Vision-Based Construction Worker Task Productivity Monitoring

Eirini Konstantinou

Wolfson College

This dissertation is submitted

for the degree of

DOCTOR OF PHILOSOPHY

October 2017

Department of Engineering

UNIVERSITY OF CAMBRIDGE

This thesis is dedicated to my parents, Christos and Katerina

&

my family, Nichole, Vasilis, Helen

&

my Pepe

DECLARATION

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

In accordance with the Department of Engineering guidelines, this thesis does not exceed 65,000 words, and it contains less than 150 figures.

Signed: _____

Date: _____

Eirini Konstantinou
October 2017

SUMMARY

Over the past decades, the construction industry lags further and further behind the manufacturing sector when productivity is considered. This is due to internal factors that take place on-site. Almost all of them are directly related to the way that productivity is monitored. Current practices for monitoring labour productivity are labour intensive, time - cost consuming and error prone. They are mainly reactive processes initiated after the detection of a negatively influencing factor. Although research studies have been performed towards leveraging these limitations, a gap still exists in monitoring labour productivity of multiple workers at the same time accurately, unobtrusively, cost and time efficiently. This thesis proposes a framework to address this gap. It hypothesizes that task productivity of construction workers can be monitored through their trajectory data. The proposed framework uses as input, video data streamed from cameras with overlapping field of view. It consists of two main methods. The output of the first is the input of the second. The first method tracks the location of workers across the range of a jobsite over time and returns their 4D trajectories. Such type of tracking requires that workers are matched under a unique ID not only across successive frames of a single camera (intra tracking) but also across multiple cameras (inter tracking). Existing tag-less studies fail to track construction workers due to the challenging nature of their working environments. Therefore, two novel computer vision-based algorithms are developed to perform both the intra and the inter camera tracking. The second method of the proposed framework converts the 4D trajectories of workers into productivity information. These trajectories are clustered into work cycles with an accuracy of 95%, recall of 76% and precision of 76%. Such work cycles depict the actual execution of tasks. The overall proposed framework features an average accuracy of 95% in terms of determining the total time workers spend on construction-related tasks.

ACKNOWLEDGMENTS

Firstly, I would like to thank my academic supervisor Dr Ioannis Brilakis, for seeing my potential and giving me the opportunity of doing a PhD in a world-renowned university. I am grateful to all my fellow students, in the Construction Information Technology Group and all the staff of the Laing O' Rourke Centre. Special thanks to Jan for all her kind support. I am extremely grateful to my industrial supervisor, Adam Locke for his excellent and amazing guidance. His help was essential throughout my studies, from data collection to the understanding of the research gap. Without his help this PhD would not have been possible. I also wish to express my sincere gratitude to my academic advisor Dr Joan Lasenby. Her help was critical all these years. Thank you to my good friends, Elia, Jason, Maria, and Ioannis. I am so grateful for all the nice moments we shared during my years in Cambridge.

My deepest gratitude goes to my parents, Christos and Katerina, my sister Nichole, my aunt Helen, my cousin Vasilis, my Pepe and my best friend Zeta. This journey was very long and challenging. I can't even think myself without their continual and unconditional support.

Finally, I would also like to acknowledge and express gratitude for financial support from the Engineering and Physical Sciences Research Council (EPSRC) and the Laing O' Rourke enterprise.

Contents

1. Introduction 20

1.1. Definition of productivity in construction 20

1.2. The problem with labour productivity in construction 21

1.3. Factors affecting labour productivity 22

1.4. Current state of practice in monitoring of labour productivity 25

1.5. Conclusions and thesis overview 27

2. Current state of research in monitoring of labour productivity 28

2.1. Region-based studies 29

2.1.1 Tagged studies 29

2.1.2 Tag-less studies 31

2.2. Activity-based studies 32

2.3. Summary of current state of research 34

2.4. Hypothesis & Proposed framework 35

2.4.1. Experimental set up 38

3. Adaptive computer vision-based 2D tracking of workers in complex environments 40

3.1. Introduction 40

3.2. Background 41

3.3. Proposed solution 43

3.4. Proposed methodology 45

3.4.1. Prediction model 45

3.4.2. Appearance model 47

3.4.3. Filtering model 51

3.4.4. Adaptive Model 57

3.5. Experiments and results 58

3.5.1. Definition of parameters 60

2.5.1.1 Definition of scale parameters 60

2.5.1.2 Definition of segmentation parameters 62

3.5.2. Quantitative Evaluation 63

3.5.3. Qualitative evaluation 68

3.6. Chapter overview 70

4. Matching of construction workers across views for automated 4D vision tracking 72

4.1. Introduction 72

4.2. Previous related work 74

4.3. Proposed solution 76

4.4. Proposed methodology 78

4.4.1. Motion-based matching method 79
4.4.2. Geometry-based matching method 83
4.4.3. Template-based matching method 86
4.5. Calculation of 3D trajectories 86
4.6. Experiments and results 88
4.6.1. Evaluation of the motion-based matching method 90
4.6.2. Evaluation of the geometry-based matching method 92
4.6.3. Evaluation of the template-based matching method 96
4.6.4. Evaluation of overall proposed matching method 97
4.6.5. Accuracy of 3D trajectories 99
4.7. Chapter overview 100
5. Detection of work cycles for monitoring labour productivity 102
5.1. Introduction to trajectory analysis for pattern recognition 102
5.2. Cluster analysis outline 105
5.3. Proposed solution 111
5.4. Proposed methodology 113
5.4.1. Smoothing of 4D trajectories 113
5.4.2. Partitioning of 4D trajectories 116
5.4.3. Classification of 4D sub-trajectories 117
5.4.4. Clustering of 4D sub-trajectories into work cycles 118
5.5. Experiments and results 119
5.5.1. Definition of parameters 120
5.5.1.1 Definition of smoothing parameters 121
5.5.1.2 Definition of classification parameters 122
5.5.2. Evaluation of work cycles' detection 123
5.6. Chapter overview 129
6. Conclusions and future work 132
6.1. Conclusions 132
6.2. Contributions 135
6.3. Recommendations for future work 135
Bibliography 137

