set of 2D surfaces to form new 2D belief maps, and at the final layer in each stage these new 2D belief

maps together with the belief maps from 2D convolutional architecture were fused into a single map f_t^p according to the Equation 4.3 that the newly generated belief map $\hat{b}_t^p$ were fused with the previous belief map b_t^p. $w_t \in [0,1]$ denotes to the weight obtained from the training. The fused map was then used in the next stage as one of the inputs to refine the 2D joint prediction. For the final estimation of the pose, the 2D belief maps produced at Stage 6 were lifted into 3D space by the probabilistic 3D model (Tome, Russell, and Agapito 2017).

$$f_t^p = w_t \bullet b_t^p + (1 - w_t) \bullet \hat{b}_t^p \tag{4.3}$$

4.6 Classifier training

For the annotation, a series of short video sequences were collected in different duration for each activity category and each video sequence contained one single scaffolder conducting one type of activity. Additionally, based on our definition of scaffolding activities,these sets of video sequences were labelled with their activity categories: scaffold erecting (direct work), transporting (essential contributory work), and idling/waiting (ineffective work). The dataset details are presented in Table 4.1

Table 4.1 Dataset details

Activity Category	No. of Samples	Sum
Class 1: Working	811	
Class 2: Transporting	779	1731
Class 3: Idling	141	

As shown in Figure 4.6, 18 3D joint positions were directly obtained from the 3D pose estimation model. These joint points are under absolute 3D coordinates. Since the facial key points are not relevant and not essential for scaffolding activity recognition, for example, the locations of eyes and ears do not provide vital information for activity recognition,the points of eyes and ears were simplified by only keeping the nose

location. This simplification can effectively increase the accuracy and save the computational cost for classification. Since one single point in 3D space provides x, y and z coordinate information, 14 3D joint points left from the elimination of the four points of ears and eyes, comprised 42 features. To facilitate the computation and make all the data comparable, the absolute coordinates were converted into relative coordinates by setting the central point as a zero point (0, 0, 0) and cutting off the same amount of x, y and z values, which are the initial values of the central point, from each of the key points, respectively. The pose codebook was adopted, where each pose of a kind of activity is viewed as a single histogram containing 42 features or vectors.

class	x0	y0	z0	x1	y1	z1	x2	y2	z2	x3	y3	z3	x4	y4	z4	x5	y5	z5
1	0	0	0	-123.009	4.039891	1.050438	-264.616	-442.641	-106.096	-209.266	-407.302	-481.461	109.4773	-4.03989	3.176284	282.6542	-461.301	-61.4646
1	0	0	0	-123.733	4.81664	0.672439	-266.419	-443.293	-106.836	-211.005	-405.939	-484.008	109.2881	-4.81664	2.789171	282.7132	-464.084	-62.1492
1	0	0	0	-119.748	-3.34965	2.62569	-251.269	-452.764	-102.388	-151.765	-410.841	-474.382	110.3013	3.349647	4.710845	287.3098	-447.941	-59.529
1	0	0	0	-150.361	-4.97629	2.787759	-242.773	-463.152	-97.827	-140.984	-429.72	-429.093	119.6792	4.976277	4.906296	300.6075	-459.002	-58.8866
1	0	0	0	-115.542	-3.82978	0.808794	-247.419	-458.373	-101.398	-143.142	-419.054	-433.332	117.5921	3.829774	5.733648	298.9911	-456.439	-59.2546
1	0	0	0	-110.395	-3.24758	0.922782	-236.169	-460.571	-52.8473	-131.878	-440.432	-418.264	122.2304	3.247569	5.733137	297.6316	-459.27	-53.589
1	0	0	0	-122.481	-13.0545	2.933637	-199.775	-478.53	-54.0747	-144.995	-465.02	-457.514	141.3068	13.05447	5.278184	327.3196	-446.354	-51.64
1	0	0	0	-115.294	-12.6677	-16.0707	-187.87	-467.486	-36.9349	-128.4	-444.292	-413.125	136.4526	12.66774	19.13157	308.4498	-440.943	-27.51
1	0	0	0	-132.526	14.11619	-18.0075	-281.719	-410.336	-86.3959	-167.965	-344.58	-444.835	131.898	-14.1162	27.05579	255.4249	-458.485	-122.2
1	0	0	0	-124.072	13.15002	-46.5411	-276.958	-388.647	-80.5397	-159.035	-314.068	-443.809	134.2473	-13.15	33.28558	262.1408	-440.387	-72.6374
1	0	0	0	-142.51	36.7136	-26.6462	-320.443	-310.202	-152.938	-194.423	-177.155	-482.085	118.8693	-36.7135	18.36021	235.4805	-437.537	-149.421
1	0	0	0	-135.413	29.50374	-25.523	-305.222	-336.71	-149.366	-180.683	-204.86	-475.568	128.0163	-29.5037	21.55687	248.4613	-441.076	-143.541
1	0	0	0	-93.5169	-87.6278	-4.95221	265.5818	-424.222	-120.833	291.8072	-373.374	-501.936	143.3239	87.62772	38.76596	354.9167	-299.239	-112.498
1	0	0	0	-99.9082	-59.7622	-73.4852	81.8245	-482.731	-74.9108	229.4437	-469.693	-426.138	155.5564	59.76211	7.782999	344.3944	-383.529	-144.148
1	0	0	0	-127.798	12.90975	-28.0482	-282.981	-404.249	-104.344	-165.502	-322.441	-471.641	135.8575	-12.9097	17.78221	265.5147	-453.196	-93.9869
1	0	0	0	-118.283	38.08803	-41.8376	-295.255	-316.994	-135.836	-213.854	-193.863	-518.637	108.1255	-38.0879	35.16317	221.8704	-441.267	-172.21
1	0	0	0	-134.965	20.20821	-25.4019	-254.045	-374.459	-149.82	-177.449	-252.165	-488.234	131.6358	-20.2081	17.57657	260.0097	-448.328	-146.298

Figure 4.6. Data frame of 3D joint positions (This figure partially displays the data frame of 3D joint positions. The actual data frame contains 43 features and more than 1700 lines).

To recognize scaffolding activity from short video sequences, a discriminative classifier was trained with the annotated data. The classifier takes the body skeleton features as the input, which provides sufficient descriptive content for activity classification. At the training stage, first, key points of the skeleton were extracted from every video sequence for each activity category and each video sequence was processed into frames, which became the input of our pose estimation model. For every image frame, a set of 42 features was generated by the pose estimator plus one manually labelled feature indicating the activity category constituted the input for supervised classifiers.

Since different supervised classifiers have varying performance on one particular classification task, major popular multi-class algorithms were adopted, including random forests (RF), decision tree (DT), artificial neural networks (ANN), one-vs-one support vector machine (one-vs-one SVM), one-vs-all support vector machine and k-

nearest neighbors (KNN). These algorithms are capable of multi-class classification as well as handling data with multiple features. To select a classifier with outstanding performance, each of the aforementioned algorithms was employed for classification through our dataset of scaffolding activities, and the training and testing process followed the principle of cross validation, which is a statistical procedure for the evaluation and validation of machine learning models.

Decision Tree (DT)

Just like its name, DT is an algorithm with a tree structure where a tree leave denotes an outcome label, and a branch represents a sub section of an entire tree or a sub-tree. DT can be used for classification and regression. For classification, a tree is constructed through the process of binary recursive partitioning, which is a procedure that iteratively splits the data into partitions. The divide and conquer algorithm was used in this process by breaking down a sophisticated group into two or more subsets with purely one type of feature, until the subsets become simple enough to be classified directly. DT is the building blocks of the random forest models (Shaikhina et al. 2019).

Random forest (RF)

RF is an ensemble learning approach for classification and regression. At the training stage, multiple individual decision trees are created by the selection of various subsets of training samples, which follows the technique of bootstrap aggregating or bagging. The selection principle allows that the same data sample can be randomly selected several times, while other sample may not be chosen at all (Chen et al. 2021). At the prediction stage, every single tree structure independently produces a class prediction, and the ultimate prediction of the RF model is that which has the majority of votes from the decision trees. The use of one single decision tree is susceptible to bias and variance, for instance, if one decision tree is too shallow, its prediction is easily influenced by bias and if one tree is too deep, it is probably overfitting with a high variance. Since RF models aggregate a multitude of decision trees, however, the models effectively reduce variance and mitigate volatility and noise because of the multiple data samples, which enhances the robustness of classification (Belgiu and Drăguţ 2016).

K-nearest neighbors (KNN)

KNN is designed to search for the K closest data points to the target to be classified. Generally, KNN contains three steps: (1) computing the distance between the target point and each point in the training dataset; (2) according to the K value that the researcher set previously, picking K data points that remain the lowest distance with the target; and (3) applying a majority vote so that the prediction result is in accordance with the majority of classes among K neighbour points.

Support vector machine (SVM)

SVM is a machine learning algorithm generally used for data classification. SVM generates a hyperplane or a group of hyperplanes in a high-dimensional space to distinctly classify the data points. There are many hyperplanes that are available to select, but SVM aims to look for a plane that maintains the maximum margin, which means this plane keeps the maximum distance between data points of both classes. SVM was initially a binary classification approach; however, after the model's extension, SVM could be used for multi-class classification. One-vs-all (one-vs-rest) and one-vs-one approaches are two common ways to solve multi-class tasks. In the case of an N-class problem, a one-vs-all approach generates N binary SVM classifiers, while each one splits one class from the rest(Li et al. 2017, Qu et al. 2020, Song et al. 2018). The ith ($i \in N$) classifier is trained with the whole training data points of ith class labelled positive, and with all the other classes labelled negative. The one-vs-one approach is developed based on the idea that a pair of distinct classes are selected each time and are trained by a binary SVM classifier, and an ensemble of binary SVM classifiers forms a one-vs-one SVM classifier. In the case of a problem of N different classes, $\frac{N(N-1)}{2}$ binary SVM classifiers are required for separating each two distinct classes. Compared to one-vs-all SVM, one-vs-one SVM is more computationally expensive since more binary SVM classifiers are created. Both one-vs-all SVM and one-vs-one SVM will be implemented for model evaluation (Hsu and Lin 2002). Figure 4.7 demonstrates the SVM classifier that divides scaffolding activities into erecting, transporting, and waiting/idling three classes.

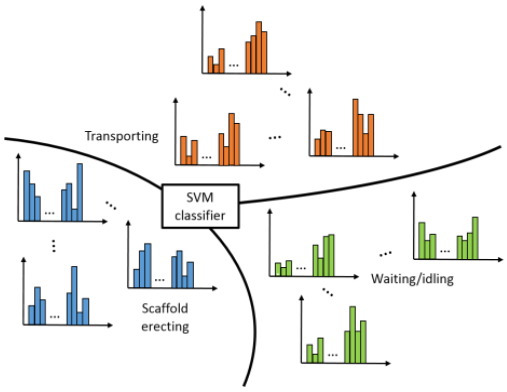


Figure 4.7 SVM classifier (it looks for a group of hyperplanes to classify data points or vectors)

Artificial neural networks (ANN)

ANN were inspired by the mechanism and structure of biological neural networks in the human brain, where the signals are received, transmitted, and emitted back and forth through multiple layers of neurons. The basic unit in an ANN is called a neuron or node. Each neuron is responsible for receiving input data, merging the input with its own parameters or processing the input with an activate function. The activate function plays the role of a value conversion that avoids great output variation as the input changes. Connections link every neuron with weights. A weight reflects the contribution percentage of an input to the output of the next layer of neurons. Typically, ANN initially sets up with randomized weights of all the neurons. Generally, neurons are organized into numerous layers, and every neuron in the layer only connects with the neurons in the closest neighbouring layers. ANN consists of an input layer where an input is received, a hidden layer, and an output layer where the result is produced. The hidden layer is sandwiched between the input and output layers and can be multiple according to the complexity of ANN architecture. At the training stage, ANN adopts the algorithms of forward propagation and backward propagation to train the model. Forward propagation refers to the procedure where the input data is fed in a forward direction from the input through the hidden layers to the output layer in the neural networks and the calculations and the storage of inter parameters are conducted. Backward propagation starts from the comparison of the deviation between the expected (labelled) output and the generated output, and it runs iteratively in order to minimize this deviation to a predetermined level. As all the weights of the neurons are randomly initialized, backward propagation regulates the contribution of every node

to the final output by nudging the weight connecting every neuron of layers from the output layer to the input layer. As a result, the ANN can automatically predict the output.

For each supervised learning classifier, the parameters of models and functions were optimized by iteratively comparing internal model performance on the same training dataset. The optimal classifier was chosen by picking the supervised classifier with the highest mean accuracy through 10-folds cross validation, which represents a globally recognized evaluation method of classification performance, and the presentation of confusion matrix with reasonable results. Next, the optimal classifier was designed to conduct a prediction of workface assessment from several video sequences comprising various scaffolding activities in actual working condition. The predicted workface assessment was compared with the manual annotation on the same video sequence for validation and discussion.

4.7 Case Study and Results Analysis

Due to the lack of relevant databases for training and testing the visual activities of scaffolding construction operations, before validating our approach it was crucial to collect a corresponding video sequence dataset. As shown in Figure 4.7 and Figure 4.8, the target scaffolding operation comprised three activities: scaffold erecting, transporting, and idling/waiting. Cameras with 12 megapixels were used to capture video data under a real construction environment in Western Australia as well as in the laboratory of the Australasian Joint Research Centre for Building Information Modelling in Curtin University. Professional volunteers were recruited to conduct the same activities both under the actual and laboratory construction conditions. Workface assessment requires supervisor to clearly observe and record entire scaffolding activities from a certain distance. This observing is conducted from distances ranging from 5 meters to 20 meters, which allows supervisor to identify distinct activities. Thus, in our approach, portable video cameras were placed 5 meters to 20 meters away from the scaffolding activities. To increase the variety of the database, the camera shot angle was set at three height levels: the first one was located between ground level and kneel level, the second was located between kneel level and eye level, and the third was located above head level. In addition, horizontal camera angles ranged from the frontal angle, three-quarter front angle, to profile angle (a view from the side). Fifty-two clips

of video regarding relevant scaffolding activities were collected; some of these videos were trimmed if they lasted more than 8 seconds and only included one single activity for training and testing purposes, and the rest of the videos covering several scaffolding activities were stored for validation. Forty clips were collected in the lab and the remaining 12 clips were collected on-site. The duration of the clips were from 3 to 10 minutes. The captured video has 25 fps and 1080p resolution. Every video clip was processed into a stack of image frames, and the whole dataset consisted of 1,731 frames. Additionally, these frames were annotated with their corresponding activity category, including erecting, transporting, and idling/waiting as the input for the purpose of supervised learning. Every video clip that only included one category of scaffolding activity was processed into image frames. For each image frame, its corresponding activity category was added as one additional feature among 42 features of body key points.

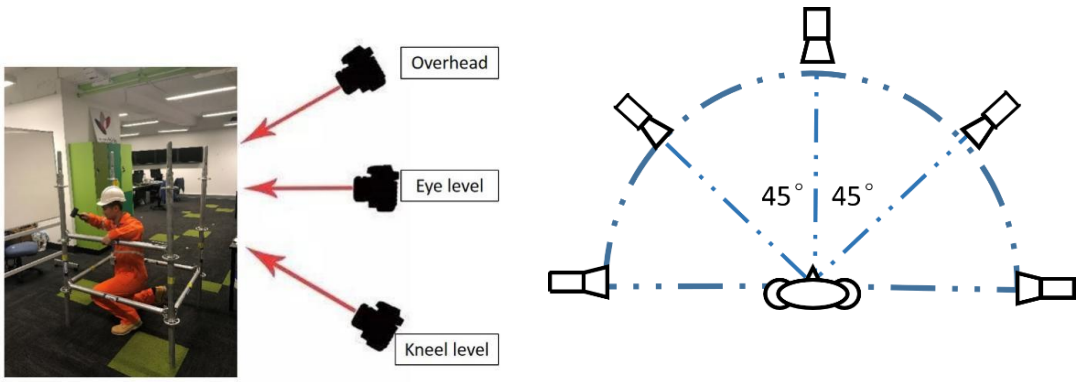


Figure 4.8. Experiment settings: the camera shot angle was set at three height levels: kneel level, eye level and overhead level, and horizontal camera angles ranged from the frontal angle, three-quarter front angle, to profile angle.

(a)

(b)

(c)

Figure 4.9. Video sequences of scaffolding activity: (a) scaffolding erecting; (b) transporting; (c) waiting or idling.

To evaluate the performance of our proposed approach for construction activity analysis, a case study was carried out. The methodology for activity classification was achieved in a Python-based environment and open-source package Scikit-learn with the assistance of a DELL workstation with a 1.9 GHz processing unit and a 32 GB RAM.

The accuracy of each frame is calculated via comparing the classification outcome generated by each machine learning classifier with the ground truth of every video frame which are manually annotated. If the outcome from machine learning classifier is the same as the ground truth, it would be counted as accurate. If the outcome is different from the ground truth, it would be taken as inaccurate. As it is shown in Table

4.2, the mean accuracy of selected supervised classifiers under 10-folds cross validation regarding to our scaffolding dataset were 96.58% (RF), 94.24% (SVM), 92.08% (DT), 96.13% (KNN), 93.12% (NN), respectively. All the classifiers' general accuracy achieved over 90%, while RF and KNN outperformed the rest with an accuracy around 96%.

Table 4.2. The performance of the classifiers3D pose classification is a transitional

Classifier	Accuracy	Macroaverage Recall	Macroaverage Precision	Average_F1 score
RF	96.58%	0.9445	0.9405	0.9422
SVM	94.24%	0.9209	0.9349	0.9273
DT	92.08%	0.9040	0.8749	0.8880
KNN	96.13%	0.9492	0.9578	0.9537
NN	93.12%	0.9339	0.9337	0.9337

step in our model and every video frame of scaffolding work can be classified as one of three scaffolding activities. The results of the 3D pose estimations included the classification results for every video frame as well as the visualization of 3D key joints and pose estimations. The classification results of every video frame were evaluated and validated via 10-fold cross validation, and the performance metrics including accuracy, precision, recall, and F1 score (Equation 4.4-4.6) indicated that all the classifiers achieved good performances. The confusion matrix of the classifiers is presented in Figure 4.10.

$$\text{Pre}cision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4.4}$$

$$Re\,c\,all = \frac{TruePositive}{TruePositive + FalseNegative} \tag{4.5}$$

$$F_1 score = \frac{2}{recall^{-1} + precision^{-1}} = 2 \bullet \frac{precision \bullet recall}{precision + recall} \tag{4.6}$$

Figure 4.10. Confusion matrix of the classifiers. The shaded boxes indicate the majority of detection results.

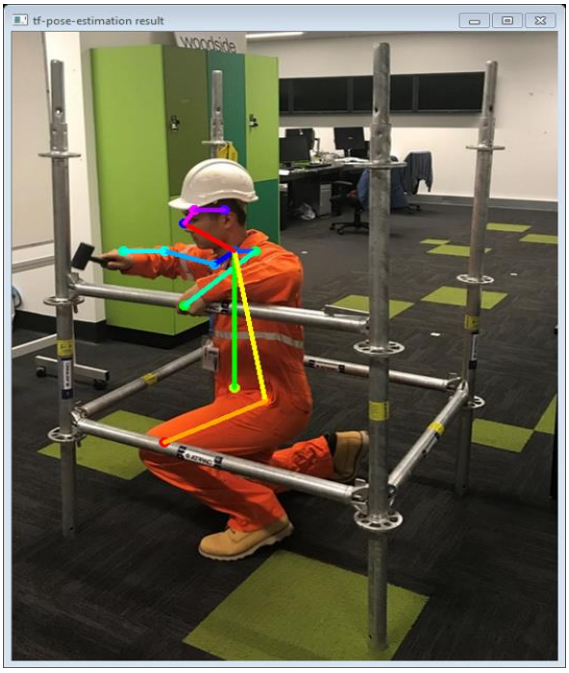
To test the capability of the formation of workface assessment, three video clips containing all activity categories were used for validation. As shown in Table 2, Video Clip 1 lasts for 206 seconds, and Clip 2 and Clip 3 last 116 seconds and 79 seconds, respectively. In addition, Clip 1 was converted into a sequence of image frames by extracting frames with 2 seconds intervals, while Clip2 and Clip3 were converted with 1 second intervals. 1,731 frames were used for training and testing and these frames went through 10-folds cross validation. 314 frames were used for the validation of the case study. For the shot angle, Clip1 was collected at the kneel level, while Clip2 and Clip3 were collected on chest level and overhead, respectively, to test the ability of generalization of our model.

The classification task is to classify scaffolding activities into working, transporting, and idling three categories and these three activities are equally important. According to the performance standard in computer vision, mean accuracy is the most suitable performance indicator for the classification that each category is equally important. Since the RF model has the highest mean accuracy and relatively high score in recall, precision, and F1 score through our training procedure, the RF model was chosen for the validation of the workface assessment. The sequences of image frames from three clips were fed into the OpenPose and 3D extraction model, and then 3D pose features were generated and transmitted through our RF classifier for activity classification. Figure 4.11 shows an example of 3D key joints and skeleton extraction. Each frame was annotated with one of our three activity categories by the RF classifier, and these annotations were placed in the order of temporal dimension to form the workface assessment. During the formation, the noise of activity status was eliminated by taking the majority status among each image frames. The duration of each activity category was determined by the following logic: 5 seconds were set as the unit of time period and picked the majority category (among the categories before and after 2 seconds) as the activity category the time point belonged to. Thus, the boundary of each activity category occurs when two different activity categories are identified before and after one certain point in time. For example, one central frame was annotated by the model as the waiting/idling category, while the rest frames next to it were annotated as

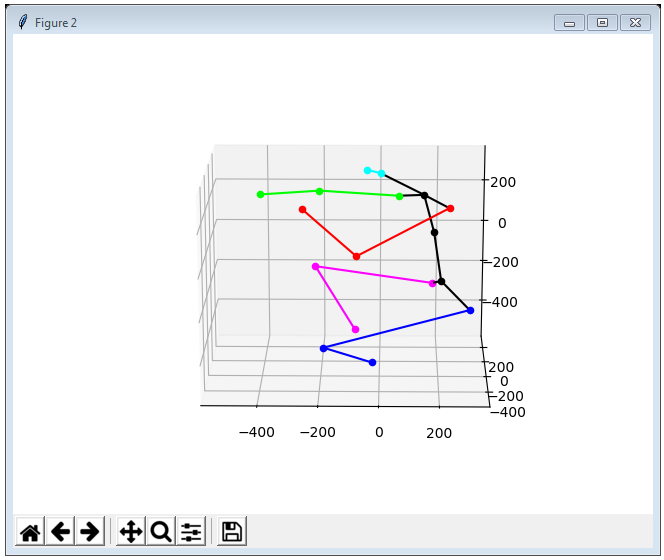
scaffolding erecting status and the erecting was regarded as the activity category of this period. The result of workface assessment is displayed in Figure 4.12.

Table 4.3. Details of video clips for validation

#Clip	Length(s)	Number of frames	Frame interval	Shot angle
1	03:26	102	2 seconds	Kneel level
2	01:56	118	1 seconds	Chest level
3	01:19	79	1 seconds	Overhead



(a)



(b)

Figure 4.11. 3D key joints and skeleton extraction: (a) key joint distribution in the image frame of scaffolding erecting and (b) the corresponding skeleton model in 3D.

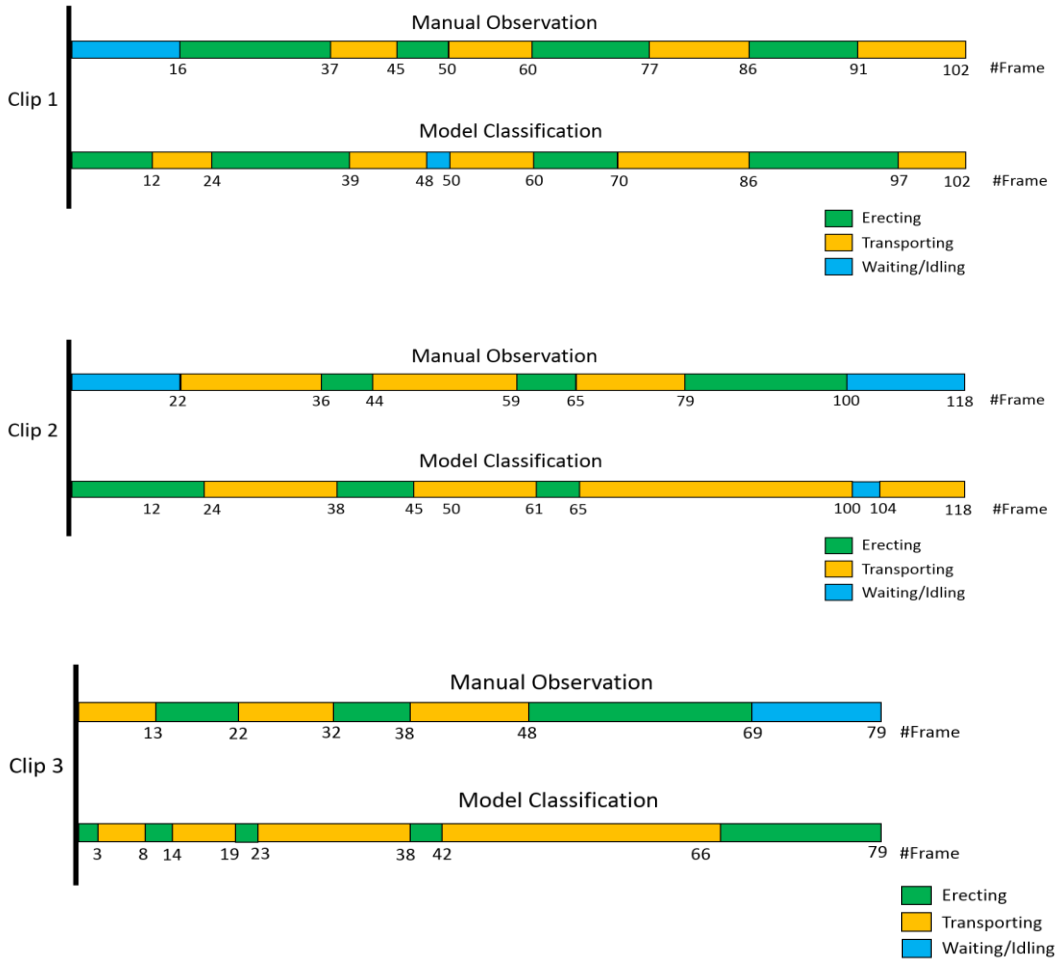


Figure 4.12. Comparison of workface assessment between manually labelled ground truth and automatic generated results from the model

According to Abhinav's research on the productivity measurement, the productivity of scaffolding can be calculated as below (Peddi, 2008):

$$P = (T_{effective} + T_{contributive})/(T_{effective} + T_{contributive} + T_{ineffective})$$

Where $T_{effective}$: Time of effective activities in scaffolding

$T_{constributive}$: Time of contributive activities in scaffolding

$T_{ineffective}$: Time of ineffective activities in scaffolding

Thus, the productivity of scaffolding was calculated based on the result of workface assessment and was listed in Table 4.4.

Table 4.4. Productivity result from workface assessment

#**Clip**	P(manual)	P(model)
1	0.84	0.98
2	0.66	0.97
3	0.87	1.00

Based on the comparison between the model outputs and the manual observation in Figure 4.12 and Table 4.4, this case study displayed several phenomena. Firstly, our model presented the feasibility and capability of discrimination between the scaffolding activities of erecting and transporting and generated relatively effective intervals between these two activities. Although a little divergence exists between the boundaries of distinct activities, which may be caused by the transitional procedure during the conduct of scaffolding activities or the occlusion of body parts, the model classification is basically capable of reflecting the activity status.

Secondly, the discrimination between the activities erecting and waiting/idling is vulnerable when the participant performed either erecting or waiting/idling activities. In three testing clips, the waiting/idling activity was misclassified as scaffolding erecting. This poor discrimination may be a result of the complexity of activity scaffolding erecting or the similarity between these two activities. In addition, due to the fact that the waiting/idling ineffective activity was misclassified as the erecting effective activity by the model, the productivity predicted by the model is slightly larger than the productivity of manual observation.

4.8 Conclusion and discussion

In conclusion, in relation to the background of the rising interest in automation and digitalization in the construction industry, a method of automatic scaffolding workface assessment has been proposed and developed in this chapter to replace repetitive man-hours of data collection and supervision. Scaffolding activities can be recorded and visually monitored through onsite video cameras and then classified into three activity categories based on the procedure of workface assessment. One video dataset of

scaffolding activities was created for training and testing. The case study demonstrated the feasibility of our developed model in workface assessment and productivity measurement, and the developed model presented enormous potential to enhance project monitoring and controlling.

At the current stage, the proposed model merely takes every video frame as the input without taking the temporal feature into consideration. As a result, our model classifies every video frame of scaffolding into individual activity categories without considering scaffolding postures in one kind of activity as an entity. scaffolding activity were simplified into three categories; however, scaffolding has greater complexity in activity transition as well as activity subdivision.

The core contribution of this study includes that 1) a video dataset was established for the research of scaffolding activities; 2) an approach was developed for automatic workface assessment and productivity measurement of scaffolding by using the method of 3D key joint extraction and machine learning classification; 3) the principle of activity analysis was followed and scaffolding activities were categorised into erecting, transporting, and idling and then explored the potential of activity recognition for the field of construction inspection and productivity measurement. This approach is designed to release on-site managers from regular inspections with the help of video surveillance.

Future work would include increasing the volume and variety of training data to increase the robustness of model prediction. In addition, improving the structure of our prediction model is another area that requires great efforts in future. Also, due to the complex nature of the scaffolding activity itself, the scaffolding activity need to be analysed and divided in more depth.

5 Scaffolding component quantity measurement

5.1 Chapter introduction

In this chapter, the approach of scaffolding quantity measurement via static images is illustrated. Unlike the approach of activity analysis, which records scaffolding activities in a dynamic mode, this approach aims to extract scaffolding components from static site images without capturing scaffolders' motion. These site images were taken periodically during the scaffolding erection. By recognizing and comparing the quantity discrepancy before and after a certain construction period, the progress and productivity of scaffolding projects is estimated.

In practice, scaffolding productivity is commonly measured by the unit *m^3/day* or *time/m^3*. To realize automatic scaffolding quantity measurement, extracting valid scaffolding structure is crucial. Chi et al. explored the automated recognition of scaffolding tubes (Chi et al. 2017), and Xu et al. investigated the reconstruction of scaffolding elements, including tubes, toeboards, and decks (Xu et al. 2018). However, these previous studies failed to convert recognition outcomes into scaffolding progress and productivity estimation. In this study, the intersections of scaffolding structure were investigated (shown in Figure 5.1) and these intersections are made up of couplers or wedges. From a perspective of image processing, scaffolding structure consists of regular cubic space, and scaffolding tubes are tied up by couplers or wedges, which presents in a form of straight lines intersecting each other. The valid recognition of the structural intersections can quantify the cubic space of a scaffolding structure.

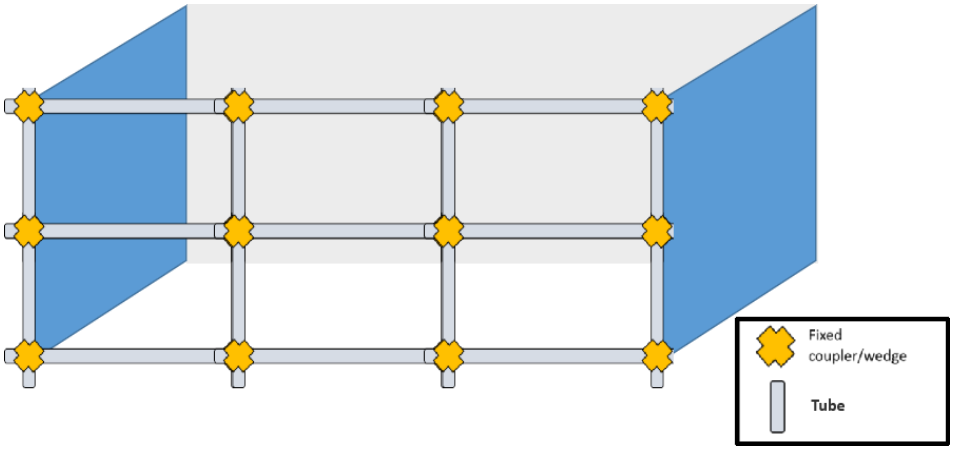


Figure 5.1The intersection of scaffolding

Figure 5.2 presents the research workflow for this approach. The model takes colour images of scaffolding structure as input and generates the 2D locations and the number of scaffolding intersections (couplers or wedges) in the images, as output. First, the vision data is captured by onsite 2D RGB camera, and every image only includes one main structure of the scaffolding project. Next, the images go through the procedure of image processing and feature extraction. Then, object detectors are designed to recognize target objects from the processed images. The model can automatically count the number of detected objects to quantify scaffolding components. Together with graph analysis and project design parameters, the model estimates the scaffolding productivity of the project in the image. Section 5.2 illustrates the procedure of image processing and feature extraction. Section 5.3 introduces the application of two object detectors in our approach. Section 5.4 displays the method of the component quantification, graph analysis, and scaffolding productivity estimation. Section 5.5 demonstrates the experiments and analysis of our model. The results are shown and discussed in Section 5.6. Finally, a conclusion is drawn in Section 5.7.

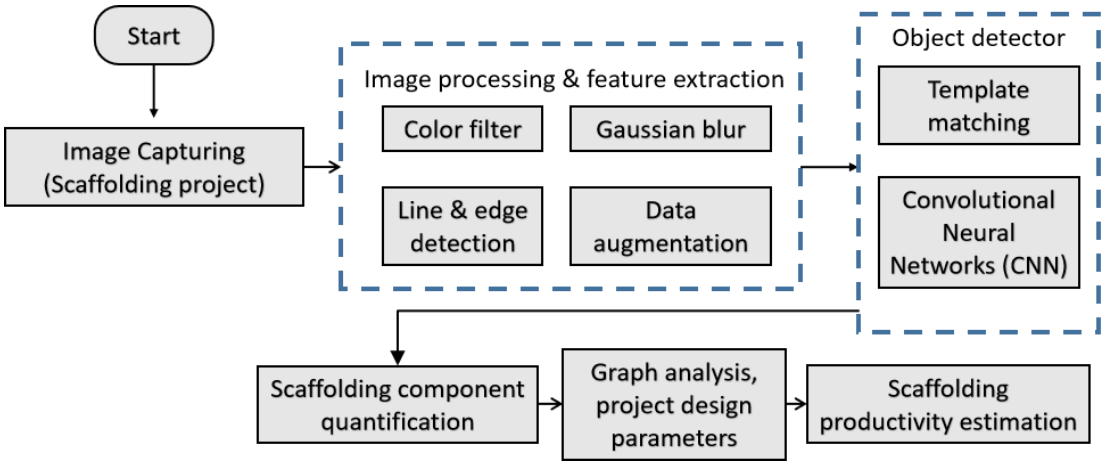


Figure 5.2. The research workflow of scaffolding component quantity measurement

5.2 Image processing and feature extraction

Noisy construction background and various sizes of scaffolding structures make automatic visual detection challenging. To enhance the efficiency of object detection, image processing, as one essential step, aims to minimize the noise and highlight the potential target if possible. Furthermore, scaffolding has unique structures and features.

From the perspective of computer vision, scaffold consists of crossed straight lines, and it also presents distinct colours apart from the background in many circumstances. Extracting useful features from irrelevant background can assist the step of object detection. In the following section, the computer vision algorithms and techniques were introduced for image processing and feature extraction.

5.2.1 Gaussian blur

Blurring is a commonly used technique of image processing where its effect works like a frosted glass covering on the image; it is also called image smoothing. Blurring can be utilized for removing image noise and lowering levels of detail. Images captured on construction sites contain a large number of sophisticated noise, and blurring can effectively reduce noise and enable core content in the image to stand out for recognition.

Gaussian blur is a type of blurring approach that transforms every pixel in an image with Gaussian normal distribution, and it is a process of weighted averaging for a whole image where the new value of every pixel is obtained through weighted averaging of that pixel and its neighbors (Flusser et al. 2015). In two dimensions, the formula of a Gaussian function can be expressed in Equation 5.1.

$$F(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5.1}$$

where x is the horizontal distance from the origin, y is the vertical distance from the origin, and σ is the standard deviation of the Gaussian normal distribution. Intuitively, in two dimensions, as shown in Figure 5.3, this formula constructs a 3D surface of weights whose contours are concentric circles with a Gaussian normal distribution from point (x, y). In practice, a convolution process is conducted, where a convolution matrix is formed based on the Gaussian distribution, and each pixel's new value is produced by weighted averaging those original pixels' values with the convolution matrix. Point (x, y) that gains the largest Gaussian value receives the heaviest weight, and its nearby pixels have smaller weight with the distances to the point (x, y) increase.

Gaussian blur is usually used in combination with edge detection and line detection because edge and line detection are sensitive to image noise and image noise presents

sharp gradient variation. The selection of a large size kernel filter can cause the loss of valuable details, while a tiny size filter cannot reach the ideal blurring effect. Thus, a 5×5 kernel filter size was chosen for our model. Edge and line detection is introduced in the following subsection.

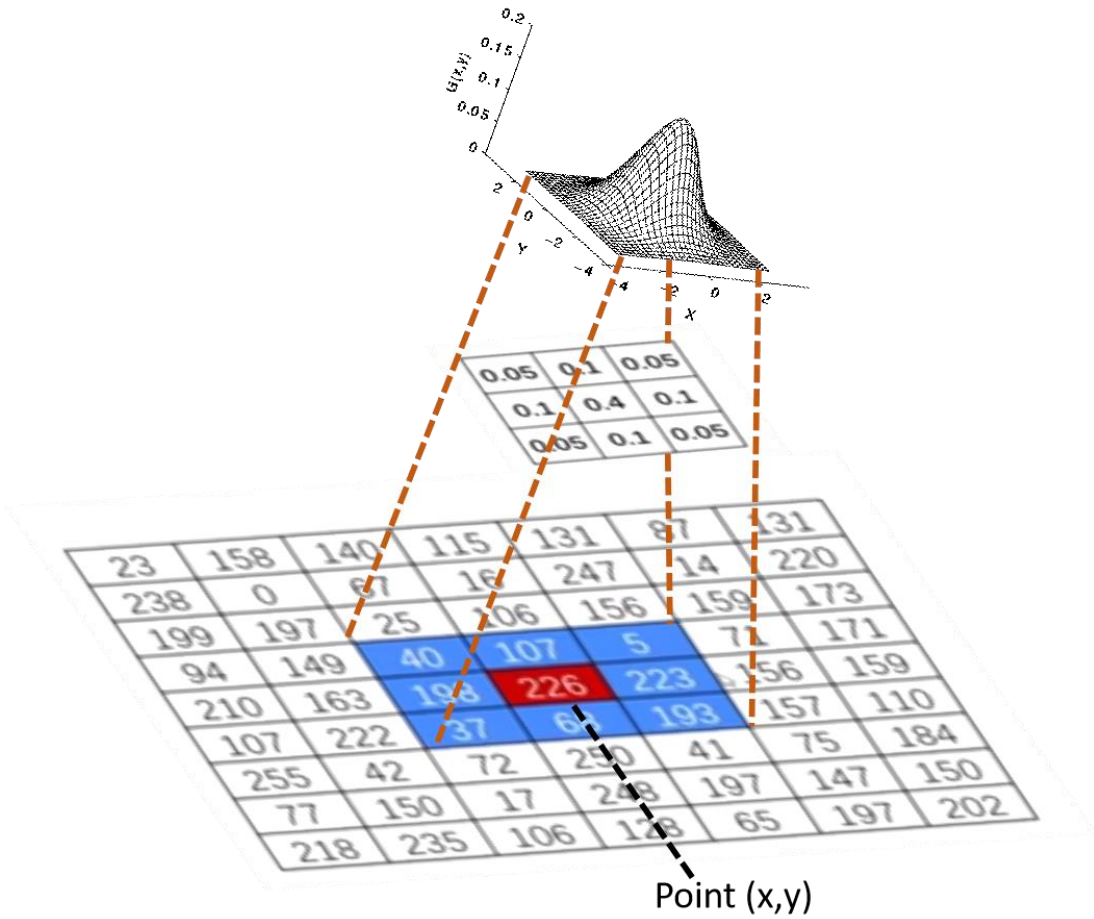


Figure 5.3. The mathematical model of Gaussian blur

5.2.2 Edge and line detection

Edge detection is utilized for recognizing and detecting an objects' edge in an image. Edge denotes the region where intensity difference changes sharply in an image, and a successful edge detector generates a series of solid lines representing the boundaries of objects. In this research, edge detection was found to be a useful technique to extract scaffolding structural information from captured images.