List of Tables

Table 1-1: Comparison “A” of factors affecting labour productivity in construction. 24
Table 1-2: Comparison “B” of factors affecting labour productivity in construction. 25
Table 3-1: Proposed features for overcoming most common tracking challenges 58
Table 3-2: Average width and walking speed values of workers per scale. 61
Table 3-3: Proposed method’s tracking performance under different segmentation 62
Table 3-4: Evaluation video samples. 63
Table 3-5: Average quantitative evaluation results *(best values are highlighted bold)* 68
Table 4-1: Expected efficiency of geometry, motion and template-based methods 77
Table 4-2: Evaluation of proposed motion matching method 91
Table 4-3: Experimentally defined $errorfi.$ 93
Table 4-4: Evaluation of proposed geometry-based matching method. 94
Table 4-5: Evaluation of proposed template based matching method. 97
Table 4-6: Evaluation of motion and geometry matching method 98
Table 4-7: Confusion matrix of the overall proposed method. 98
Table 5-1: Smoothing step k of the method proposed 121
Table 5-2: Manually collected ground truth of semantic “stops” of worker “1” (data set steel). 124
Table 5-3: Manually collected ground truth of semantic “stops” of worker “2” (data set electrical). 126
Table 5-4: Manually collected ground truth of semantic “stops” of worker “3” (data set electrical). 126
Table 5-5: Confusion matrix of proposed method for detecting work cycles. 128
Table 5-6: Quantitative summary of the labour input. 129

List of Figures

Figure 1-1: Relationship between performance and productivity in construction 21
Figure 1-2: Labour productivity growth rates of the construction sector and non–farm industries..... 22
Figure 1-3: Factors affecting labour productivity (Nasirzadeh & Nojedehi, 2013)..... 23
Figure 1-4: Factors affecting labour productivity in construction. 24
Figure 1-5: Crew Balance Chart of a pre-cast beam installation operation. (Al-qahtani et al. 2007)... 26
Figure 2-1: (a) GPS mounted on earthmoving equipment (RTD FasTracks, 2014) 29
Figure 2-2**:** Monitoring of labour productivity given worker presence 30
Figure 2-3: Operation process model of a concrete pouring task (Gong & Caldas, 2010, 2011) 31
Figure 2-4: Productivity monitoring of earthmoving equipment (Golparvar-Fard et al., 2013)..... 32
Figure 2-5: Methodology of physiological-based studies for recognizing activities 33
Figure 2-6: Overall framework for automated construction worker task productivity monitoring. 36
Figure 2-7: Assumptions of proposed framework 36
Figure 2-8: Camera centred coordinate system..... 38
Figure 3-1: Flowchart of proposed computer vision-based 2D tracking method 45
Figure 3-2**:** Prediction of the position of a tracked target in the following frame. 47
Figure 3-3: HSV vs RGB channels for (a) orange & (b) yellow Hi-Vis apparel..... 47
Figure 3-4: Proposed appearance model..... 48
Figure 3-5: Training data of a multi-colour SVM classifier. 48
Figure 3-6: Linear classification of colour image patches with an SVM..... 49
Figure 3-7: Template matching of the same worker between non-successive frames..... 50
Figure 3-8: Proposed filtering method..... 51
Figure 3-9: Failure of the proposed prediction model under the appearance of outliers. 51
Figure 3-10: Motion-based classification of pixels..... 52
Figure 3-11: Example of frame differencing between successive frames 53
Figure 3-12: Motion contour for filtering outliers. 54
Figure 3-13: Motion contour while target walks (a-b), partially moves (c), and bends (d)..... 54
Figure 3-14: Classification of tracked target as moving or stationary 56
Figure 3-15: Activation of proposed filtering model in (a) and deactivation in (d)..... 57
Figure 3-16: Workers' appearance variations due to posture (a-c), scale (d) and occlusion (e)..... 57
Figure 3-17: Distance error performance metric..... 59
Figure 3-18: F-measure performance metric 60
Figure 3-19: Cumulative probability of tracked targets' walking speed per scale..... 60
Figure 3-20: Screenshots of the evaluation of the proposed prediction model..... 61
Figure 3-21: Examples of tracking performance 62
Figure 3-22: Comparison results of this chapter's proposed visual tracking method..... 64