The Canny algorithm was first developed by John F. Canny in 1986. Although it has been a long time since this algorithm first appeared, the Canny edge detector is still

regarded as the state of art for edge detection. The process of the Canny algorithm mostly includes four steps: 1) apply Gaussian blur to remove image noise, 2) calculate the intensity gradient of the image, 3) apply non-maximum suppression, and 4) set hysteresis thresholds and connect edges together (Zhao, Qin, and Wang 2010).

For the calculation of intensity gradient, a Sobel operator was applied, which consists of two kernels in order to acquire first-order derivatives in both the horizontal and vertical directions.

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \tag{5.2}$$

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \tag{5.3}$$

G_x and G_y are the two kernels that are practically applied to convolve the image in x and y directions, respectively.

Then, the gradient strength and direction can be obtained with:

$$G = \sqrt{G_x^2 + G_y^2} \tag{5.4}$$

$$\theta = arctan\left(\frac{G_y}{G_x}\right) \tag{5.5}$$

In order to simplify calculations, the direction θ is rounded to one of four angles including 0, 45, 90 and 135.

Non-maximum suppression, just as its name implies, is utilized to filter out pixels that are less than the local maximum values. After obtaining the gradient strength and direction, the full range of pixels in the image were checked to establish which pixel becomes the one with local maximum gradient strength in its gradient direction. Then, those pixels with local maximum values were selected as edge candidates.

In the step of hysteresis thresholding, double thresholds were applied, where a high threshold was established to allow a group of pixels with strong gradient values to be classified as edge points, and a low threshold was applied to filter out a series of pixels with weak gradient values, such as image noise and colour variation. Additionally, those pixels between the low and high thresholds were retained only if they continued or reinforced the connectivity of previous determined edge points (Medina-Carnicer et al. 2009). This step can effectively remove noise and build up the connectivity of edge points.

An image of scaffolding structures commonly consists of straight lines, as scaffolding tubes present the feature of assembled straight lines in the image. An as-built or completed scaffolding structure normally presents a series of intersectional straight lines. Therefore, accurate line detection enables us to effectively recognize the scale and quantity of scaffolding tubes. Line detection is a practical function that reveals straight line in an image.

Hough transform is a technique of feature extraction in computer vision, and it has been widely used for detecting simple geometrical shapes, such as straight lines and circles. It is a given that an infinite number of lines can pass through a single point and that this group of lines passing through the single point only represent one line in the parameter space. To fix the problem that vertical lines would output infinite values of the slope parameter when the parameter space is another Cartesian coordinate, polar coordinates were used to describe image points and lines in the Hough transform.

Assume point (x1, y1) passes through a line in Figure 5.1, and r is the distance from the origin to the point (x1, y1) and θ is the angle between the x axis and the line connecting the origin with the point (x1, y1).

$$r \cdot \cos\theta = x_1$$

$$r \cdot \sin\theta = y_1$$

$$x_1 \cdot \cos\theta = r \cdot \cos^2\theta$$

$$y_1 \cdot \sin\theta = r \cdot \sin^2\theta$$

$$x_1 \cdot \cos\theta + y_1 \cdot \sin\theta = r(\cos^2\theta + \sin^2\theta)$$

$$x_1 \cdot \cos\theta + y_1 \cdot \sin\theta = r \tag{5.6}$$

$$r = x\cos\theta + y\sin\theta \tag{5.7}$$

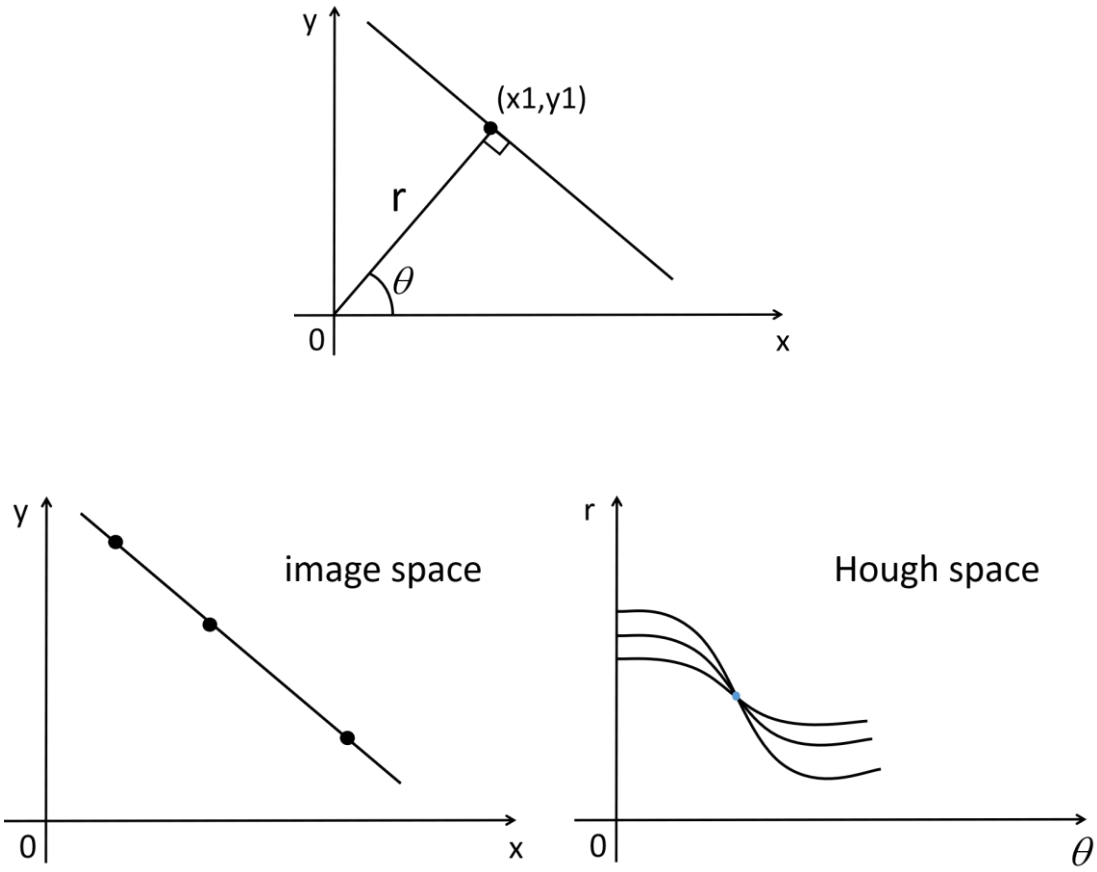


Figure 5.4. Hough transform

Thus, given a single point (x, y) in the Cartesian coordinates, all the lines that pass through that point (x, y) are represented by a unique sinusoidal curve in the polar coordinates. Additionally, these polar coordinates can be referred to as Hough space. For instance, in the figure XX, there are three points falling in a straight line in the image space, and in Hough space their mapping from the image space, three corresponding sinusoidal curves, intersect at one single point which represents that straight line in the image space. Therefore, every point in a two-dimensional (grey scale) image can be converted into sinusoidal curves in the Hough space, and by determining the intersection of two sinusoidal curves in the Hough space, consequently, the corresponding straight line can be determined in the image space, which is the mapping of the intersection in Hough space.

In the process of Hough line detection, three parameters need to be adjusted: accumulator threshold, minimum line length, and maximum line gap. The accumulator

threshold is an intensity parameter allowing those lines to be selected when they reach enough intensity. Minimum line length rejects those line segments that are shorter than this pre-set parameter. Maximum line gap connects two points on the same straight line when the gap between them is shorter than this parameter. Adjusting these thresholds can change the number of detected lines in the image (Aggarwal and Karl 2006).

Specific steps for Hough line detection can be concluded as follows: 1) convert coloured image to grey image, 2) apply Gaussian blur, 3) edge detection, 4) binary conversion, 5) Hough transform, 6) set thresholds and filter noise, and 7) draw straight lines.

5.2.3 Colour filter

Colour is one of crucial features that objects present in RGB images. Every object has its own colour, and objects belonging to the same category commonly possess the same colour, while objects from different categories often have distinct colours. According to this phenomenon, it is proposed to extract scaffold from other construction objects and background in RGB images.

An image is made up of pixels. A coloured RGB image is created by three channels of pixels (red, green, and blue) in identical size merged. Each pixel value ranges from 0 to 255, which stands for the brightness of each colour. For convenience, RGB channels was converted to HSV (hue, saturation, value) channels, which is another colour space broadly applied in computer vision. In the HSV model, these parameters of hue, saturation, and value can be conveniently adjusted, and the result of colour extraction is intuitively displayed.

On construction sites, scaffolds normally have the colours of silver or grey, yellow, dark red, depending on the scaffold type the site used, the aging degree, the degree of rust, and the illumination condition. A panel is designed and programmed to adjust the minimum and maximum values of HSV, so as to only extract useful colour information.

To enhance the performance of the colour filter, the input frame was first smoothed by Gaussian blur to remove irrelevant image noise. Next, the input frame was converted from RGB channels to HSV colour space. The subsequent step aims to extract

scaffolding regions by selecting colour thresholds. Since different illumination conditions change the colour that scaffolding components present, colour thresholds included maximum and minimum values of HSV to ensure the scaffolding colour fell in the range of our colour filter. Then, contours were drawn along the boundary of all the pixels that shared the same colour. The results of the contouring were filtered by eliminating those contours which are too small and were treated as image noise. Background objects, such as cranes and protective boards, sharing the same colour with scaffolding structures can yield contours with abnormal sizes and shapes. These contours can be filtered out by setting the maximum pixel size of the contour. Finally, the filtered contour results were sent to our detectors for object detection.

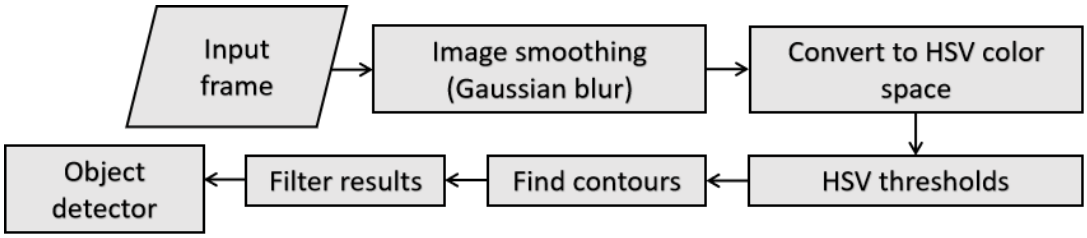


Figure 5.5. The workflow of colour filter

Colour is regarded as a vital object feature in computer vision. Regarding scaffolding detection, a colour filter is helpful to capture relevant scaffolding components whose colour is sharply depicted against the background. A colour filter need to be preset according to scaffold types and illumination conditions. After initially presetting the colour filter, scaffolding images of one project were taken at a same period of a day to ensure HSV value were similar. For instance, the photos of a scaffolding project were taken at 9AM in the morning every two weeks.

Scaffolding sheeting (or covering), consisting of plastic mesh, has been broadly applied in scaffolding projects for the purpose of dust prevention, personnel safety, and site tidiness. Scaffolding sheeting presents a background of a relatively invariant colour, which is applicable for colour filters to remove as background information. Figure XX shows the results obtained from a colour filter. The HSV colour filter controls the colour range from 49 to 71 (hue), 45 to 255 (saturation), 41 to 255 (value). As is shown in Figure 5.6, the colour filter enabled removal of the most irrelevant background information and only the area of scaffolding with bright green sheeting remained.

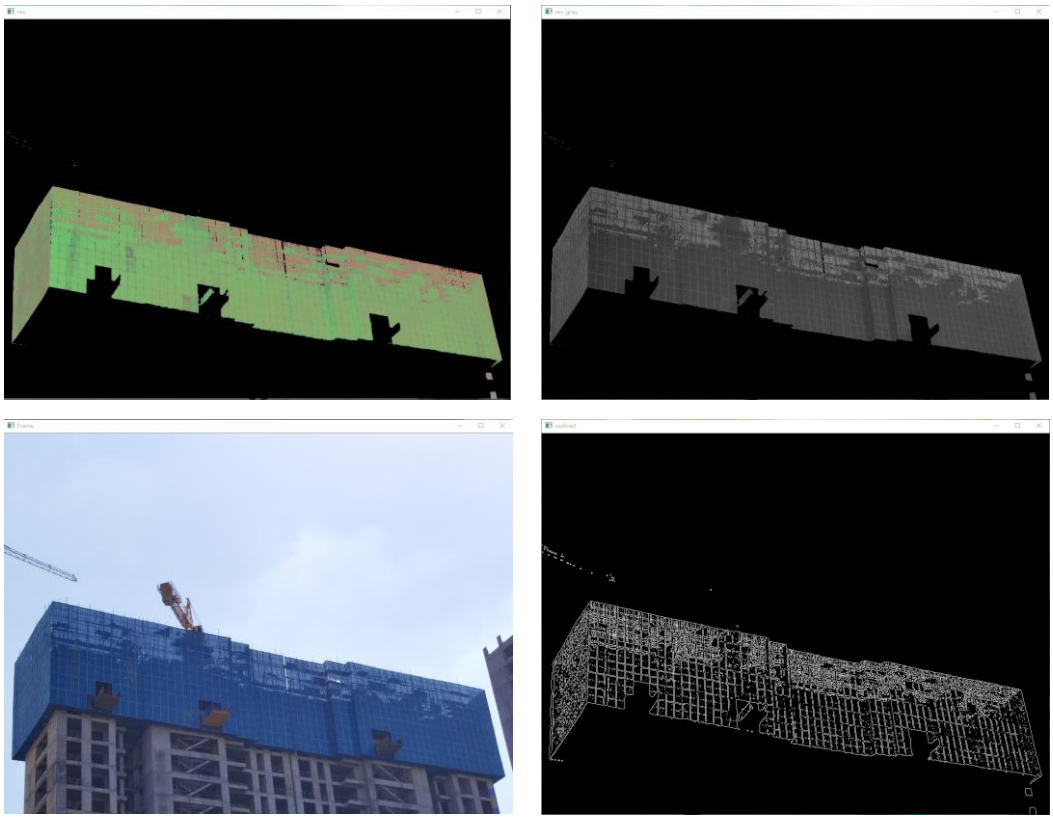


Figure 5.6. The results from colour filter

5.2.4 Data augmentation

Convolutional neural network (CNN) is selected and utilized as one of the object detectors. CNN is a type of deep learning structure and remains the state of the art in object detection for digital images. To achieve excellent performance, deep learning models commonly require a huge training dataset. However, current open-source datasets in construction are mainly focused on workers and vehicles, such as cranes, excavators, and trucks. Also, datasets of scaffolding structure are not available online. Though our own dataset of scaffolding images was manually collected, this dataset was not sufficient to train the detection model.

Data augmentation is used in our approach to increase the variety of the dataset and boost the performance of CNN. It expands the size of the dataset based on the original image dataset through different processing ways or a combination of multiple processing methods. Common image augmentation methods include crop, flip, rotation, and shift, etc.

5.3 Object detector

An object detector plays a role in discovering the locations of objects in an image. The detected objects are usually highlighted with bounding boxes or coloured contours. Another way of highlighting is to cover the regions of detected objects with coloured shadow. An object detector not only highlights the objects in an image, but also provides the coordinates of the object in the image. For example, regarding a rectangular bounding box, the x and y coordinates of four endpoints can be obtained from the object detector. This section introduces two object detectors utilized in our study.

5.3.1 Template matching

Template matching is an approach for scanning and searching a template image region constituting the matching image. In other words, this algorithm moves the template over the whole image and compares the deviation between the template and the regions on the image.

The algorithm of template matching is implemented with 2D convolution, and it works similarly to the sliding window that slides the template image over the candidate image and compares the image patch with the template using specified modes and stores the comparison results. The matching process can be described as searching global minimums in function (1) or finding maximums in function (2).

$$G(x,y) = \sum_{x',y'} (T(x',y') - I(x+x',y+y'))^2 \tag{5.8}$$

$$F(x,y) = \sum_{x',y'} \left(T(x',y') \bullet I(x+x',y+y') \right) \tag{5.9}$$

where T and I denote the template and the image, respectively, and x', y' are the image patches.

The main challenging aspects of template matching are scale and background changes. Images of scaffolding structures present regular and repetitive structural shapes, and it is applicable to use template matching when the bulk of scaffolding image constitutes a large number of matching images of the same size.

5.3.2 Convolutional neural networks (CNN)

As introduced in the previous chapter, CNN is a type of artificial neural network with excellent performance in image processing, and it is the state of art algorithm for image classification, object detection, and facial recognition. Wedges and couplers are crucial components in scaffolding structures, and they present the main nodes and intersections of scaffolding structures in a macroscope view. Therefore, it is proposed to automatically detect and recognize the parts of pins or wedges and couplers by utilizing CNN. YOLOv4 is an advanced model of CNN architecture, and it achieves both outstanding accuracy and recognition speed for object detection. YOLOv4 was implemented as the CNN model for object detection and collected the training and testing data for the model (Bochkovskiy, Wang, and Liao 2020).

Basically, the training data is of great importance in the training process of CNN. As an initial step for CNN model to achieve a good performance, the training data needs to fulfill two crucial aspects. First, there must be a relatively sufficient volume of images. Although the required number of images varies by project scope and complexity, a few hundred images to millions of images is a general figure to achieve high levels of performance. Second, there has to be diversity; training images are required to include the object of interest under various illumination conditions, different colours, orientations, and locations because these aspects allow the machine learning and the understanding of the objects' characteristics and they also enable the machine to classify objects in unknown environments (Gong, Zhong, and Hu 2019).

Since wedges and couplers are selected as our object, the initial step was to collect a relatively sufficient training data of scaffolding images, and then all parts of wedges and couplers were marked as our target objects for recognition, as shown in Figure . The annotation procedure was to manually use rectangular boxes to mark out wedges and couplers one by one in every scaffolding images. For the next step, those annotated training images were fed into the CNN model. Ideally, the CNN model is able to automatically detect and recognize wedges and couplers after model parameter adjustment. The recognition results were detected and highlighted with bounding boxes surrounding each intersection and also came with the confidence rate of every detection in the image.

Figure 5.7. The wedges and couplers marked out for training

5.4 Component quantification, graph analysis and productivity estimation

Scaffolding quantity measurement to a great extent relies on scaffolding component recognition. Accurate and successful component recognition provides a solid

foundation for scaffolding quantity measurement. In our approach, the number of scaffolding intersections is measured. And by combining the project design parameters from site managers, the volume of the scaffolding project can be estimated, shown in the images. The common unit reflecting scaffolding productivity in construction industry is m^3/day, so the scaffolding productivity is estimated with the unit of m^3/day.

A mathematics models was established to estimate the surface area of our target scaffolding project. The surface area of the scaffolding project was initially assumed to be rectangular, which meets regular construction requirements and was feasible for modelling. The total number of central intersections in scaffolding structure was set as C, which excluded those marginal points on edges and were highlighted in red. The value of C was proposed to be obtained automatically from our computer vision program, including the result of true positive plus false negative. With regard to two common types of scaffolding structure, two different models through graph analysis were proposed, which were displayed in Figure 5.8(a)(b). The first model was built regarding the scaffolding projects with diagonal structure (see Figure 5.8(a)). The module highlighted in orange was taken as the basic scaffolding surface unit and the red points indicated the scaffolding intersections which were the objects for recognition. The other model targeted the scaffolding projects with rectangular structures (see Figure 5.8(b)) and took the module highlighted in blue as the surface unit and the red points indicated the scaffolding intersections as well. A successful object detector was proposed to precisely detect the scaffolding intersections (red points). Once the total number of intersections C was obtained, the number of scaffolding surface unit Q could be calculated through combining the project design parameters. The number of layers of surface unit was defined as N and the value of N could be gained from project design parameters.

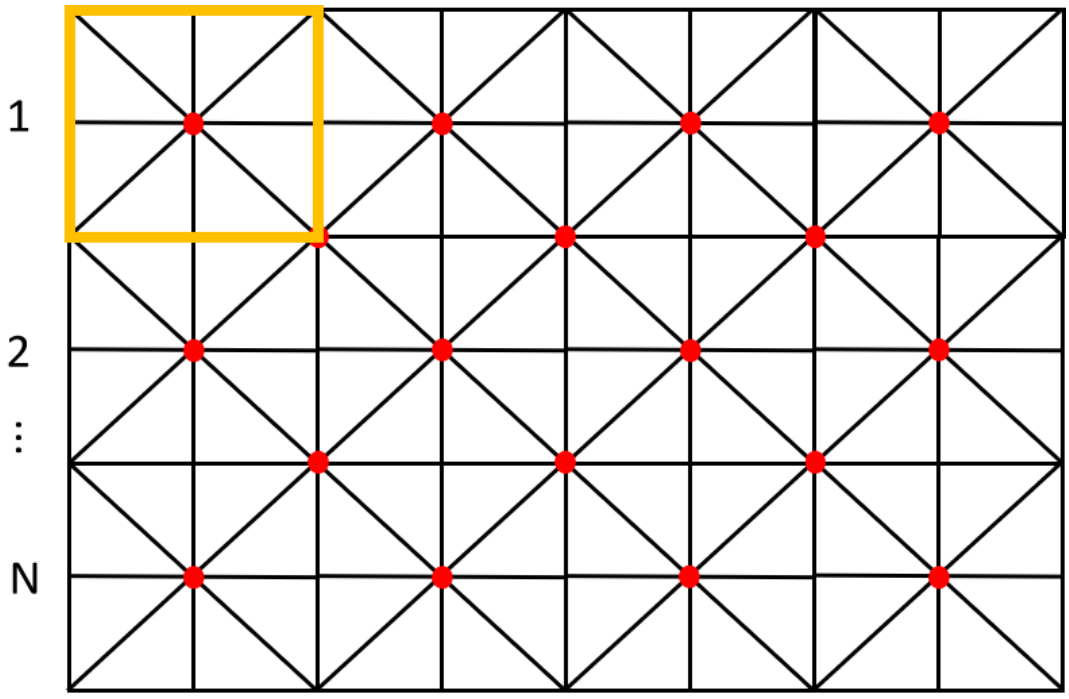


Figure 5.8(a) The model of diagonal scaffolding structures

For the model with diagonal structure, the number of scaffolding surface unit Q could be expressed in Equation 5.11.

$$C = Q + (N-1)\left(\frac{Q}{N} - 1\right) \tag{5.10}$$

$$Q = \frac{C * N + N^2 - N}{2N - 1} \tag{5.11}$$

For the model with rectangular structure, the number of scaffolding surface unit Q could be expressed in Equation 5.12.

$$Q = (N+1) * \left(\frac{C}{N} + 1\right) \tag{5.12}$$

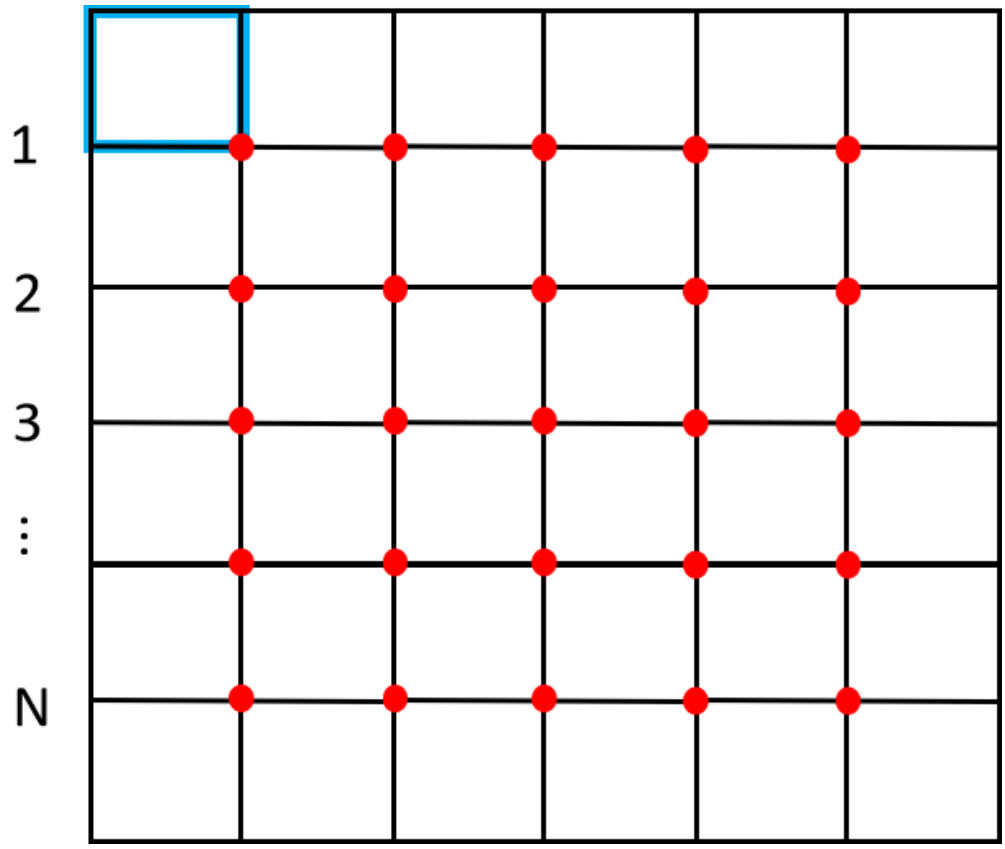


Figure 5.9(b) The model of rectangular scaffolding structures

Additionally, v denotes the volume of scaffolding surface unit and it takes the width parameter of a project into account, which cannot be identified in a 2D image. The total volume of the scaffolding structure is defined as V. As 2D images only present the length and height of one surface of the structure and cannot display the width information, the value of v varies on different scaffolding projects and is gained from project design parameters. T denotes the time that the scaffolding project consumed, and P denotes the actual scaffolding productivity. Hence, the total volume of scaffolding structure V can be obtained as the volume of scaffolding surface unit v multiplying the number of scaffolding surface unit Q. The average scaffolding productivity can be expressed as V divided by the time T that this scaffolding project consumed, shown in Equation 5.13 and Equation 5.14.

$$V = v * Q \tag{5.13}$$

$$P = \frac{V}{T} = v * \frac{Q}{T} \tag{5.14}$$

5.5 Case study and analysis

This section illustrates the experiment setup and the process of scaffolding detection. The experiments were conducted under PYTHON 3 environment and open-source computer vision library OPENCV. The training process of CNN was conducted on the basis of Google TensorFlow open-source library and is capable of running on Google Colaboratory as well as our local computer Dell Workstation, with a 1.9 GHz processing unit and a 32 GB RAM. To effectively validate our proposed research model, a database of scaffolding images was collected, and one case study was conducted and analysed. In addition, the outputs of every step for scaffolding detection and quantity measurement were displayed.

5.5.1 Data collection, annotation and augmentation

As our research mainly relied on the technique of computer vision, the research data included an enormous collection of scaffolding images. Great efforts were made in the data collection, and 376 images were captured in 21 construction sites, including industrial plant maintenance and residential buildings. The capture devices, listed in Table 5.1 included smart phones ranging from 8 million to 40 million pixels of camera resolution and a handheld GoPro camera, which provided high digital image stream. Also, with the help of a drone, it is feasible to capture photographs of scaffolding projects from a greater distance and a higher platform with few obstacles. These collected photos covered different scaffolding projects under a variety of scaffold colours, lighting conditions, and shooting angles. The shooting distances ranged from 20 meters to 200 meters. The diversity of captured images can help enhance the robustness of the proposed model in object detection. Moreover, the project design parameters of six scaffolding projects were collected for validation.

Before the training process of YOLOv4, the CNN model, all the input data went through the annotation process. Every intersection of scaffolding structure in the images was labelled with a bounding box and the category tag of intersection. The annotation process was implemented via the Python package *LabelImg*. Additionally, the input images were resized to 416×416 pixels to fit the requirement of YOLOv4.

Once the dataset was resized and annotated, cropped, saturation changes, brightness changed and noise addition, four type of augmentation techniques were applied to the dataset, which increased the size of dataset four-fold. The number of training images before and after image augmentation were 120 and 477, respectively.

Table 5.1. The specification of the image capture devices.

Camera	Leica 40MP, Cine Camera	GoPro 20MP	DJ 1/2.3" CMOS12MP
Photo definition	3648*2736	3648*2736	4000*3000
Photo format	JPEG, RAW		

5.5.2 Case study and discussion

In the case study, six different scaffolding projects, which included their scaffolding design parameters, were selected from the collected dataset for validation. In order to evaluate the proposed model, the scaffolding design parameters and the actual productivity of these six projects were regarded as the ground truth for the comparison with the results from the automatic model.

The YOLOv4 CNN model, template matching with image processing package (TMIP), (Gaussian blur, edge and line detection, colour filter), and the method of template matching (TM) these three object detectors were evaluated in this case study. Every testing image only contained one scaffolding project. Moreover, these testing images were captured under different shooting angles and illumination conditions. The shooting distance ranged from 20 meters to 200 meters, and also different colours and types of scaffolds were involved in the testing images. The specifications of six project images are displayed in Table 5.2. The results were displayed in Figure 5.10-5.15.

Table 5.2. The specification of six scaffolding projects in the case study

Project #	Scaffolding Volume(m3)	Shooting Distance (m)	Scaffolding colour	Structure type
Project 1	16340	20	Yellow	Diagonal
Project 2	4200	20	Yellow	Rectangular
Project 3	6400	200	Blue	Diagonal
Project 4	1120	30	White	Rectangular
Project 5	7920	60	Yellow	Diagonal
Project 6	1440	20	Brown	Rectangular

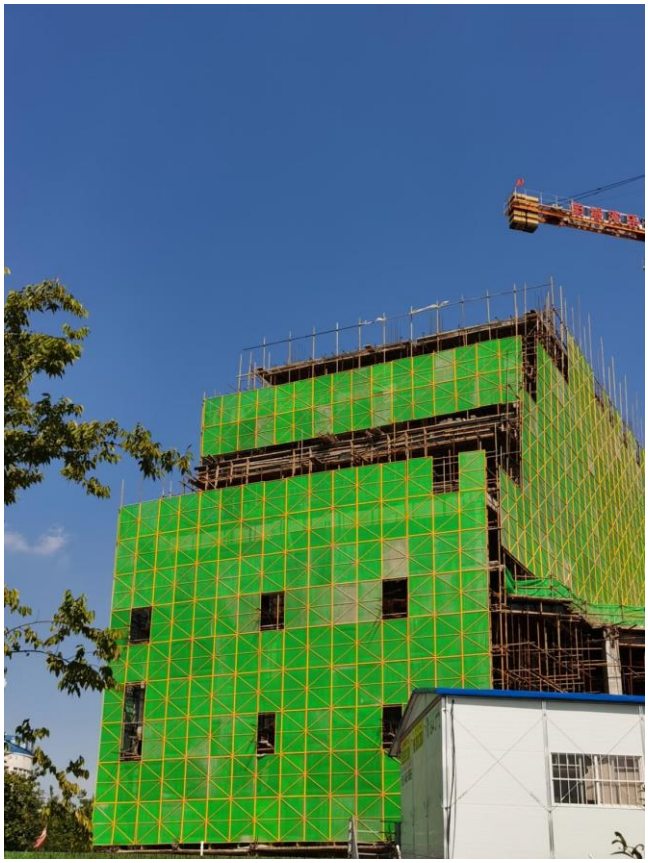


Figure 5.10(a). The original photo of Project 1

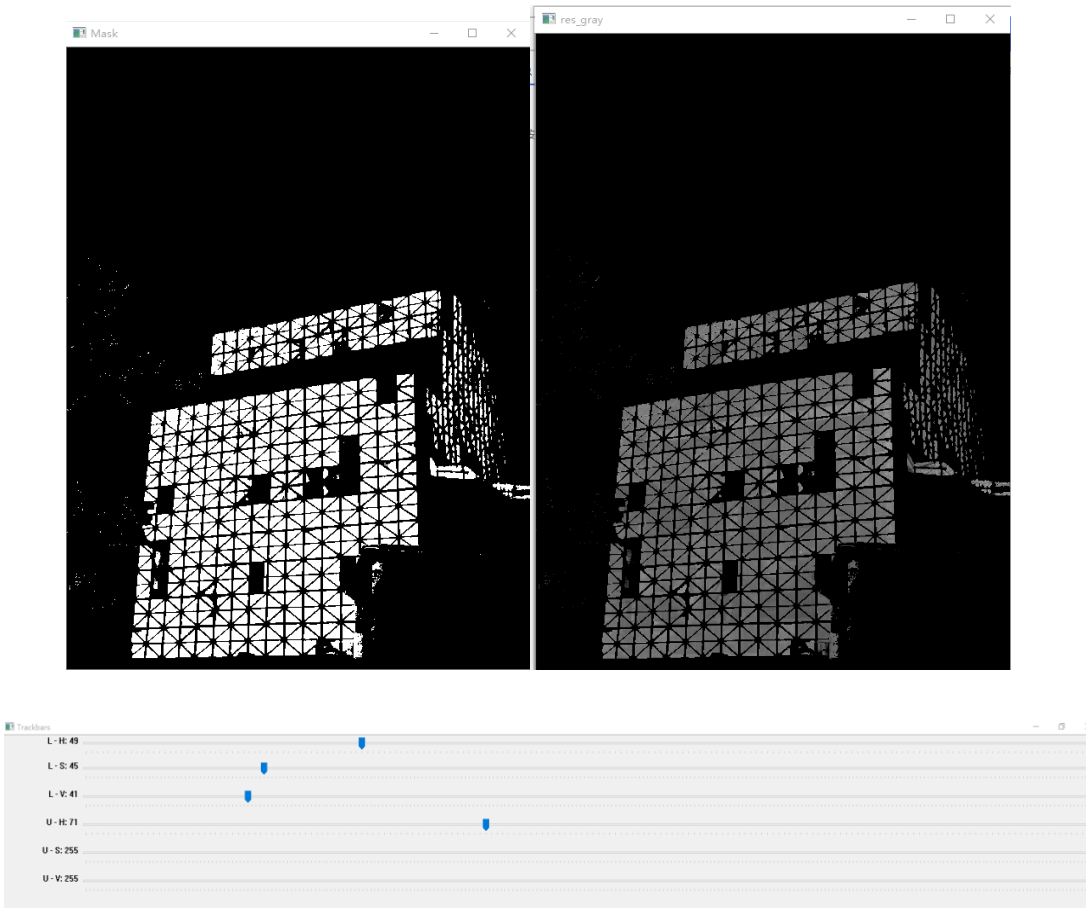


Figure 5.11(b). The processed photo of Project 1 and the colour filter

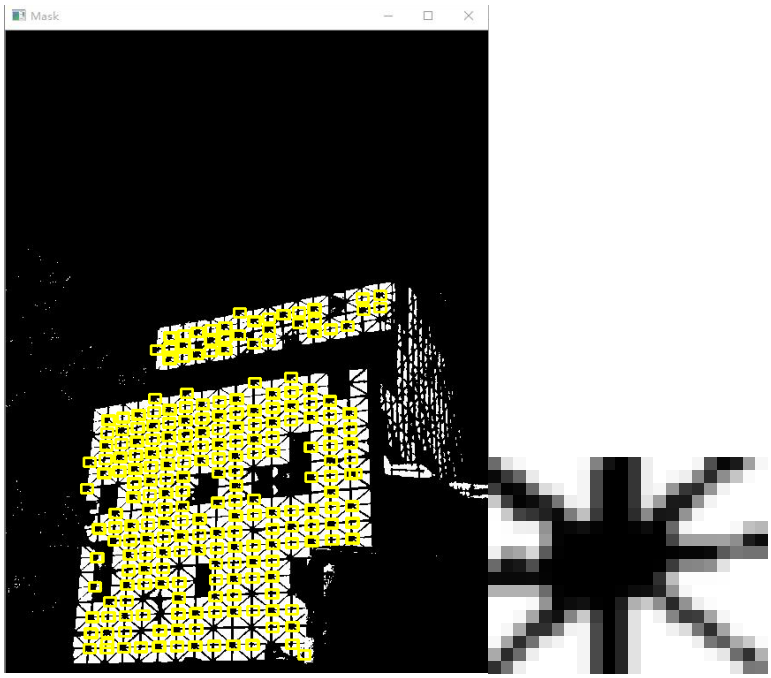


Figure 5.12(c). The detection result of TMIP with the template on the right.

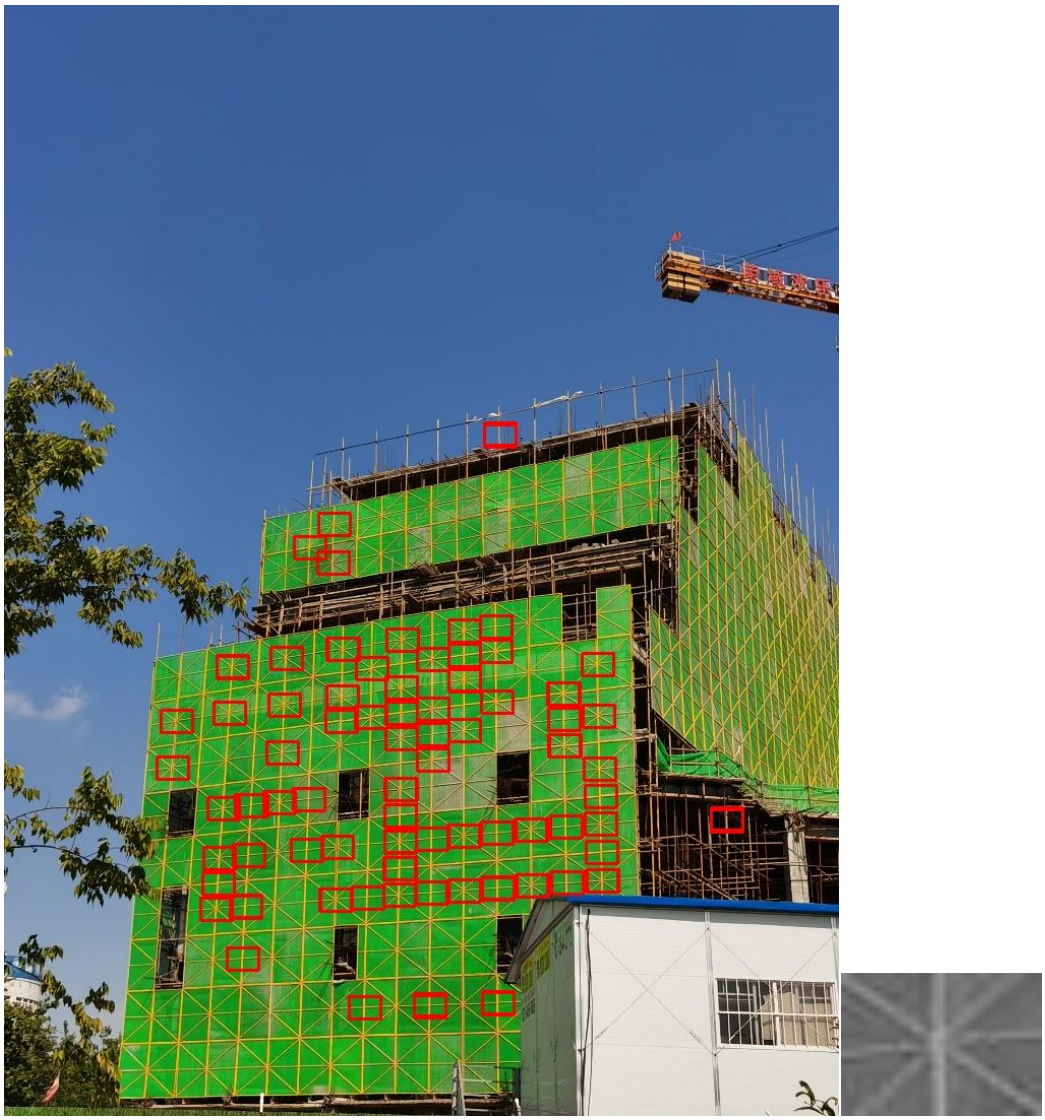


Figure 5.13(d). The detection result of TM with the template on the right.

Figure 5.14(e). The detection result of YOLOv4 CNN model.

Figure 5.10 (b)(c) displays the results from TMIP. The image processing package (colour filter: Hue: 49 to 71; Saturation: 45 to 255; Value: 41 to 255) effectively blocks the background noise and present the structure of scaffolding.

Figure 5.10(c) exhibits the results of scaffolding intersections (wedges and couplers) detected by the colour filter. Most of scaffolding intersections under the scaffolding sheeting can be detected except those intersections far away from the shooting camera that lacks sharp colour deviation.