Figure 3-23: Screenshots of the performance of the proposed tracking method 65
Figure 3-24: Comparison results of this chapter's visual tracking method..... 66
Figure 3-25: Screenshots of the performance of the proposed tracking method 67
Figure 3-26: Screenshots of the performance of the proposed method..... 69
Figure 3-27: Screenshots of the performance of the proposed method..... 69
Figure 4-1: Worker's motion through the surveillance cameras of a jobsite. 73
Figure 4-2: Typical workforce in a construction site. 73
Figure 4-3: Examples of occluded (a-d) and similarly dressed (e) construction workers..... 75
Figure 4-4: Flowchart of computer vision-based method for matching construction workers 77
Figure 4-5: Image and world coordinate systems of a pinhole camera model..... 78
Figure 4-6: Worker's 2D motion data across time. 79
Figure 4-7: Candidates correlation table (CCT) 79
Figure 4-8: Proposed searching algorithm for the "strongest" candidate. 81
Figure 4-9: Projection of tracked 2D motion paths..... 82
Figure 4-10: Reference coordinate system for motion matching between non-conjugated cameras.... 82
Figure 4-11: Posture variations between frames of non-conjugated cameras..... 83
Figure 4-12: False negative (FN) matching of a single worker between two cameras (a-b). 84
Figure 4-13: Drifting issue of a computer vision tracking method..... 84
Figure 4-14: Geometry-based matching search band..... 85
Figure 4-15: Multiple candidates within the proposed matching search band..... 86
Figure 4-16: Mid-point triangulation method 87
Figure 4-17: Experimental video data sets. (a) Data set A. (b) Data set B. 88
Figure 4-18: Detection of corresponding points for stereo camera calibration..... 89
Figure 4-19: Stereo calibration accuracy 89
Figure 4-20: Display of previous motion data (black line) over time..... 90
Figure 4-21: Total TP and FP matches over lengths of past motion data (per camera)..... 91
Figure 4-22: Performance examples of the motion-based matching method..... 92
Figure 4-23: Cumulative distribution of the fluctuation error ($errorf$)..... 93
Figure 4-24: TN & FN missed matches over the probability ($Perrorfi$) of $errorfi$ values. 94
Figure 4-25: Performance examples of the geometry-based matching method..... 95
Figure 4-26: Illustration of compared colour templates with the HSV colour space..... 96
Figure 4-27: Performance examples of the template based matching method 97
Figure 4-28: Performance of the overall vision-based matching method. 99
Figure 4-29: Ground truth trajectory..... 100
Figure 4-30: Euclidean calculation of triangulated trajectory..... 100
Figure 5-1: Trajectory clustering for detecting abnormal human behaviour (Wiliem et al., 2008).... 103
Figure 5-2: Hausdorff distance between trajectories TA, TB. 103

Figure 5-3: Hidden states (1-10) of an HMM within a store (Suzuki et al., 2007) 104
Figure 5-4: Cluster analysis of objects into three groups (clusters) 106
Figure 5-5: Locality In-between Polylines (LIP) distance of trajectories (Q, S) 107
Figure 5-6: Types of clusters (Tan et al., 2005) 108
Figure 5-7: Partitioning of trajectories based on preciseness and conciseness (J. Lee et al., 2008) ... 109
Figure 5-8: Partitioning of trajectories based on spatiotemporal changes. 110
Figure 5-9: Semantic stop regions of trajectory data (Palma et al. 2008) 110
Figure 5-10: Flowchart of proposed method for monitoring the labour productivity of workers 112
Figure 5-11: Screenshots of a tracked worker who remains still 114
Figure 5-12: Unsmoothed trajectory data of a non-moving worker 114
Figure 5-13: Fitting a line to a time series 115
Figure 5-14: Range of workers' movements while at "stop" event. 117
Figure 5-15: Clustering of 4D sub-trajectories into three work cycles ci (blue, yellow, green) 118
Figure 5-16: Tested data sets (from top to bottom: data set steel fixing, data set electrical) 119
Figure 5-17: Tracked areas 120
Figure 5-18: 3D speed values of the smoothed trajectory data of a non-moving worker 121
Figure 5-19: Smoothed trajectory data of the almost still worker of previous Figure 5-11. 122
Figure 5-20: Precision, recall, and accuracy graphs 123
Figure 5-21: Normalized speed values of worker "1" along the floor plane $vixzi = 1 \dots N.$ 123
Figure 5-22: Detected work cycles of worker "1" from data set steel fixing part A (a), and B (b) 125
Figure 5-23: Significant fluctuation of implemented computer vision-based 2D tracking method 126
Figure 5-24: Detected work cycles of worker "2" 127
Figure 5-25: Detected work cycles of worker "3" 128

List of Abbreviations and Acronyms

SVM: Support Vector Machine

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

NCC: Normalized Cross Correlation

1

Introduction

This chapter presents the labour productivity gap between construction and other industries, its social and economic impact and identifies its causes.

1.1. Definition of productivity in construction

In general, productivity is defined as the ratio of output to input (Lim, 1996). Productivity rates are used by project managers during planning and scheduling in order to reduce the labour cost and improve the performance of workers (Alinaitwe et al., 2006). Several models have been proposed in order to quantify these productivity rates in construction. Such models are the following (Thomas et al.,1990):

- The economic model that expresses both the input and the output in monetary units (e.g. dollars $):

$$Productivity = \frac{Total\ output\ in\ \$}{Total\ input\ in\ \$} = \frac{Total\ output\ in\ \$}{(Labour+Materials+Equipment+Energy+Capital)\ in\ \$} \quad (1\text{-}1)$$

- The project-specified model that estimates the productivity based on the size of the project (e.g. square meters m^2):

$$Productivity = \frac{Total\ output\ \ in\ m^2}{Partial\ input\ in\ \$} = \frac{Total\ output\ \ in\ m^2}{(Labour+Equipment+Materials)\ in\ \$} \quad (1\text{-}2)$$