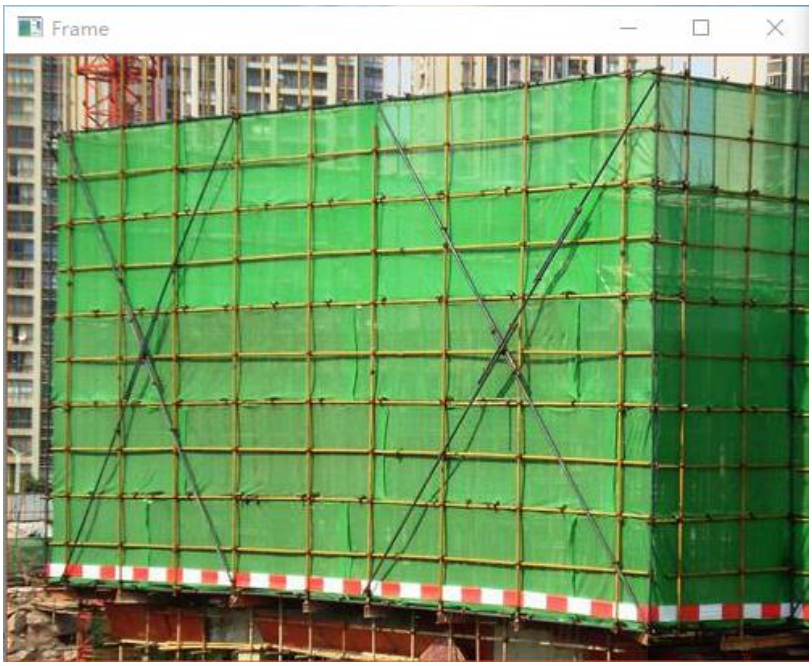


Figure 5.11(a). The original photo of Project 2

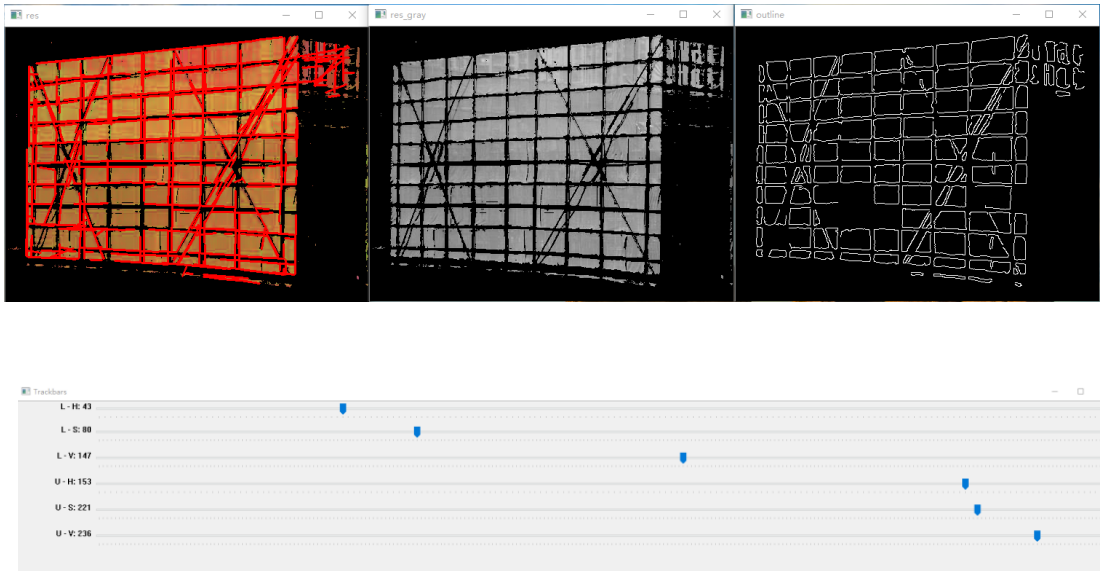


Figure 5.11 (b). The processed photo of Project 2 and the colour filter

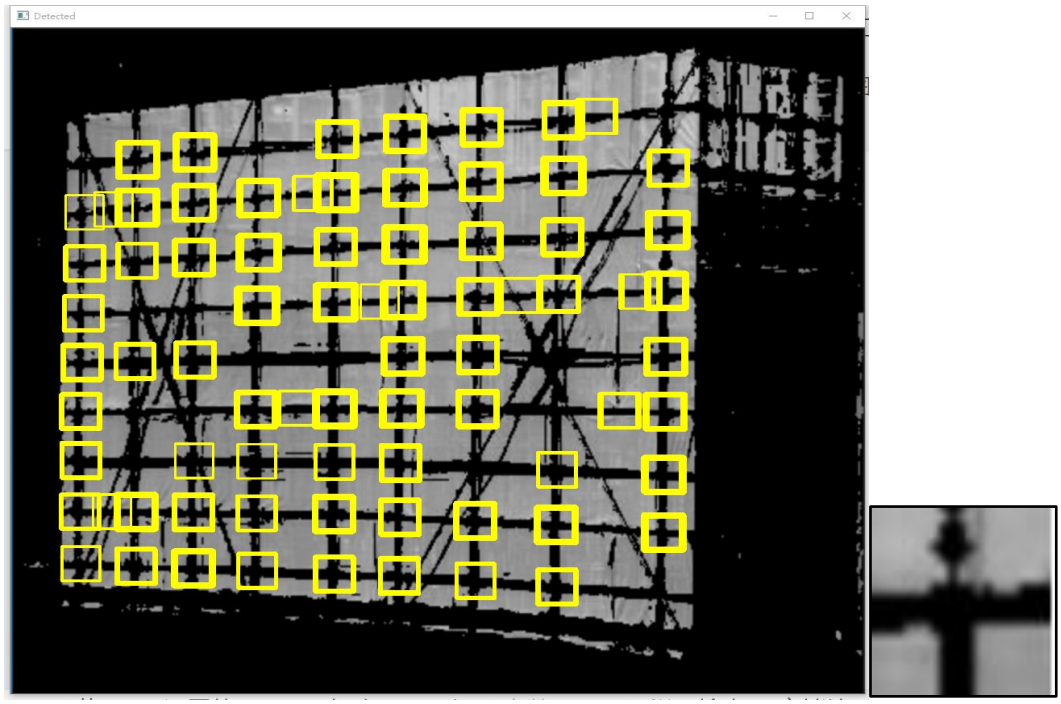


Figure 5.11(c). The detection result of TMIP with the template region on the right

Figure 5.11(d). The detection result of TM with the template photo on the right

Figure 5.11(e). The detection result of YOLOv4 CNN model

As shown in Figure 5.11, the method of TMIP performs better than TM method. However, it fails to detect the region reinforced with diagonal structures. The YOLOv4 CNN model outperforms the other two object detectors that not only precisely detect the joints reinforced with diagonal structures as well as a large number of intersections, but also reveals the joints on the lateral surface under different shape and illumination condition.

Figure 5.12(a) The original photo of Project 3

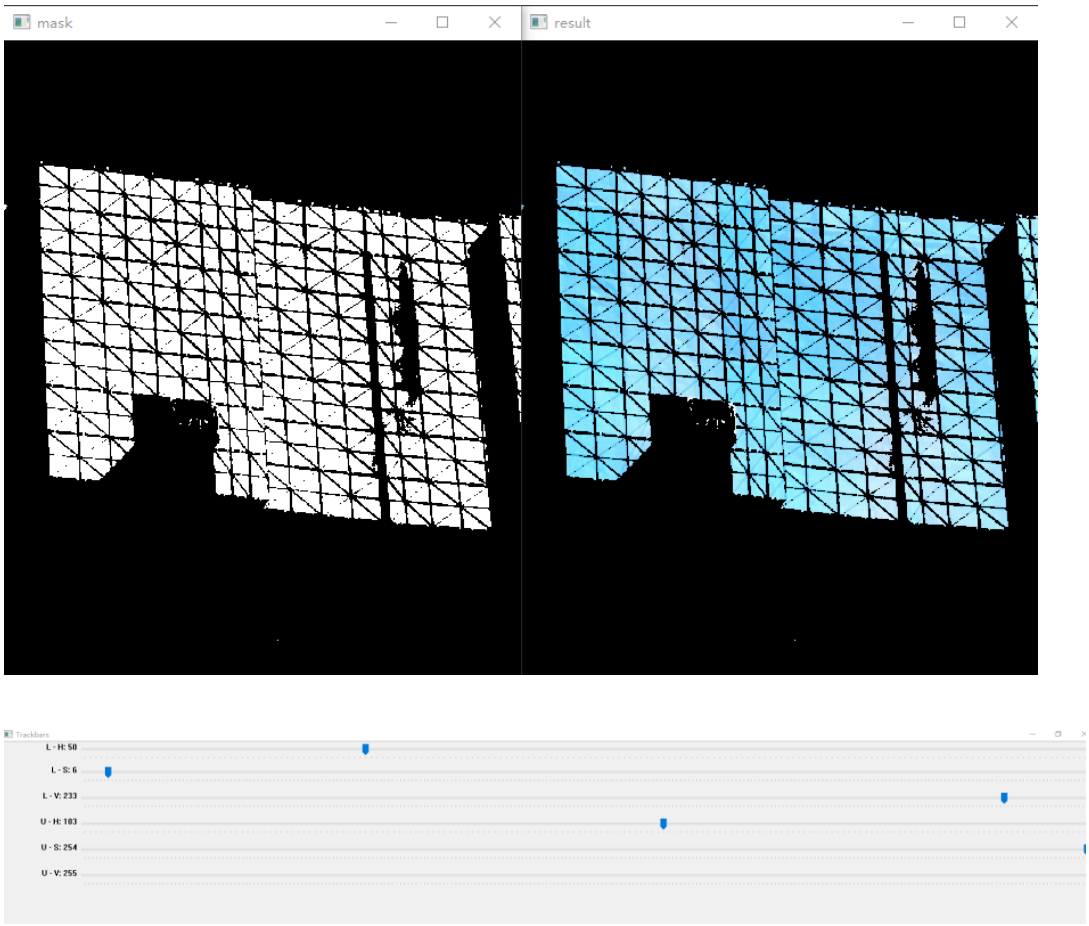


Figure 5.12(b) The processed photo of Project 3 and the colour filter

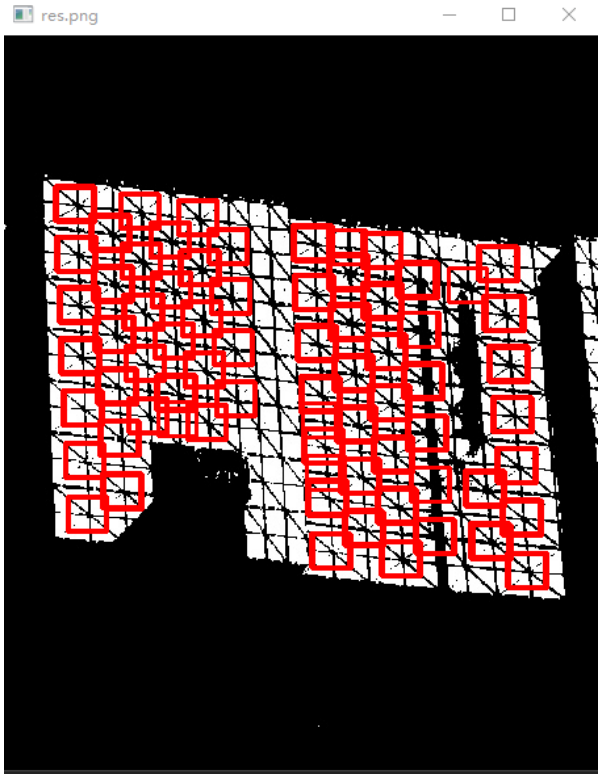


Figure 5.12(c) The detection result of TMIP with the template region on the right

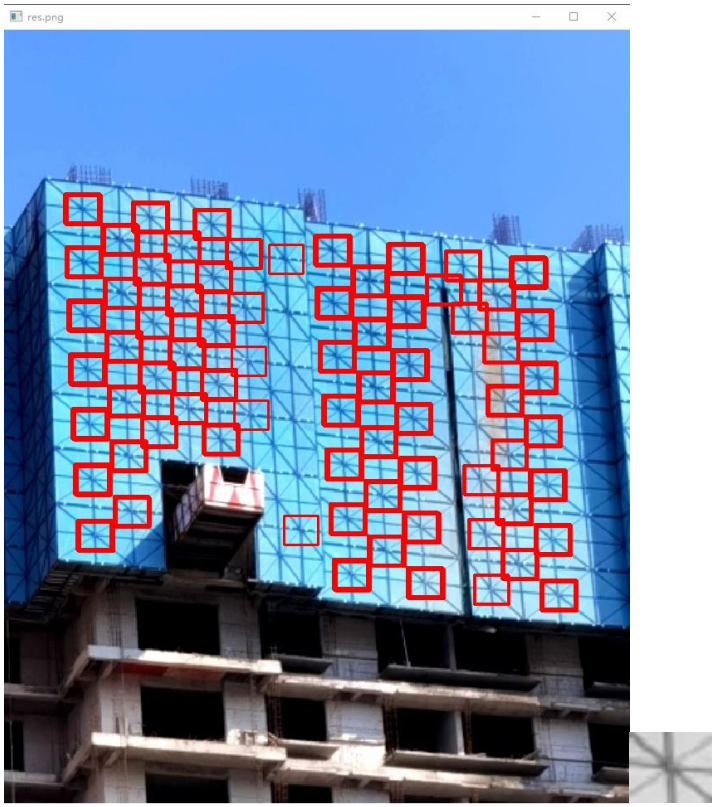


Figure 5.12(d) The detection result of TM with the template photo on the right

Figure 5.12(e) The detection result of YOLOv4 CNN model

Figure 5.13(a) The original photo of Project 4

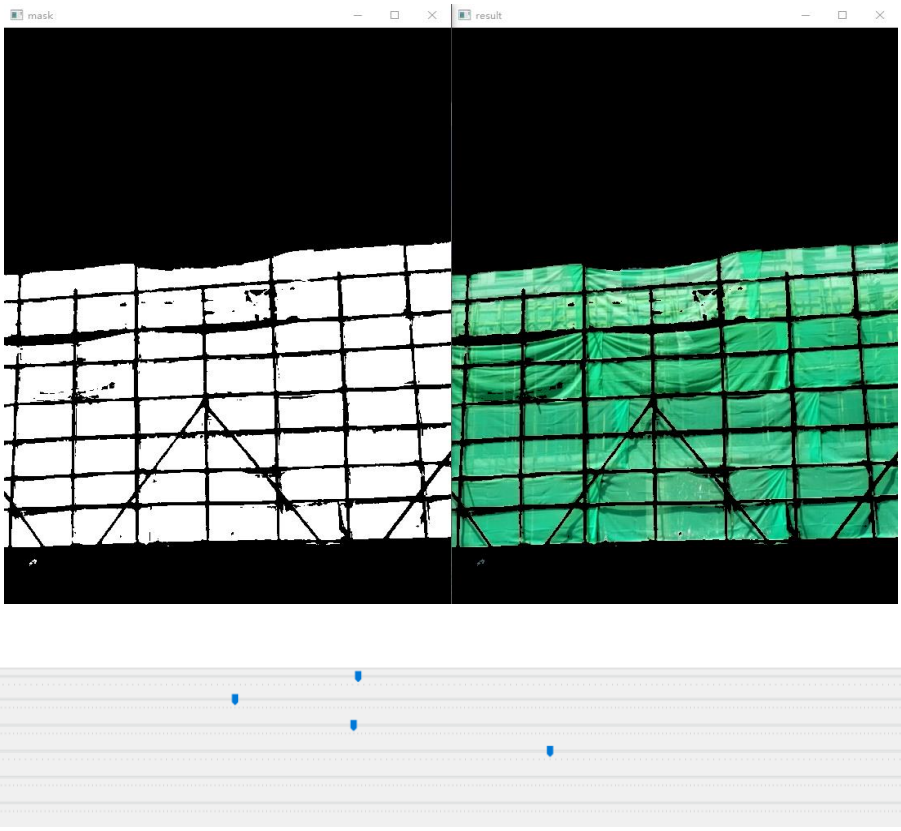


Figure 5.13(b) The processed photo of Project 4 and the colour filter

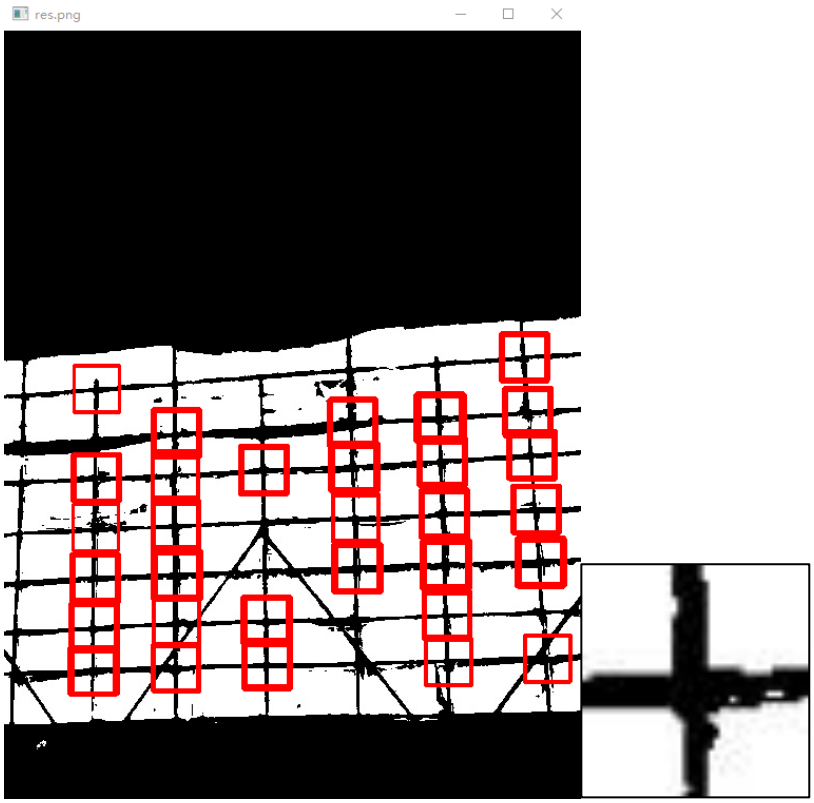


Figure 5.13(c). The detection result of TMIP with the template region on the right