- The activity-oriented model that focuses on the labour input i.e. paid work hours and the installed quantity as output (e.g. cubic meters of soil excavated, meters of brick wall constructed, number of concrete buckets transferred, steel cages assembled):

$$Labour\ Productivity = \frac{Partial\ output\ as\ installed\ quantity}{Partial\ input\ in\ work\ hours} \quad (1\text{-}3)$$

The economic and the project-oriented models are mainly used by governmental agencies and the private sector whilst the activity-oriented model is preferred by contractors (Shehata & El-Gohary, 2011). This is because the latter model provides a better insight of the performance of workers. Performance refers to excellence. It is not only related to productivity but also to profitability, and several other metrics such as speed, quality and flexibility (see Figure 1-1). Another advantage of the activity-oriented model is that it relies on the labour input that can be managed much easier compared to other resources such as materials and equipment (Park, 2006). In addition, it has been reported that labour productivity in construction is not easily predictable as it usually returns big variations (Halligan et al., 1994). Due to these reasons, this thesis will focus on the labour productivity as defined by the activity-oriented model (see Equation 1-3).

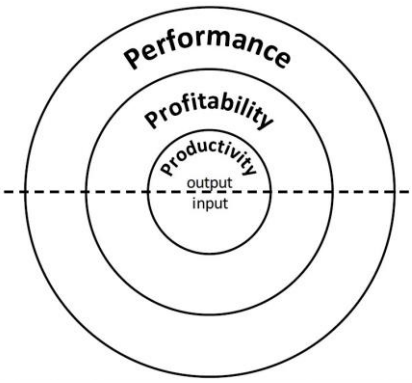

Figure 1-1: Relationship between performance and productivity in construction (Pekuri et al., 2011).

1.2. The problem with labour productivity in construction

The construction sector has gradually created a significant labour productivity gap compared to other industries over the past five decades. It is estimated that that only 50% of the total construction time is productive (Horman & Kenley, 2005; Picard, 2004). Data taken from the US Bureau of Labour Statistics show that labour productivity in construction is not improving over time, with an annual 0.59% declining rate in USA during the period 1964-2003 (Teicholz, 2004) and 0.32% during the 2003 to 2012 period (Teicholz, 2013). In the UK for the period between 1998 to 2008, labour productivity was 11% less than at USGC (US Gulf Coast) and 6% less than Western Europe (Merrow et al., 2009). On the other hand, non-farm industries (i.e. the part of the domestic economy that does not include activities related to private households, government, farm and no profit organizations) in the US doubled their labour productivity over the past 50 years (see Figure 1-2). The resulting gap leaves a lot of room for innovation to improve construction productivity. Therefore, the question that arises is: "*Why is the gap that big in the first place*?"

Before answering this question it is essential to define the impact that low productivity has on society. This can be estimated by considering the significance of the role that the construction sector

holds in national economies. According to the annual report of FIEC (European Construction Industry Federation, 2017), 8.6% of the gross domestic product (GDP) of the European Union (EU) belongs to the construction industry sector. This contributed 1278 billion Euros (€) in 2016 (EU28), and supplied almost 6.4% of Europe's total employment (42.9 million workers). Almost 8% (11 million workers) of the total US workforce was engaged in the construction sector in 2008 just before the onset of the recession (Bureau of Labor Statistics, 2008). If labour productivity in construction is improved, then the construction sector will automatically increase its revenue, considering that 33-50% of the entire cost of a typical project is spent on labour (Hanna et al., 2005).

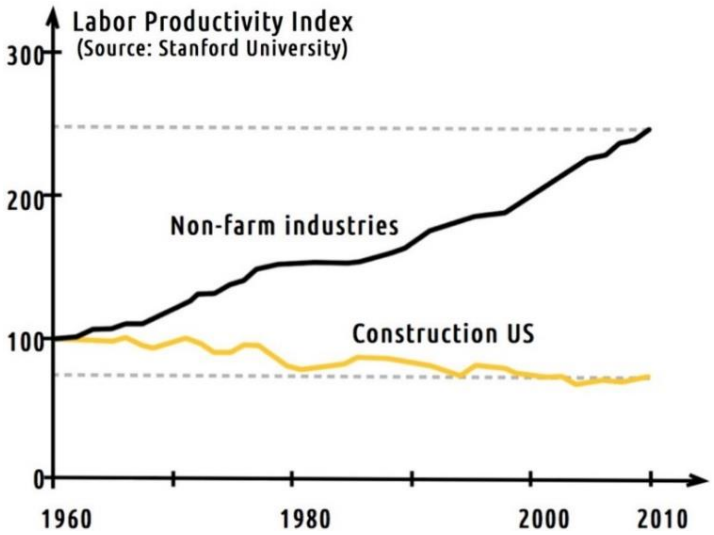

Figure 1-2: Labour productivity growth rates of the construction sector and non–farm industries over the last 5 decades (Frinault, 2015).

1.3. Factors affecting labour productivity

Previous studies have performed a number of surveys in order to identify the factors that affect labour productivity. Figure 1-3 illustrates the large range of these factors. A survey carried out in the US (Dai et al., 2009) evaluated the impact of 83 factors as derived from a previous study (Dai et al., 2005) and concluded that the most important factors are mismanagement of construction equipment (e.g. availability, quality, lack), materials, tools and consumables. This study questioned in total 1996 craft workers from 28 industries. Lack of materials, overcrowded areas and rework can reduce the productive labour input by a factor of 5.1 to 13.6 man-hours per week based on a survey carried out in Hong Kong (Ng et al., 2004). Another survey conducted in the US concluded that most of the factors are related to a lack of efficient management (Wambeke et al., 2011). These factors were: equipment and senior management coordination, crew/labour force/material management, prerequisites and constructability, tools and PPE (personal protective equipment), supervisor skills and communication.