Figure 5.13(d) The detection result of TM with the template photo on the right

Figure 5.13(e) The detection result of YOLOv4 CNN model

Figure 5.14(a). The original photo of Project 5

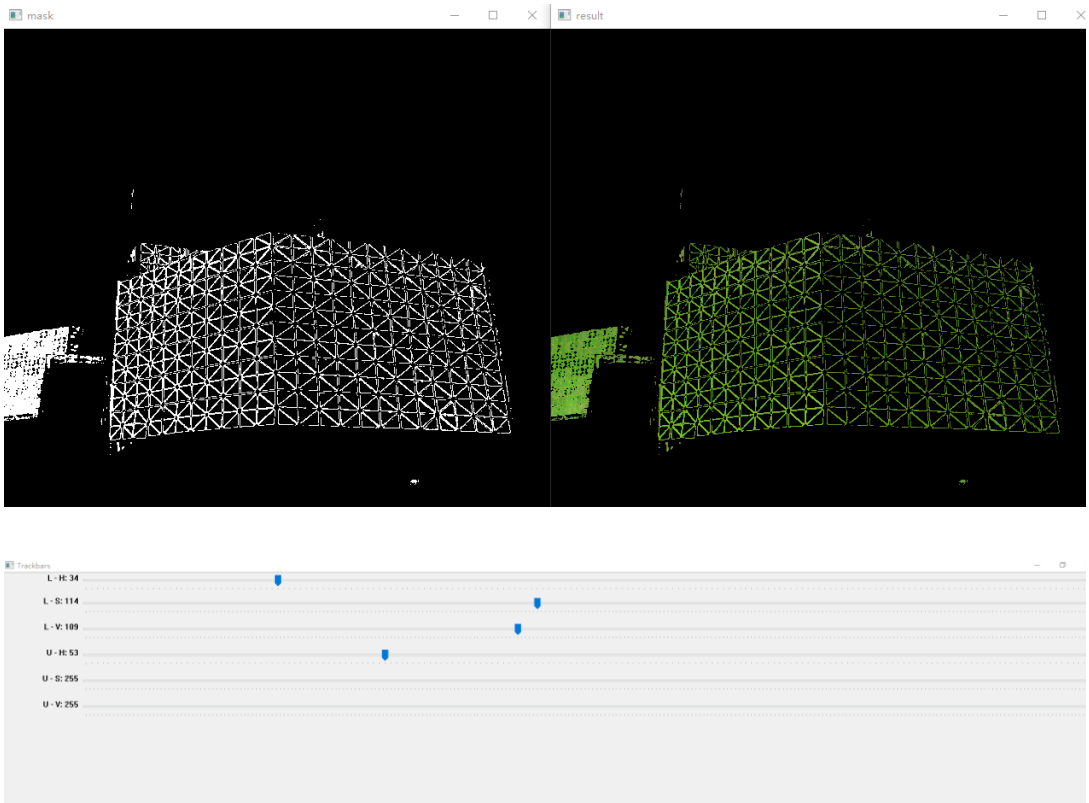


Figure 5.14(b). The processed photo of Project 5 and the colour filter

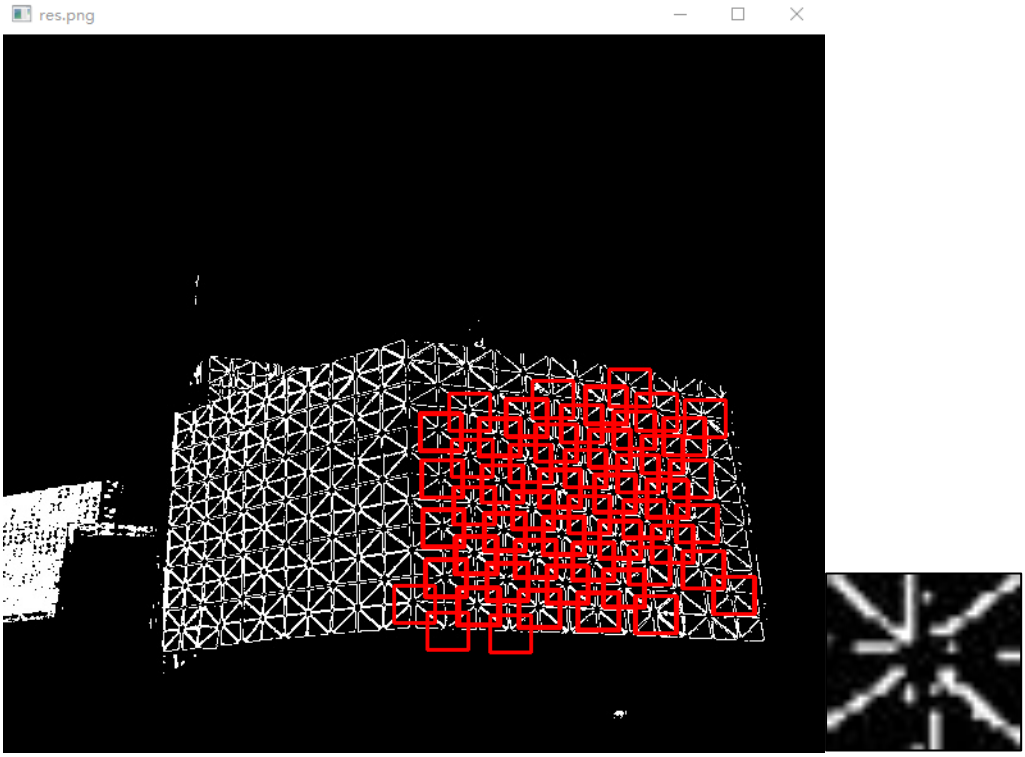


Figure 5.14(c). The detection result of TMIP with the template region on the right

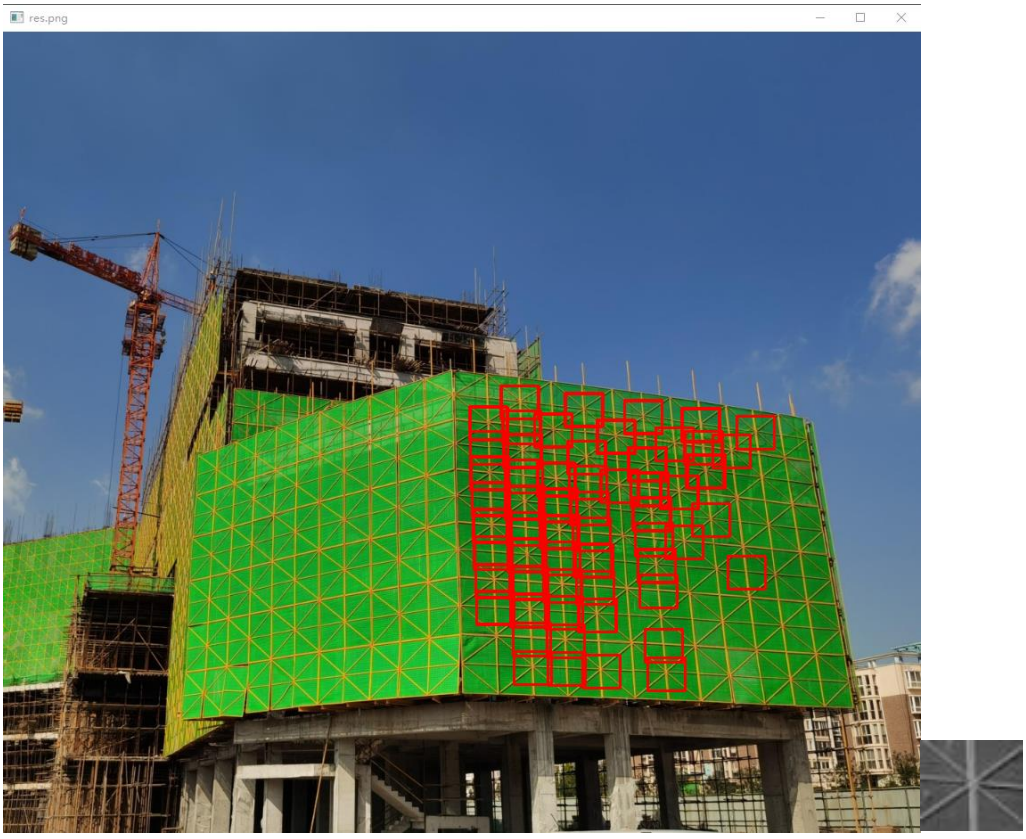


Figure 5.14(d). The detection result of TM with the template photo on the right

Figure 5.14(e) The detection result of YOLOv4 CNN model

In Figure 5.14, the scaffolding project consists of two surfaces with a little angle difference in the middle. Compared to the other two detectors, YOLOv4 CNN model shows relatively good robustness of detecting scaffolding intersections. The left structure surface cannot be recognized by TM and TMIP because the angle difference of two scaffolding surfaces causes a considerable divergence between the applied template and the actual intersections.

Figure 5.15(a) The original photo of Project 6

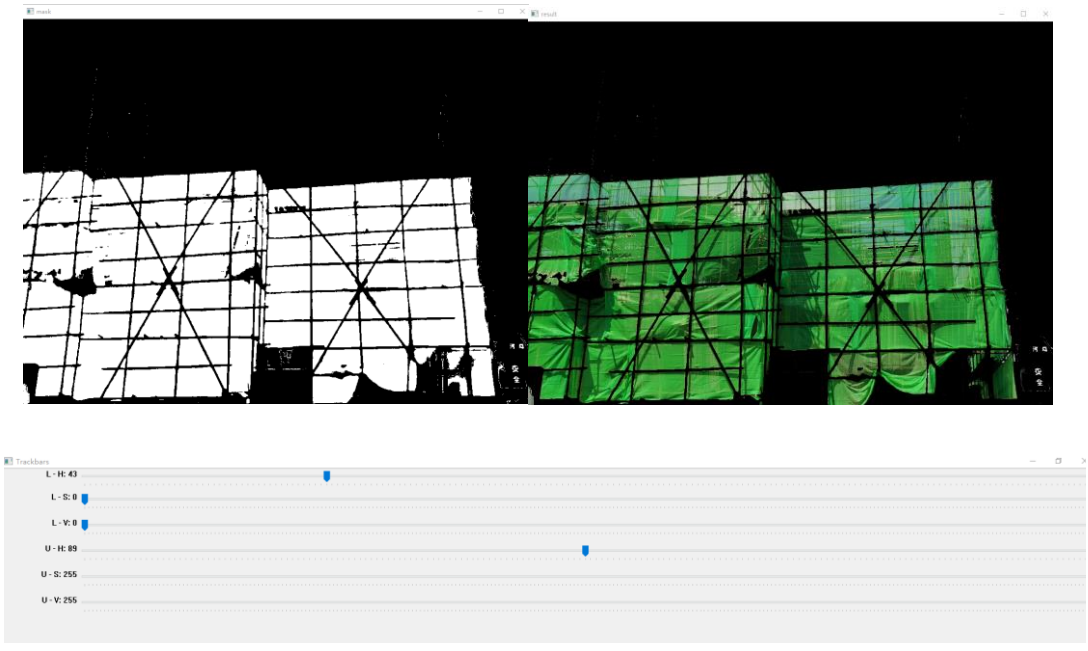


Figure 5.15(b) The processed photo of Project 6 and the colour filter

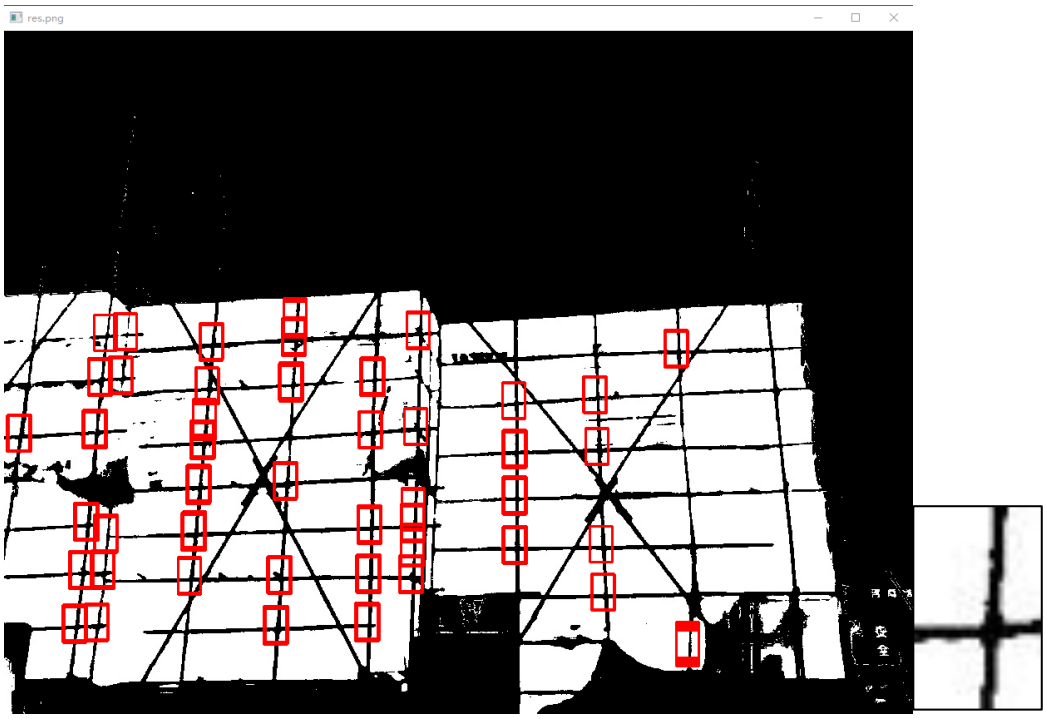


Figure 5.15(c) The detection result of TMIP with the template region on the right

Figure 5.15(d) The detection result of TM with the template photo on the right

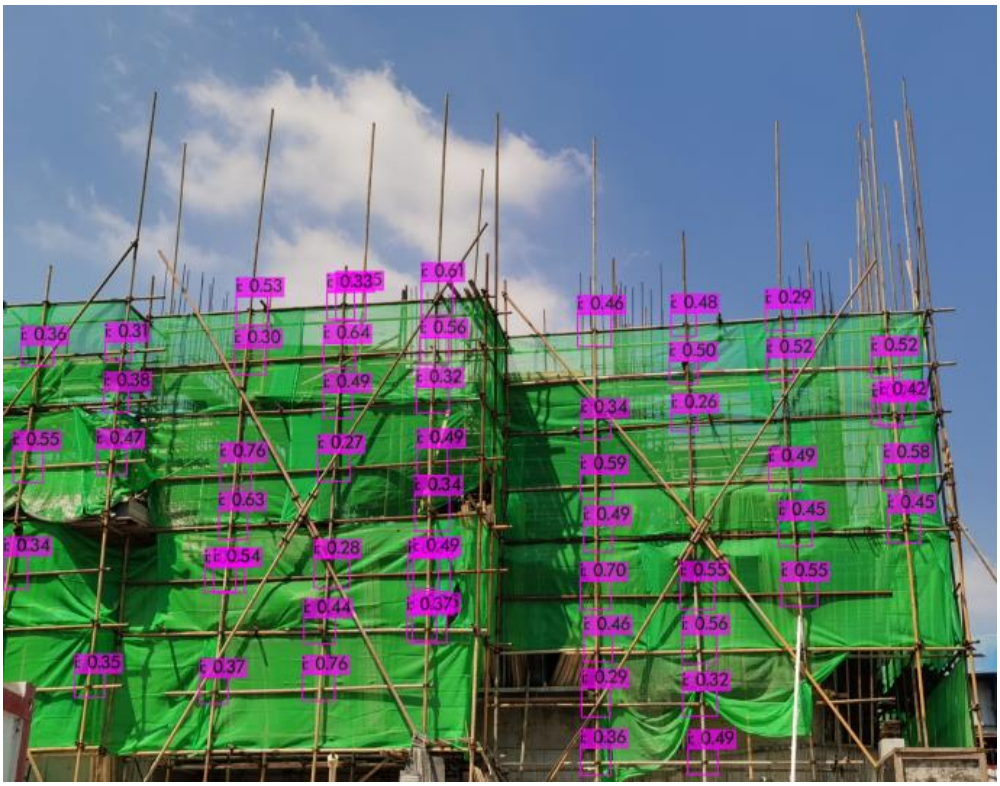


Figure 5.15(e) The detection result of YOLOv4 CNN model

In Figure 5.15, the TM and TMIP detectors achieve relatively poor performance because there is complicated background noise caused by shadows and scaffolding sheeting. Also, irregular shapes of scaffolding structures might impact the application of TMIP. Comparatively, YOLOv4 CNN model presents better robustness for detection.

Compared with the ground truth that were manually labeled and collected; the performance of our approach is listed in the Table 5.4-6. The Precision, Recall and F1 Score, widely accepted performance indicators in object detection, were used in our research. Also, Figure 5.16-18 displays the performance results.

$$\text{Pre}cision = \frac{\text{TP}}{TP + \text{FP}} \tag{5.15}$$

$$Re\,c\,all = \frac{TP}{TP + FN} \tag{5.16}$$

$$F_1 score = \frac{2}{recall^{-1} + precision^{-1}} = 2 \bullet \frac{precision \bullet recall}{precision + recall} \tag{5.17}$$

Table 4.4 The performance of TMIP

#Project	TMIP Detector					
	True Positive(TP)	False Positive(FP)	False Negative(FN)	Precision	Recall	F1 Score
Project 1	129	52	6	0.71	0.96	0.82
Project 2	67	9	14	0.88	0.83	0.85
Project 3	68	5	15	0.93	0.82	0.87
Project 4	31	0	13	1.00	0.70	0.83
Project 5	50	2	61	0.96	0.45	0.61
Project 6	36	5	20	0.88	0.64	0.74

Table 4.5. The performance of TM

#Project	TM Detector					
	True Positive(TP)	False Positive(FP)	False Negative(FN)	Precision	Recall	F1 Score
Project 1	45	24	90	0.65	0.33	0.44
Project 2	31	10	50	0.76	0.38	0.51
Project 3	71	1	12	0.99	0.86	0.92
Project 4	25	10	19	0.71	0.57	0.63
Project 5	34	2	77	0.94	0.31	0.46
Project 6	8	2	48	0.80	0.14	0.24

Table 4.6 The performance of YOLOv4

#Project	YOLOv4 CNN model					
	True Positive(TP)	False Positive(FP)	False Negative(FN)	Precision	Recall	F1 Score
Project 1	75	0	60	1.00	0.56	0.71
Project 2	66	4	15	0.94	0.81	0.87
Project 3	62	0	21	1.00	0.75	0.86
Project 4	38	0	6	1.00	0.86	0.93
Project 5	61	4	50	0.94	0.55	0.69
Project 6	46	2	10	0.96	0.82	0.88