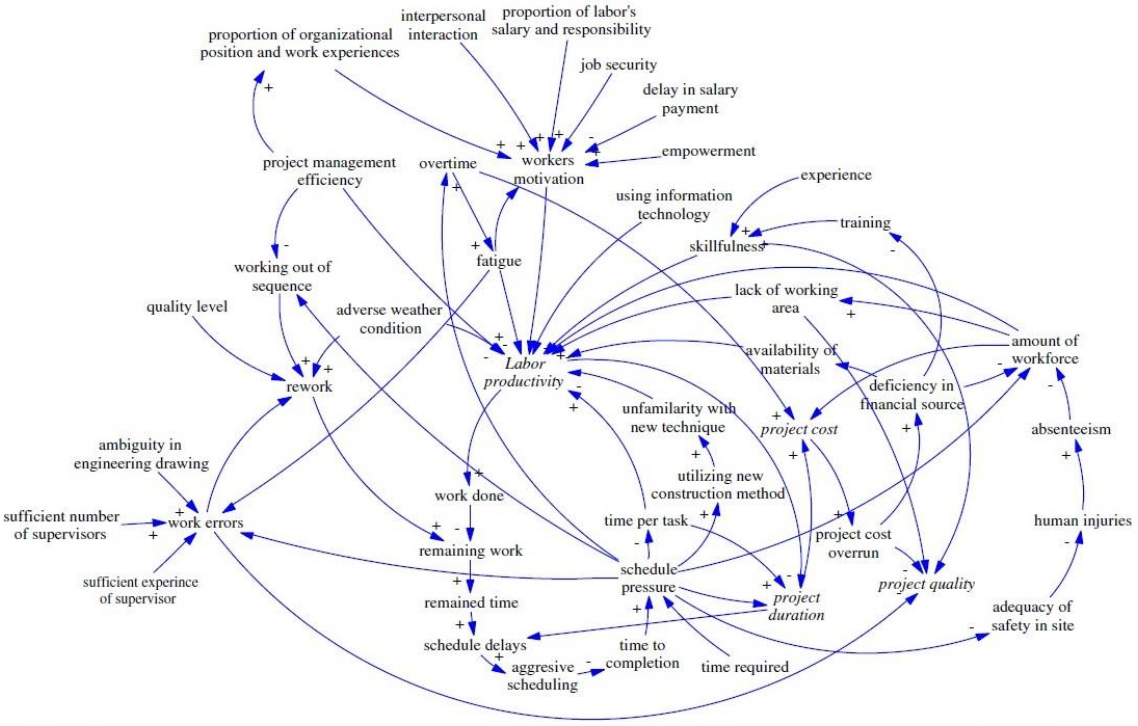

Figure 1-3: Factors affecting labour productivity (Nasirzadeh & Nojedehi, 2013).

The problem with the above studies is that they do not reach the same conclusions about the factors that affect labour productivity in construction. Inconclusivity arises due to ambiguities about which factors are the most important. This chapter performs an analytic comparison of previously identified factors (Cundecha, 2012; El-gohary & Aziz, 2014; Jarkas & Bitar, 2012; Kuykendall, 2007; Lim & Alum, 1995; Makulsawatudom et al., 2004) in order to resolve the ambiguity. To achieve this, all factors from each of these studies are grouped into four categories based on the categorization proposed by Jarkas & Bitar (2012). These categories are:

- The management category: supervision, overtime, turnover, safety, resources, scheduling, rest areas, transportation, payment, congestion, and disruptions.
- The technological category: designs, drawings, site layout, rework, and site access.
- The workforce category: skills, fatigue, motivation, absenteeism, late arrivals, age, unscheduled breaks, and personal issues.
- The external category: weather, law regulation, and owner (change of orders)

The horizontal lines of Table 1-1 illustrate the resulting influence rates of the four categories for each study. It appears that 61% of the factors on average are related to management issues such as supervision incompetency (e.g. delays, communication problems), resource management (e.g. lack of materials, workforce), disruption of labourers and congestion implications. The next most significant factors deal with technological challenges (15.71%) like rework and design complexity. Workforce problems exerted the same influence (15.55%) and included factors such as absenteeism, skills, age, and late arrivals. Lastly, external influences proved to have a small contribution of just 7.64% in total

because of their random nature (e.g. weather, law regulations). The knowledge that management is the main reason behind low labour productivity still does not explain why project managers fail to take the correct decisions. Hence, further analysis is required in order to achieve a clearer understanding of the problem. On this basis, an alternative comparison is proposed.

Table 1-1: Comparison "A" of factors affecting labour productivity in construction.

Survey Data	Management	External	Workforce	Technological
(Cundecha, 2012)	47.12	9.75	20.72	22.41
(Kuykendall, 2007)	66.91	4.79	22.70	5.60
(Jarkas & Bitar, 2012)	45.96	11.44	10.93	31.67
(Makulsawatudom et al., 2004)	73.20	6.35	5.12	15.33
(E. Lim & Alum, 1995)	72.85	6.76	20.39	0.00
(El-gohary & Aziz, 2014)	60.52	6.79	13.43	19.27
Average (%)	**61.09**	**7.64**	**15.55**	**15.71**

This second ("B") comparison categorizes the factors that affect labour productivity based on their level of influence on on-site tasks. In this chapter, the term *on-site* refers to construction jobsites. These categories are: a) the internal on-site that consists of factors that are directly related to on-site tasks (i.e. overtime, safety, resources, scheduling, rest areas, transportation, congestion, disruptions, site layout, supervision, rework, skills, fatigue, absenteeism, late arrivals and unscheduled breaks), b) the internal off-site that includes factors which have an indirect side-effect relationship with on-site tasks (i.e. age, motivation, turnover, payment, personal issues), and c) the external off-site that contains factors which are out of the range of a jobsite (i.e. weather, owner change of orders, law regulations, site access and designs). Figure 1-4 illustrates this categorization graphically.

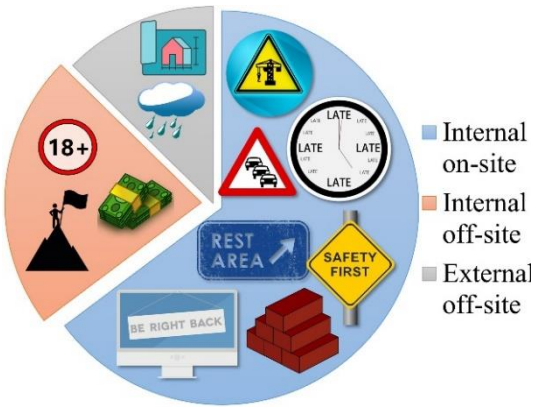

Figure 1-4: Factors affecting labour productivity in construction.

Table 1-2 presents the results of the second ("B') comparison of factors that affect labour productivity. This table shows that the internal on-site category is the most important as it contains 64.77% of the total factors. The other two categories, the internal on-site and the external off-site, are less significant with 22.56% and 12.67% respectively. In particular, the external off-site factors are considered fixed

risk in construction since weather and higher hierarchy levels (e.g. owners) can be mitigated but not fully controlled. The same holds for the internal offsite factors. For instance, age does not affect performance of workers in the same way, as it also depends on parameters such as their physical condition, personal weight. Additionally, although lack of motivation is responsible for 5-14 wasted work-hours per week, it has been suggested that lack of materials, overcrowded areas and rework are important factors behind this demotivation (Ng et al., 2004). Hence, it is logical to focus on the characteristics of the factors of the internal on-site category. An interesting observation about this category is that all its factors can be easily detected and eliminated during productivity monitoring. For example, although disruptions during a hoisting operation, caused from tracks traversing the same region, result from congestion, they can still be identified by project managers while they monitor this hoisting task's productivity. Given this, the questions that arise are: "*Are we monitoring labour productivity proactively? If not, why?"*

Table 1-2: Comparison "B" of factors affecting labour productivity in construction.

Survey Data	Internal on-site	Internal off-site	External off-site
(Cundecha, 2012)	42.95	18.41	38.64
(Kuykendall, 2007)	82.42	12.79	4.79
(Jarkas & Bitar, 2012)	48.46	40.03	11.51
(Makulsawatudom et al. 2004)	75.61	22.49	1.90
(E. Lim & Alum, 1995)	75.06	18.18	6.76
(El-gohary & Aziz, 2014)	64.13	23.48	12.39
Average (%)	**64.77**	**22.56**	**12.67**

1.4. Current state of practice in monitoring of labour productivity

The current state of practice in monitoring construction workers is mainly based on manual observation and work sampling techniques (AMAC Consultants, 2004; Carrasco, Hall, & Sweany, 2013; Dozzi & AbourRizk, 1993; Shehata & El-Gohary, 2011). This section groups existing practices in two categories given the level of detail they return about worker performance.

The first category does not identify any management issues. It only provides an indication about a worker's performance state as "productive/unproductive" or "working/non-working" based on an observation sample. This category contains: a) the five minute rating that relies on observations of tasks for a small period of time, b) the field rating that calculates the fraction of productive workers over the total number of productive and unproductive ones in order to pinpoint whether something is wrong with a task's productivity (i.e. $\frac{\sum productive}{\sum(productive+unproductive)} \leq 60\%$), and c) the work sampling that analyses a small sample of data collected based on statistical sampling theory and returns a general assessment about worker performance e.g. effective, contributory and non-effective.

The second category comprises practices that monitor workers in greater detail. Such practices are: a) the crew balance charts that illustrate each worker's performance over time using a stopwatch, b) the field surveys that use questionnaires (e.g. foreman delay survey, craftsman questionnaire) in order to provide an understanding of the possible causes of bad performance, c) the photographic and video based techniques that rely on human operators to review video data and still images, and lastly d) the Method Productivity Delay Models (MPDM) where an observer fills in a form regarding the cycle time of a leading resource along with the causes of delays. This second category provides a lot more information about potential productivity issues compared to the first, but it is still labour intensive and time consuming considering that a large number of workers must be monitored on a daily basis for the entire duration of their work shifts. For instance an expert engineer needs on average 3.5 hours to extract the productivity of a 14 minute video capturing a worker while installing a scaffold (Gong & Caldas, 2011).