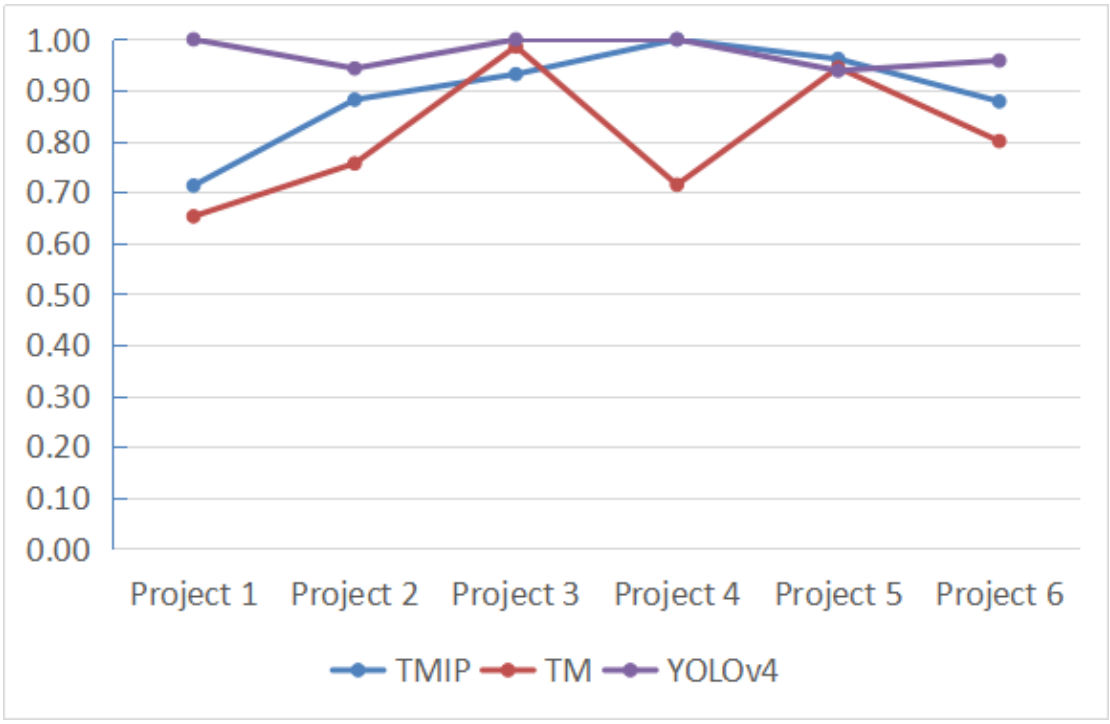


Figure 5.16 The Precision of the three detectors

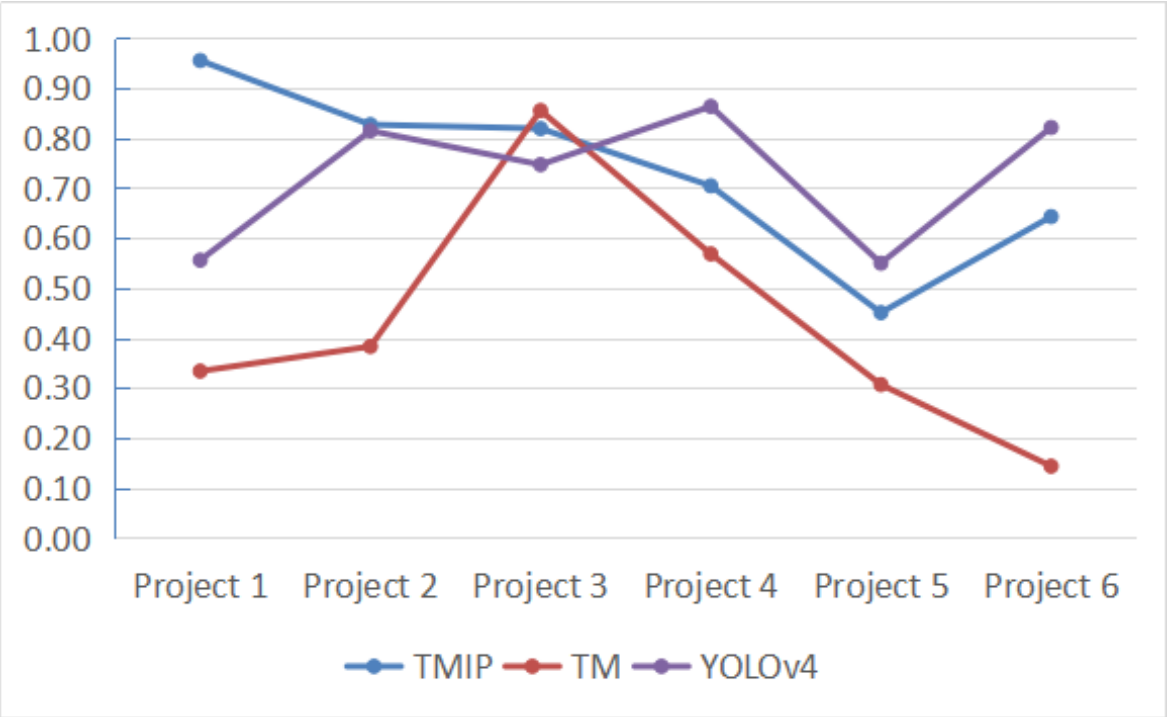


Figure 5.17 The Recall of the three detectors

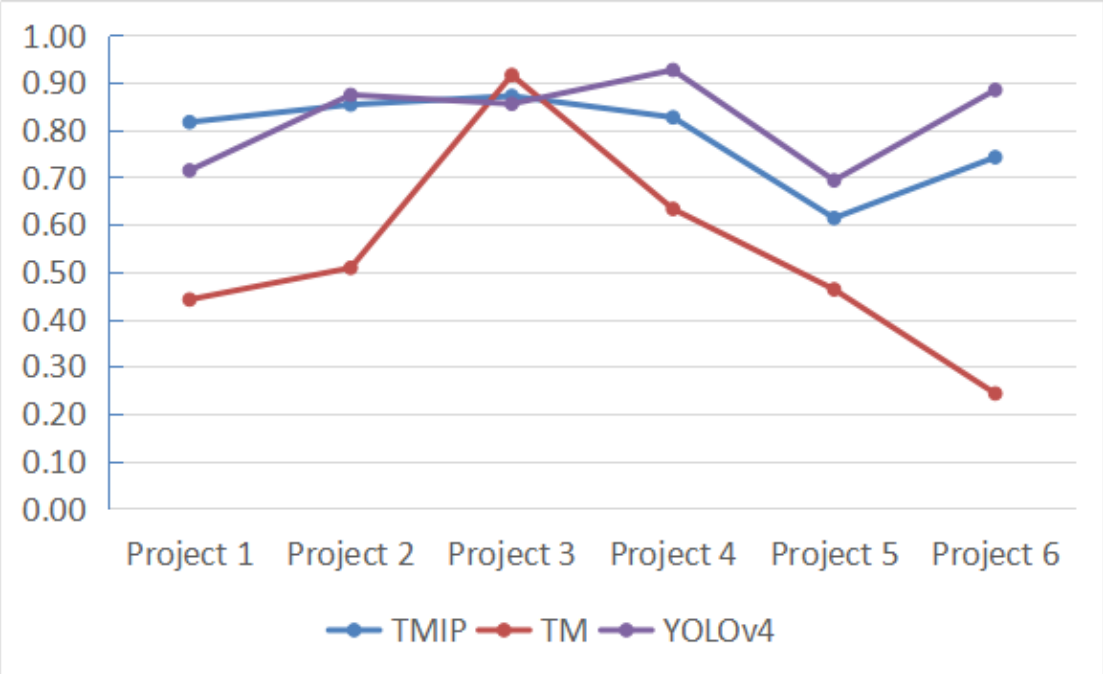


Figure 5.17 The F1 score of the three detectors

YOLOv4 CNN model keeps relatively high-level (over 0.9) Precision in the six projects which means over 90 percent of detected positive results are correct. However, this Recall rate of YOLOv4 drops to about 0.55 in Project 1 and Project 5 and remain around 0.8 for the rest of projects, indicating that it fails to identify 20% to 45% intersections in the images. This might be caused by the CNN's limited capability of detecting tiny objects in a image. YOLOv4 achieves at least 0.7 F1 score in the six projects and presents better stability than TM and TMIP.

For the TMIP detector, a great deal of background noise and irrelevant colors have been removed with the help of image processing package. Consequently, there is a considerable performance elevation after the application of image processing. TMIP produces excellent results when scaffolding intersections with regular shapes and united size appear in the image. However, a tilted shooting angle can distort intersection regions and makes the close regions present bigger and the far regions present smaller. This impact together with the illumination changes result in the difficulty in matching the selected template with the intersection regions in a image.

The performance of TM fluctuated in the six scaffolding projects, and it produces low-quality results in the most projects except in Project 3. Impacted by the background noise, TM is not able to effectively identify many intersections in Project 1, Project 5 and Project 6, resulting in the low scores in Recall and F1 score. The fluctuation in performance is not a desirable feature for a robust object detector.

YOLOv4 CNN model is selected for the next step of productivity measurement, because YOLOv4 achieves the highest F1 scores in Project 2, 4 5 and 6 and presents relatively excellent robustness and stability than the other two detectors. Firstly, the YOLOv4 model automatically output the number of detected intersections C (C=TurePositive+FalsePositive). Next, the number of scaffolding surface unit Q is derived from the equations of component quantification. Then, the total volume of scaffolding project V equals the volume v of scaffolding surface unit times Q. By comparing the V results with the ground truth project volume, the predication error is gained in percentage. Finally, with the time each specific project consumed, the productivity of scaffolding project is calculated. For example, in Project 1, $C = 75$, the volume v of scaffolding surface unit and the number of layer of central intersections N is obtained from project design parameters, $v = 1 \times 1 \times 2m^3$ and $N = 11$.

Then the value of Q is gained $Q = 44.5$, and the scaffolding volume $V = v \bullet Q = 300m^3 \times 44.5 = 13357.1m^3$. Moreover, the ground truth project volume equals 16340 cubic meters. Finally, the project productivity $P = V / T = 13357.1m^3 / 34 = 393m^3 / day$. The results of productivity measurement are listed in the Table 4.7.

Table 4.7. The productivity measurement of six scaffolding projects

Project #	Structure type	C=TP+FP	V (m3)	N	Q	V(m3)	Ground Truth (m3)	Prediction error	Time consumed (day)	Productivity (m3/day)
1	Diagonal	75	300	11	44.5	13357.1	16340	-18.25%	34	393
2	Rectangular	70	40	9	87.8	3511.1	4200	-16.40%	11	319
3	Diagonal	62	160	7	36.6	5858.5	6400	-8.46%	20	293
4	Rectangular	38	20	8	51.8	1035.0	1120	-7.59%	7	148
5	Diagonal	65	120	6	38.2	4581.8	6742	-31.86%	20	230
6	Rectangular	48	20	8	63.0	1260.0	1440	-12.50%	8	158

As shown in the Table 4.7, the predicted project volume is smaller than the ground truth in the six projects and the percent predication errors range from -32% to -7.59%. This occurs because the YOLOv4 detector outputs detected intersections less than the ground truth, rather than generating results more than the ground truth.

There are some limitations exposed in this study during the case study. First, there is still a large space for the elevation of detection accuracy and reliability. As displayed in the detection results, some intersection regions with distortion or background noise could not be recognized by the model. Second, the model is developed under the assumption that the scaffolding project belongs to one of diagonal and rectangular structures, while a scaffolding project contains more complexity in structure and variety of background. Third, this study only places research interest and build models on scaffolding projects. However, other construction works or structures that have regular and repetitive shape, such as concert structure, can also be explored in research.

With the discussion about limitations, the future work will be placed as follows. First, efforts need to be devoted to the dataset expansion. The current captured data is not sufficient if the performance of the CNN model need to be optimised. A training dataset of hundreds to ten thousand of images as well as with a significant variety is ideal for object detection. The work of collecting a large variety of scaffolding photos is one of the research directions. Second, the optimization of our model. The detection performance is elevated by adjusting the hyper parameters of the CNN model or selecting the latest model, which requires iterative practices. Third, steel, concert and other structures that have regular shapes and obvious outlines and features for recognition and detection could be studied viewed as a research direction in the future.

5.6 Conclusion

This study has presented a computer-vision based method for scaffolding component quantity and productivity measurement. To automatically identify scaffolding components and measure scaffolding productivity, this study has taken 2D color images of scaffolding structure as input and has aimed to extract scaffolding intersections (couplers or wedges) from the structures. Image processing techniques, including Gaussian blur, edge and line detection, and colour filter have been employed to remove background noise and extract useful features for detection. Additionally, for the training of a deep CNN-based learning, data augmentation was utilized to boost the size and variety of the dataset. Two object detectors, including template matching and CNN, were selected for the detection of scaffolding intersections. The object detector was designed to generate the 2D locations and the number of scaffolding intersections as well as to highlight the results in the image. To build connections with scaffolding volume and project productivity, a mathematical model was established through graph analysis. Regarding diagonal and rectangular two different structures, two models were built. The number of scaffolding surface units was inferred from the number of scaffolding intersections. Consequently, the volume and productivity of a scaffolding project was calculated by considering project design parameters in accordance with the number of scaffolding surface units. A case study was conducted to validate and evaluate the detection process. YOLOv4 CNN model, TM, and TMIP three object detectors were tested in the case study. Six individual project images were used as the input, and the performance indicators precision, recall and F1 score, were

employed to evaluate the detectors. A YOLOv4 CNN model presented good stability and performance in these indicators in the case study, compared to TM and TMIP detectors. Then, the results from YOLOv4 CNN model was used to calculate the volume and productivity of the scaffolding projects. The percent predication errors ranged from -32% to -7.59% in the case study. Hence, this method showed potential in reliable scaffolding component detection to measure scaffolding component quantity and scaffolding productivity.

This study has combined image processing methods with a deep CNN model in scaffolding component detection and extraction. Compared with current practices of construction component detection such as windows and columns, this study explored an efficient and feasible way to extract more complicated features from images. Also, this study explored and compared the performance of conventional image processing tools and a deep CNN model. In the scaffolding images, conventional image processing tools outperformed when the scaffold appeared in a repetitive and homogeneous feature. While the CNN model was more capable of detect scaffold with distortion. The main contributions of this study include the following: first, automated scaffolding intersection detection in 2D images; second, the connections between the scaffolding intersection and scaffolding structure volume; and third, automated scaffolding intersection quantification and scaffolding volume and productivity measurement. By changing the training dataset and object definition, the proposed method can be applied in other construction structures which have a regular and repetitive appearance.

6 Conclusion

Scaffolding work consumes a large percentage of man-hours and cost in a construction project. A failure in delivery of scaffolding work could cause the serious delay of a whole construction project and consequent cost increase. Despite the high importance of scaffolding work, it suffers from low productivity and high cost in Australia. This research investigated vision-based methods for monitoring and measuring scaffolding productivity to provide onsite managers with timely and effective feedback on scaffolding work.

6.1 Summary and contribution

With the wide application of digital cameras in the construction industry, many scenarios on a construction site are captured by video surveillance or cameras for safety and production reasons. This research employed digital cameras to capture onsite vision data of scaffolding work and to extract semantic information from vision data. Two approaches were proposed in this research, and they utilized video frames and static images as data input, respectively.

The first approach proposed used video data which recorded scaffolder's activities over a period. By implementing the structure of OpenPose, 14 key points of the human skeleton were extracted from the video frames, and the key points were lifted into 3D coordinates. The 3D coordinates of these key points were transmitted to machine learning classifiers for classifier training. The scaffolding activities were defined as direct work (erecting), essential contributory work (transporting), and ineffective work (waiting/idling). A series of classifiers, including SVM, RF, DT, KNN, and ANN were trained to identify scaffolding activities, and the workface assessment was proposed to be generated from the results of classifiers. A case study was conducted, and all the selected classifiers achieved 90% average accuracy and 0.88 average F1 score in the scaffolding activity classification. Workface assessment was derived from the results of RF, which outperformed the other classifiers in the case study. Finally, the productivity indicator was generated based on the workface assessment.

Rather than taking dynamic video data as input in the first approach, the second proposed method processes static scaffolding photos as input. Scaffolding

intersections (couplers and wedges) were set as the object for detection. By detecting and calculating the number of scaffolding intersections in a scaffolding photo, the number of scaffolding surface units could be calculated, and then the volume of scaffolding structure in the image could be inferred by taking the project design parameters into calculation. Consequently, the scaffolding productivity was proposed to be derived from the total volume of scaffolding structure divided by the time consumption. This approach implements a YOLOv4 CNN model and template matching as the object detectors. A case study was conducted to validate and evaluate the detection process. A YOLOv4 CNN model, TM, and TMIP three object detectors were tested in the case study. Six individual project images were used as the input, and the performance indicators precision, recall, and F1 score were employed to evaluate the detectors. A YOLOv4 CNN model presents good stability and performance in these indicators in the case study, compared to TM and TMIP detectors. Next, the results from YOLOv4 CNN model was used to calculate the volume and productivity of the scaffolding projects. The percent predication errors range from -32% to -7.59% in the case study. Hence, this method shows potential as reliable scaffolding component detection to measure scaffolding component quantity and scaffolding productivity.

Finally, both approaches showed great potential and feasibility in automated monitoring and measuring scaffolding productivity in real time, and the research is expected to assist onsite project management and as a result release managers from regular manual inspections.

This study took onsite vision data and employed computer vision, deep learning algorithms for semantic information extraction and interpretation. It contributed to the onsite PMC in the following aspects.

1) 3D information extraction and interpretation of scaffolding activity in video

The proposed model obtained semantic information from scaffolding activities recorded in videos, by extracting key joints of a scaffolder and projecting these key joints into 3D. By applying machine learning classifiers, this model classified scaffolding activities into predefine categories. As a result, it reduced the time spent on manual interpretation.

2) Automatic scaffolding component detection

The proposed approach automatically detected and quantified the scaffolding couplers and wedges from 2D images. It released onsite managers from periodic inspection and assists in PMC.

3) Theoretical connections between the recognition results and scaffolding productivity

Theoretical connections were established between the recognition outcomes and scaffolding productivity. The developed approach automatically detected scaffolding intersections from images and recognise scaffolding activities and it could produce a quantitative result of scaffolding productivity based on the detection and recognition results. By following the principles of activity analysis and workface assessment, scaffolding productivity is linked to scaffolding activities. By conducting graph analysis, the scaffolding productivity is connected to the number of scaffolding intersections.

This study contributed to the top of construction productivity measurement that it presented a practical approach to employ RGB cameras to automate the process of scaffolding productivity measurement. Rather than employing various sensors and RGB-D cameras, this study explored the potentials of using RGB cameras of capturing and measuring scaffolding activities. This study developed a practical framework and a feasible approach to automatically measure scaffolding productivity through both dynamic videos and static images. Furthermore, deep learning and computer vision techniques were utilised in this study to automate the proposed process and improve recognition performance. It filled the research gap in automated scaffolding productivity measurement and provided practical guidance to other construction activities.

6.2 Limitations and future work

First, future work will be placed in the data collection. Our model achieved 0.7 overall F1 score in the detection of scaffolding intersection and the detection scores are relatively low, but there are still potentials to elevate its performance. Since the technique of deep learning was utilized in this study, the variety and the volume of

training dataset directly impacted the performance and robustness of our model. To enhance the model performance in the future, the variety and the volume of the training data need to be extended.

In terms of the recognition of scaffolding intersection, the object shapes are relatively simple, more efforts will be placed in the generalisation of this method towards other construction components.

Our model employed one single camera to capture 2D videos and images. Occlusion is an inevitable concern for the object recognition and semantic information extraction. Multi-angle capturing with several cameras might solve the problem of occlusion and it will be another direction of future work.

This study only focused on the scaffolding projects. However, the developed models can be applied to other types of construction project. Regarding the activity analysis and workface assessment, other construction activities such as bricklaying, which is another repetitive activity performed by workers can be a research direction. In terms of construction component recognition, construction materials or components with a regular and repetitive feature, for example concrete structure, would be investigated in the future.

Reference

AbouRizk, S, P Knowles, and UR Hermann. 2001. "Estimating labor production rates for industrial construction activities." *Journal of construction engineering and management* 127 (6):502-511.

Aggarwal, Nitin, and W Clem Karl. 2006. "Line detection in images through regularized Hough transform." *IEEE transactions on image processing* 15 (3):582-591.

Ali, Saad, and Mubarak Shah. 2008. "Human action recognition in videos using kinematic features and multiple instance learning." *IEEE transactions on pattern analysis and machine intelligence* 32 (2):288-303.

Allmon, Eric, Carl T Haas, John D Borcherding, and Paul M Goodrum. 2000. "US construction labor productivity trends, 1970–1998." *Journal of construction engineering and management* 126 (2):97-104.

Almukhtar, Avar, Zaid O Saeed, Henry Abanda, and Joseph HM Tah. 2021. "Reality capture of buildings using 3D laser scanners." *CivilEng* 2 (1):214-235.

Aloui, Saifeddine, Christophe Villien, and Suzanne Lesecq. 2015. "A new approach for motion capture using magnetic field: models, algorithms and first results." *International Journal of Adaptive Control and Signal Processing* 29 (4):407-426.

Alwasel, Abdullatif, Ali Sabet, Mohammad Nahangi, Carl T Haas, and Eihab Abdel-Rahman. 2017. "Identifying poses of safe and productive masons using machine learning." *Automation in Construction* 84:345-355.

Amendola, Sara, Luigi Bianchi, and Gaetano Marrocco. 2015. "Movement Detection of Human Body Segments: Passive radio-frequency identification and machine-learning technologies." *IEEE Antennas and Propagation Magazine* 57 (3):23-37.

Assadzadeh, Amin, Mehrdad Arashpour, Alireza Bab‐Hadiashar, Tuan Ngo, and Heng Li. 2021. "Automatic far‐field camera calibration for construction scene analysis." *Computer‐Aided Civil and Infrastructure Engineering*.

Bangaru, Srikanth Sagar, Chao Wang, and Fereydoun Aghazadeh. 2020. "Data Quality and Reliability Assessment of Wearable EMG and IMU Sensor for Construction Activity Recognition." *Sensors* 20 (18):5264.

Bangaru, Srikanth Sagar, Chao Wang, Sri Aditya Busam, and Fereydoun Aghazadeh. 2021. "ANN-based automated scaffold builder activity recognition through wearable EMG and IMU sensors." *Automation in Construction* 126:103653.

Belgiu, Mariana, and Lucian Drăguţ. 2016. "Random forest in remote sensing: A review of applications and future directions." *ISPRS Journal of Photogrammetry and Remote Sensing* 114:24-31.

Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934*.

Bosché, Frédéric, Adrien Guillemet, Yelda Turkan, Carl T Haas, and Ralph Haas. 2014. "Tracking the built status of MEP works: Assessing the value of a Scan-vs-BIM system." *Journal of Computing in Civil Engineering* 28 (4):05014004.

Brilakis, Ioannis K, and Lucio Soibelman. 2008. "Shape-based retrieval of construction site photographs." *Journal of Computing in Civil Engineering* 22 (1):14-20.

Brilakis, Ioannis, Lucio Soibelman, and Yoshihisa Shinagawa. 2005. "Material-based construction site image retrieval." *Journal of computing in civil engineering* 19 (4):341-355.

Cao, Zhe, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A Sheikh. 2019. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chae, Myung Jin, Gyu Won Lee, Jung Yoel Kim, Jae Woo Park, and Moon Young Cho. 2011. "A 3D surface modeling system for intelligent excavation system." *Automation in construction* 20 (7):808-817.

Chen, Yanyu, Wenzhe Zheng, Wenbo Li, and Yimiao Huang. 2021. "Large group Activity security risk assessment and risk early warning based on random forest algorithm." *Pattern Recognition Letters* 144:1-5.

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

Cheng, Tao, Jochen Teizer, Giovanni C Migliaccio, and Umberto C Gatti. 2013. "Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data." *Automation in Construction* 29:24-39.

Chi, Hung-Lin, Jian Chai, Changzhi Wu, Junxiang Zhu, Xiangyu Wang, and Chongyi Liu. 2017. "Scaffolding progress monitoring of LNG plant maintenance project using BIM and image processing technologies." 2017 International Conference on Research and Innovation in Information Systems (ICRIIS).

Chong, Heap-Yih, Cen-Ying Lee, and Xiangyu Wang. 2017. "A mixed review of the adoption of Building Information Modelling (BIM) for sustainability." *Journal of Cleaner Production* 142:4114-4126.

Christian, John, and Daniel Hachey. 1995. "Effects of delay times on production rates in construction." *Journal of Construction Engineering and Management* 121 (1):20-26.

CII. 2010. "Guide to activity analysis." *IR252–2a, Construction Industry Institute*.

Dang, L Minh, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. 2020. "Sensor-based and vision-based human activity recognition: A comprehensive survey." *Pattern Recognition* 108:107561.

Dictionary, Oxford. 2000. Oxford Advanced Learner's Dictionary. Oxford: Oxford university press.