In summary, collecting data with high frequency and extent is a cumbersome process due to the manual procedures involved in current practices. Given the fact that a construction site has multiple activities that take place simultaneously and spread across a large area, i.e. excavation works, concrete pouring, the task of recording everything in detail becomes time consuming and labour intensive. Therefore, most of the time problems are first detected (e.g. delays, congestion, lack of materials, absenteeism) and then reported by the project managers, after which corrective actions can be taken. However, this strategy entails a delayed reaction time. As highlighted by Al-qahtani et al. (2007), the wasted time during a hoisting operation had first to be detected through a crew balance chart (Figure 1-5), before the project manager was able to implement the appropriate changes. In addition, given the fact that most tasks are not fixed processes and that influencing factors are not periodic phenomena, a significant amount of time may be spent until the malfunctions are redetected. Therefore, project managers should be monitoring labour productivity on a constant basis. However, this is not currently feasible given the existing practices. For all these reasons, project managers do not have a clear and solid idea of the state of the productivity of workers or the reasons behind the problems (Navon & Sacks, 2006). Hence, the question that arises is: "*How can labour productivity be monitored proactively?*"

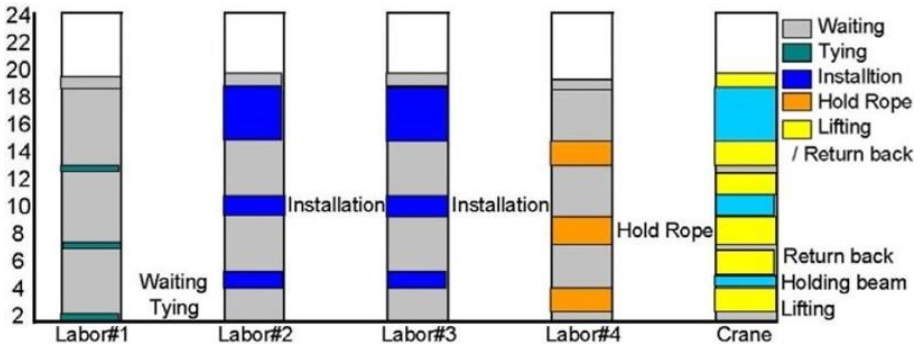

Figure 1-5: Crew Balance Chart of a pre-cast beam installation operation. (Al-qahtani et al. 2007).

1.5. Conclusions and thesis overview

To date, the construction sector has not yet managed to improve labour productivity over the past 5 decades. This is due to factors that affect on-site construction tasks negatively. Almost all of them are related to the way that productivity is monitored. Construction project managers currently evaluate worker performance based on questionnaires, manual observations, and work sampling practices. Construction requires proactive monitoring of labour productivity in order to detect issues sufficiently early. However, this is not feasible as current practices are labour intensive and time consuming due to the large number of employees and the long lasting tasks. Therefore, this thesis presents a method that performs proactive monitoring of labour productivity regardless of the type of tasks the workers are involved in. The remainder of this thesis is structured as follows: Chapter 2 discusses the current state of research on automated monitoring of labour productivity in construction. The chapter closes with a presentation of the overall proposed framework that aims to address the existing gap in knowledge and the objectives. This framework consists of three intermediate methods which are separately presented and evaluated in Chapters 3, 4 and 5. Then, Chapter 6 presents the conclusions of the research presented in this thesis along with recommendations for future work.

2

Current state of research in monitoring of labour productivity

This chapter reviews the latest studies that focus on monitoring of labour productivity. Current studies are divided in two main categories based on the methods they employ to infer productivity. The first contains the region-based studies that link the location of workers to regions of management interest (work zones) such as steel fixing zone, concrete pouring zone. The second consists of the activity-based studies that detect and link activities such as bending, hammering, and drilling to specific tasks. The aim of this chapter is to conclude whether existing research has managed to address the limitations of the current state of practice as presented in Chapter 1 and to devise a solution that addresses the existing gap in knowledge.

At this point, it is important to explicitly define some terms that will be repeatedly used throughout this thesis. These terms are the following:

- Task is a construction related operation such as brick laying, scaffolding, hoisting etc.
- Sub-task is an operation related to a task. For example, a steel task consists of sub-tasks such as placing, fixing, and picking the reinforcement bars (re-bars).
- Activity is the physical description of a sub-task such as bending, stretching, sound, strain etc.
- Pattern is a repetitive activity of a task.
- Construction entity is any worker or earthmoving equipment (e.g. trucks, excavators, cranes) that performs a construction related operation.

The remainder of this chapter is structured as follows. Sections 2.1 and 2.2 analyse the current state of research in monitoring of labour productivity in construction. Section 2.3 discusses the objectives, scope, and aims of this thesis. Section 2.4 presents an overview of the overall proposed framework presented in this thesis.

2.1. Region-based studies

Region-based studies monitor labour productivity through the time construction entities spend at zones of management interest (e.g. excavation zone, concrete pouring zone). In order to achieve this, the location of monitored entities is tracked across the jobsite. The studies of this category are sub-divided into tagged and tag-less given the methods they use for calculating the location data of construction entities.

2.1.1 Tagged studies

The tagged (RF tagged) studies employ tags to extract the location of construction entities. These tags are physically attached on workers and earthmoving equipment. The most frequently used tags are based on technologies such as the Global Positioning System (GPS) (see (a) in Figure 2-1), the Radio Frequency Identification system (RFID) (see (b) in Figure 2-1), and of the Ultra-Wide band system (UWB) (see (c) Figure 2-1).