Eastman, Chuck, and Rafael Sacks. 2005. "Relative Productivity in the AEC Industries in the US for Onsite and Off-site Activities." *Journal of construction Engineering & Management*.

Ekanayake, Biyanka, Johnny Kwok-Wai Wong, Alireza Ahmadian Fard Fini, and Peter Smith. 2021. "Computer vision-based interior construction progress monitoring: A literature review and future research directions." *Automation in Construction* 127:103705.

Ezeldin, A Samer, and Lokman M Sharara. 2006. "Neural networks for estimating the productivity of concreting activities." *Journal of construction engineering and management* 132 (6):650-656.

Fang, Yihai, Yong K Cho, Sijie Zhang, and Esau Perez. 2016. "Case study of BIM and cloud–enabled real-time RFID indoor localization for construction management applications." *Journal of Construction Engineering and Management* 142 (7):05016003.

Fayek, Aminah Robinson, and Ayodele Oduba. 2005. "Predicting industrial construction labor productivity using fuzzy expert systems." *Journal of construction engineering and management* 131 (8):938-941.

Fığlalı, Nilgün, Ahmet Cihan, Hatice Esen, Alpaslan Fığlalı, Davut Çeşmeci, Mehmet Kemal Güllü, and Mustafa Kerim Yılmaz. 2015. "Image processing-aided working posture analysis: I-OWAS." *Computers & Industrial Engineering* 85:384-394.

Flusser, Jan, Sajad Farokhi, Cyril Höschl, Tomáš Suk, Barbara Zitova, and Matteo Pedone. 2015. "Recognition of images degraded by Gaussian blur." *IEEE transactions on Image Processing* 25 (2):790-806.

Golparvar-Fard, Mani, Jeffrey Bohn, Jochen Teizer, Silvio Savarese, and Feniosky Peña-Mora. 2011. "Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques." *Automation in construction* 20 (8):1143-1155.

Gong, Jie, and Carlos H Caldas. 2010. "Computer vision-based video interpretation model for automated productivity analysis of construction operations." *Journal of Computing in Civil Engineering* 24 (3):252-263.

Gong, Zhiqiang, Ping Zhong, and Weidong Hu. 2019. "Diversity in machine learning." *IEEE Access* 7:64323-64350.

Gouett, Michael C, Carl T Haas, Paul M Goodrum, and Carlos H Caldas. 2011. "Activity analysis for direct-work rate improvement in construction." *Journal of Construction Engineering and Management* 137 (12):1117-1124.

Guraliuc, Anda R, Paolo Barsocchi, Francesco Potortì, and Paolo Nepa. 2011. "Limb movements classification using wearable wireless transceivers." *IEEE Transactions on Information Technology in Biomedicine* 15 (3):474-480.

Han, SangUk, SangHyun Lee, and Feniosky Peña-Mora. 2013. "Vision-based detection of unsafe actions of a construction worker: Case study of ladder climbing." *Journal of Computing in Civil Engineering* 27 (6):635-644.

Hanna, Awad S, Chul-Ki Chang, Kenneth T Sullivan, and Jeffery A Lackney. 2008. "Impact of shift work on labor productivity for labor intensive contractor." *Journal of construction engineering and management* 134 (3):197-204.

Herbsman, Zohar, and Ralph Ellis. 1990. "Research of factors influencing construction productivity." *Construction management and economics* 8 (1):49-61.

Hou, Lei, Changzhi Wu, Xiangyu Wang, and Jun Wang. 2014. "A framework design for optimizing scaffolding erection by applying mathematical models and virtual simulation." In *Computing in Civil and Building Engineering (2014)*, 323-330.

Hou, Lei, Chuanxin Zhao, Changzhi Wu, Sungkon Moon, and Xiangyu Wang. 2017. "Discrete firefly algorithm for scaffolding construction scheduling." *Journal of Computing in Civil Engineering* 31 (3):04016064.

Hsu, Chih-Wei, and Chih-Jen Lin. 2002. "A comparison of methods for multiclass support vector machines." *IEEE transactions on Neural Networks* 13 (2):415-425.

Hu, Xin, Heap-Yih Chong, and Xiangyu Wang. 2019. "Sustainability perceptions of off-site manufacturing stakeholders in Australia." *Journal of cleaner production* 227:346-354.

Hughes, Josie, Jize Yan, and Kenichi Soga. 2015. "Development of wireless sensor network using bluetooth low energy (BLE) for construction noise monitoring." *International Journal of Smart Sensing and Intelligent Systems* 8 (2):1379-1405.

Hui, Linda, Manwoo Park, and Ioannis Brilakis. 2014. "Automated in-placed brick counting for façade construction progress estimation." In *Computing in Civil and Building Engineering (2014)*, 958-965.

Ibrahim, YM, Tim C Lukins, X Zhang, Emanuele Trucco, and AP Kaka. 2009. "Towards automated progress assessment of workpackage components in construction projects using computer vision." *Advanced Engineering Informatics* 23 (1):93-103.

Ionescu, Catalin, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments." *IEEE transactions on pattern analysis and machine intelligence* 36 (7):1325-1339.

Jarkas, Abdulaziz M. 2010. "Critical investigation into the applicability of the learning curve theory to rebar fixing labor productivity." *Journal of Construction Engineering and Management* 136 (12):1279-1288.

Jiang, Hao, Mark S Drew, and Ze-Nian Li. 2006. "Successive convex matching for action detection." 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06).

Joshua, Liju, and Koshy Varghese. 2011. "Accelerometer-based activity recognition in construction." *Journal of computing in civil engineering* 25 (5):370-379.

Kang, Shih-Chung, Hung-Lin Chi, and Eduardo Miranda. 2009. "Three-dimensional simulation and visualization of crane assisted construction erection processes." *Journal of Computing in Civil Engineering* 23 (6):363-371.

Khosrowpour, Ardalan, Igor Fedorov, Aleksander Holynski, Juan Carlos Niebles, and Mani Golparvar-Fard. 2014. "Automated worker activity analysis in indoor environments for direct-work rate improvement from long sequences of RGB-D images." Construction Research Congress 2014: Construction in a Global Network.

Khosrowpour, Ardalan, Juan Carlos Niebles, and Mani Golparvar-Fard. 2014. "Vision-based workface assessment using depth images for activity analysis of interior construction operations." *Automation in Construction* 48:74-87.

Kim, Changyoon, Byoungil Kim, and Hyoungkwan Kim. 2013. "4D CAD model updating using image processing-based construction progress monitoring." *Automation in Construction* 35:44-52.

Kim, Kyungki, and Yong Cho. 2015. "BIM-based planning of temporary structures for construction safety." In *Computing in Civil Engineering 2015*, 436-444.

Kim, Kyungki, Yong Cho, and Sijie Zhang. 2016. "Integrating work sequences and temporary structures into safety planning: Automated scaffolding-related safety hazard identification and prevention in BIM." *Automation in Construction* 70:128-142.

Kim, Kyungki, and Jochen Teizer. 2014. "Automatic design and planning of scaffolding systems using building information modeling." *Advanced Engineering Informatics* 28 (1):66-80.

Kopsida, Marianna, Ioannis Brilakis, and Patricio Antonio Vela. 2015. "A review of automated construction progress monitoring and inspection methods." Proc. of the 32nd CIB W78 Conference 2015.

Lee, Jae Hwan, and Sang Oh Park. 2019. "Machine learning-based automatic reinforcing bar image analysis system in the internet of things." *Multimedia Tools and Applications* 78 (3):3171-3180.

Li, Huanhuan, Diyi Chen, Hao Zhang, Changzhi Wu, and Xiangyu Wang. 2017. "Hamiltonian analysis of a hydro-energy generation system in the transient of sudden load increasing." *Applied Energy* 185:244-253.

Li, Xiao, Wen Yi, Hung-Lin Chi, Xiangyu Wang, and Albert PC Chan. 2018. "A critical review of virtual and augmented reality (VR/AR) applications in construction safety." *Automation in Construction* 86:150-162.

Liu, Guohua, Liangyu Li, and Bingle Liu. 2015. "Study on recognition method of adhering bars based on support vector machine." *International Journal of Signal Processing, Image Processing and Pattern Recognition* 8 (9):363-370.

Liu, Jingen, Saad Ali, and Mubarak Shah. 2008. "Recognizing human actions using multiple features." 2008 IEEE Conference on Computer Vision and Pattern Recognition.

Liu, Kaijian, and Mani Golparvar-Fard. 2015. "Crowdsourcing construction activity analysis from jobsite video streams." *Journal of Construction Engineering and Management* 141 (11):04015035.

Lukins, Tim C, and Emanuele Trucco. 2007. "Towards Automated Visual Assessment of Progress in Construction Projects." BMVC.

Luo, Hanbin, Chaohua Xiong, Weili Fang, Peter ED Love, Bowen Zhang, and Xi Ouyang. 2018. "Convolutional neural networks: Computer vision-based workforce activity assessment in construction." *Automation in Construction* 94:282-289.

Marks, Eric Daniel, and Jochen Teizer. 2013. "Method for testing proximity detection and alert technology for safe construction equipment operation." *Construction Management and Economics* 31 (6):636-646.

Medina-Carnicer, R, Francisco José Madrid-Cuevas, A Carmona-Poyato, and Rafael Muñoz-Salinas. 2009. "On candidates selection for hysteresis thresholds in edge detection." *Pattern Recognition* 42 (7):1284-1296.

Memarzadeh, Milad, Mani Golparvar-Fard, and Juan Carlos Niebles. 2013. "Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors." *Automation in Construction* 32:24-37.

Moon, Sungkon, John Forlani, Xiangyu Wang, and Vivian Tam. 2016. "Productivity study of the scaffolding operations in liquefied natural gas plant construction: Ichthys project in Darwin, Northern Territory, Australia." *Journal of Professional Issues in Engineering Education and Practice* 142 (4):04016008.

Narazaki, Yasutaka, Vedhus Hoskere, Tu A Hoang, Yozo Fujino, Akito Sakurai, and Billie F Spencer Jr. 2020. "Vision - based automated bridge component recognition with high - level scene consistency." *Computer - Aided Civil and Infrastructure Engineering* 35 (5):465-482.

Nie, Zuoxian, Mao-Hsiung Hung, and Jing Huang. 2016. "A Novel Algorithm of Rebar Counting on Conveyor Belt Based on Machine Vision." *J. Inf. Hiding Multim. Signal Process.* 7 (2):425-437.

Omar, Hany, Lamine Mahdjoubi, and Gamal Kheder. 2018. "Towards an automated photogrammetry-based approach for monitoring and controlling construction site activities." *Computers in Industry* 98:172-182.

Park, Hee-Sung, Stephen R Thomas, and Richard L Tucker. 2005. "Benchmarking of construction productivity." *Journal of construction engineering and management* 131 (7):772-778.

Park, JeeWoong, Eric Marks, Yong K Cho, and Willy Suryanto. 2016. "Performance test of wireless technologies for personnel and equipment proximity sensing in work zones." *Journal of Construction Engineering and Management* 142 (1):04015049.

Peddi, Abhinav, Luke Huan, Yong Bai, and Seonghoon Kim. 2009. "Development of human pose analyzing algorithms for the determination of construction productivity in real-time." Construction Research Congress 2009: Building a Sustainable Future.

Pitelis, Nikolaos, Chris Russell, and Lourdes Agapito. 2013. "Learning a manifold as an atlas." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Qu, Xiaobo, Yang Yu, Mofan Zhou, Chin-Teng Lin, and Xiangyu Wang. 2020. "Jointly dampening traffic oscillations and improving energy consumption with electric, connected and automated vehicles: A reinforcement learning based approach." *Applied Energy* 257:114030.

Rahman, CM. 2014. "Study and analysis of work postures of workers working in a ceramic industry through rapid upper limb assessment (RULA)." *International Journal of Engineering* 5 (3):8269.

Ratay, Robert T. 2012. *Temporary structures in construction*: McGraw-Hill Education.

Ray, Soumitry J, and Jochen Teizer. 2012. "Real-time construction worker posture analysis for ergonomics training." *Advanced Engineering Informatics* 26 (2):439-455.

Rebolj, Danijel, Nenad Čuš Babič, Aleš Magdič, Peter Podbreznik, and Mirko Pšunder. 2008. "Automated construction activity monitoring system." *Advanced engineering informatics* 22 (4):493-503.

Roh, Seungjun, Feniosky Peña-Mora, Mani Golparvar Fard, and SangUk Han. 2009. "Visualization application for interior progress monitoring in 3D environment." Construction Research Congress 2009: Building a Sustainable Future.

Ryu, JuHyeong, JoonOh Seo, Houtan Jebelli, and SangHyun Lee. 2019. "Automated action recognition using an accelerometer-embedded wristband-type activity tracker." *Journal of construction engineering and management* 145 (1):04018114.

Sarafianos, Nikolaos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 2016. "3d human pose estimation: A review of the literature and analysis of covariates." *Computer Vision and Image Understanding* 152:1-20.

Seo, JoonOh, Richmond Starbuck, SangUk Han, SangHyun Lee, and Thomas J Armstrong. 2015. "Motion data-driven biomechanical analysis during construction tasks on sites." *Journal of Computing in Civil Engineering* 29 (4):B4014005.

Shaikhina, Torgyn, Dave Lowe, Sunil Daga, David Briggs, Robert Higgins, and Natasha Khovanova. 2019. "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation." *Biomedical Signal Processing and Control* 52:456-462.

Sheikh, Yaser, Mumtaz Sheikh, and Mubarak Shah. 2005. "Exploring the space of a human action." Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1.

Shou, Wenchi, Jun Wang, Peng Wu, Xiangyu Wang, and Heap-Yih Chong. 2017. "A cross-sector review on the use of value stream mapping." *International Journal of Production Research* 55 (13):3906-3928.

Simonyan, Karen, and Andrew Zisserman. 2014. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*.

Singh, Vishal, Ning Gu, and Xiangyu Wang. 2011. "A theoretical framework of a BIM-based multi-disciplinary collaboration platform." *Automation in construction* 20 (2):134-144.

Son, Hyojoo, Changmin Kim, and Changwan Kim. 2015. "Fully automated as-built 3D pipeline extraction method from laser-scanned data based on curvature computation." *Journal of Computing in Civil Engineering* 29 (4):B4014003.

Son, Hyojoo, and Changwan Kim. 2010. "3D structural component recognition and modeling method using color and 3D data for construction progress monitoring." *Automation in Construction* 19 (7):844-854.

Song, Yongze, Xiangyu Wang, Graeme Wright, Dominique Thatcher, Peng Wu, and Pascal Felix. 2018. "Traffic volume prediction with segment-based regression kriging and its implementation in assessing the impact of heavy vehicles." *Ieee transactions on intelligent transportation systems* 20 (1):232-243.

Sonmez, Rifat, and James E Rowings. 1998. "Construction labor productivity modeling with neural networks." *Journal of construction engineering and management* 124 (6):498-504.

Srinavin, Korb, and Sherif Mohamed. 2003. "Thermal environment and construction workers' productivity: some evidence from Thailand." *Building and Environment* 38 (2):339-345.

Stoeter, Sascha A, Frederic Le Mauff, and Nikolaos P Papanikolopoulos. 2000. "Real-time door detection in cluttered environments." Proceedings of the 2000 IEEE International Symposium on Intelligent Control. Held jointly with the 8th IEEE Mediterranean Conference on Control and Automation (Cat. No. 00CH37147).

Tam, Vivian WY, and Ivan WH Fung. 2011. "Tower crane safety in the construction industry: A Hong Kong study." *Safety science* 49 (2):208-215.

Thomas, H Randolph, and Jeffrey Daily. 1983. "Crew performance measurement via activity sampling." *Journal of construction engineering and management* 109 (3):309-320.

Thomas, H Randolph, and Ahmet S Sakarcan. 1994. "Forecasting labor productivity using factor model." *Journal of Construction Engineering and Management* 120 (1):228-239.

Thomas, H Randolph, and Iacovos Yiakoumis. 1987. "Factor model of construction productivity." *Journal of construction engineering and management* 113 (4):623-639.

Tome, Denis, Chris Russell, and Lourdes Agapito. 2017. "Lifting from the deep: Convolutional 3d pose estimation from a single image." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Torres Calderon, Wilfredo, Dominic Roberts, and Mani Golparvar-Fard. 2021. "Synthesizing Pose Sequences from 3D Assets for Vision-Based Activity Analysis." *Journal of Computing in Civil Engineering* 35 (1):04020052.

Trucco, Emanuele, and Ammar P Kaka. 2004. "A framework for automatic progress assessment on construction sites using computer vision." *International Journal of IT in Architecture Engineering and Construction* 2:147-164.

Turkan, Yelda, Frederic Bosche, Carl T Haas, and Ralph Haas. 2012. "Automated progress tracking using 4D schedule and 3D sensing technologies." *Automation in construction* 22:414-421.

Van Tam, Nguyen, Nguyen Quoc Toan, Dinh Tuan Hai, and Nguyen Le Dinh Quy. 2021. "Critical factors affecting construction labor productivity: A comparison between perceptions of project managers and contractors." *Cogent Business & Management* 8 (1):1863303.

Wang, Qian, and Min-Koo Kim. 2019. "Applications of 3D point cloud data in the construction industry: A fifteen-year review from 2004 to 2018." *Advanced Engineering Informatics* 39:306-319.

Wang, Zhaojing, Hao Hu, and Wei Zhou. 2017. "RFID Enabled Knowledge - Based Precast Construction Supply Chain." *Computer - Aided Civil and Infrastructure Engineering* 32 (6):499-514.

Wu, Peng, Jun Wang, and Xiangyu Wang. 2016. "A critical review of the use of 3-D printing in the construction industry." *Automation in Construction* 68:21-31.

Wu, Y, and H Kim. 2004. "Digital imaging in assessment of construction project progress." *Proc. Of the 21th ISARC*:537-542.

Xu, Yusheng, Sebastian Tuttas, Ludwig Hoegner, and Uwe Stilla. 2018. "Reconstruction of scaffolds from a photogrammetric point cloud of construction sites using a novel 3D local feature descriptor." *Automation in Construction* 85:76-95.

Yan, Xuzhong, Heng Li, Angus R Li, and Hong Zhang. 2017. "Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention." *Automation in Construction* 74:2-11.

Yan, Xuzhong, Heng Li, Chen Wang, JoonOh Seo, Hong Zhang, and Hongwei Wang. 2017. "Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion." *Advanced Engineering Informatics* 34:152-163.

Yang, Jun, Patricio Vela, Jochen Teizer, and Zhongke Shi. 2014. "Vision-based tower crane tracking for understanding construction activity." *Journal of Computing in Civil Engineering* 28 (1):103-112.

Yang, Xiaodong, and Yingli Tian. 2010. "Robust door detection in unfamiliar environments by combining edge and corner features." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops.

Yi, Wen, and Albert PC Chan. 2014. "Critical review of labor productivity research in construction journals." *Journal of management in engineering* 30 (2):214-225.

Yu, Yantao, Hongling Guo, Qinghua Ding, Heng Li, and Martin Skitmore. 2017. "An experimental study of real-time identification of construction workers' unsafe behaviors." *Automation in Construction* 82:193-206.

Zhang, Hong, Xuzhong Yan, and Heng Li. 2018. "Ergonomic posture recognition using 3D view-invariant features from single ordinary camera." *Automation in Construction* 94:1-10.

Zhao, Huili, Guofeng Qin, and Xingjian Wang. 2010. "Improvement of canny algorithm based on pavement edge detection." 2010 3rd International Congress on Image and Signal Processing.

Zhou, Yimin, Ling Tian, Ce Zhu, Xin Jin, and Yu Sun. 2019. "Video coding optimization for virtual reality 360-degree source." *IEEE Journal of Selected Topics in Signal Processing* 14 (1):118-129.

Zhu, Junxiang, Xiangyu Wang, Mengcheng Chen, Peng Wu, and Mi Jeong Kim. 2019. "Integration of BIM and GIS: IFC geometry transformation to shapefile using enhanced open-source approach." *Automation in construction* 106:102859.

Zhu, Junxiang, Peng Wu, Mengcheng Chen, Mi Jeong Kim, Xiangyu Wang, and Tingchen Fang. 2020. "Automatically processing IFC clipping representation for BIM and GIS integration at the process level." *Applied sciences* 10 (6):2009.

Zhu, Zhenhua, and Ioannis Brilakis. 2010a. "Concrete column recognition in images and videos." *Journal of computing in civil engineering* 24 (6):478-487.

Zhu, Zhenhua, and Ioannis Brilakis. 2010b. "Parameter optimization for automated concrete detection in image data." *Automation in Construction* 19 (7):944-953.