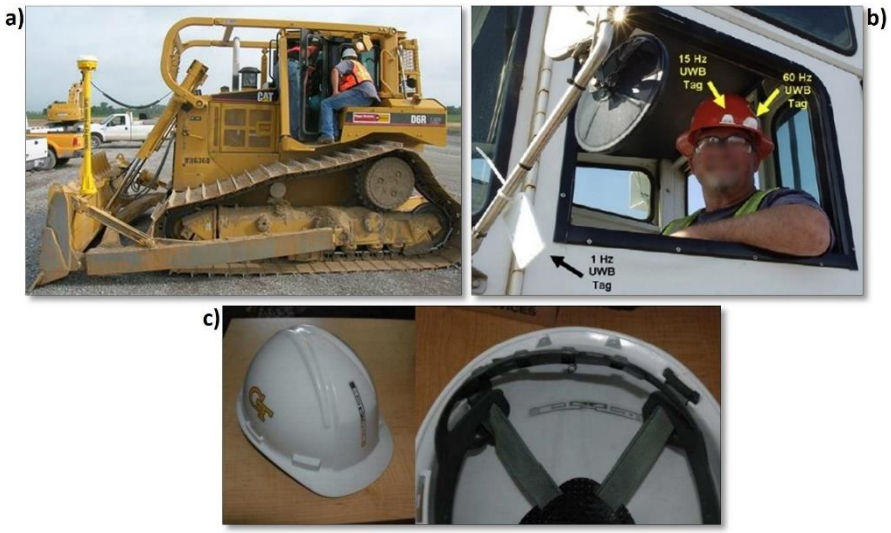

Figure 2-1: (a) GPS mounted on earthmoving equipment (RTD FasTracks, 2014). (b) Passive tags on worker hardhats (Sedehi, 2010). (c) UWB tags placed on worker hardhats (T. Cheng et al., 2011).

The GPS navigation system was developed by the US and is based on a constellation of 24 orbiting satellites. A receiver requires a clear view of at least 4 satellites in order to calculate its location on earth. Because of this, GPS has a poor performance indoors and in dense urban environments due to the multipath effect (Kos et al., 2010), but has an unlimited coverage outdoors with an accuracy that ranges from 1m to 2.5m (Pawłowski, 2015). The RFID positioning systems locate a tagged object within the range of readers with an accuracy of 15cm (Chawla et al., 2010). A full installation requires receivers for detecting the tags and antennas for transmitting the signal between them. Tags are either passive or active (Beinat et al. 2007). Passive tags can be detected only within a range of 1m from the readers

(Ubisense.net, 2017). On the other hand, active tags that are powered from an energy source (e.g. battery) are located within a significantly larger range of 100m (Ubisense.net, 2017). An important advantage of RFID over GPS systems is that the former perform better indoors. Lastly, UWB systems are based on active tags (battery – powered) and sensors. Such systems can locate targets within 100m range (Ubisense.net, 2017) with an accuracy from 5cm to 10cm (Connell, 2015).

The above systems provide the input data of tagged studies. On this basis, the speed and the location of a haul truck were both combined for monitoring its productivity while performing an earthmoving operation (Hildreth et al. 2005). If the haul truck's location was within the range of fixed known distances from specific work zones (e.g. load and dump zones), then the time during which its speed was equal to zero was converted into labour input. On the other hand, the labour productivity of workers, was monitored by linking their presence at predefined work zones (T. Cheng et al., 2011; T. Cheng et al., 2013; Jiang et al., 2015; Navon & Goldschmidt, 2003; Sedehi, 2010). For instance, if a concrete worker is located at zones “A” and “B” which are scheduled for concrete pouring, then the total time the worker spent in these zones is considered productive and equal to his/her labour input. The studies of this category also sub-divide the areas between the actual work zones into waiting and travelling zones, for a more detailed insight of worker productivity (see Figure 2-2).

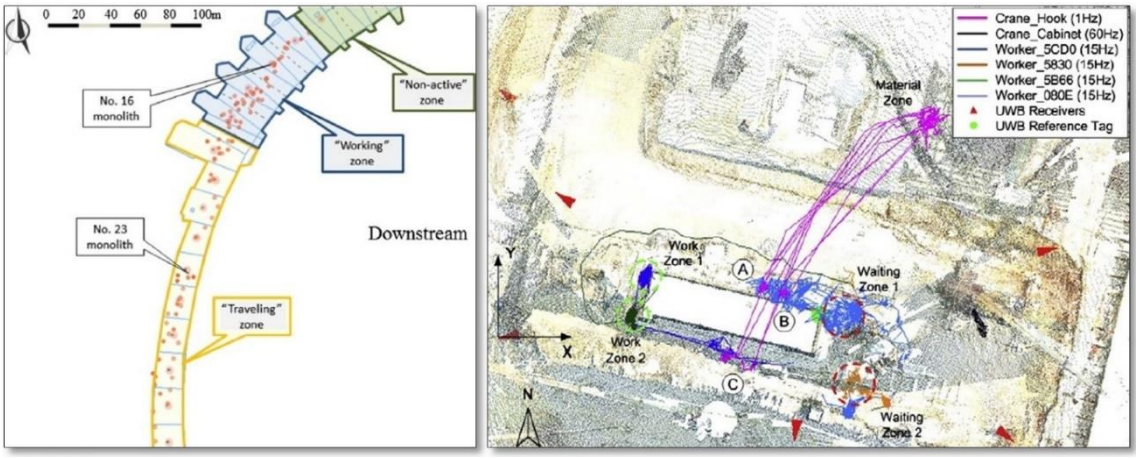

Figure 2-2**:** Monitoring of labour productivity given worker presence within predefined work zones. (left: Jiang et al., 2015, right: Cheng et al., 2011)

The most important disadvantage of the tagged studies is that they can neither identify the unproductive time (idle time) nor the low productivity pace. For example, even if a worker is located in the correct work zone, but without performing any task due to shortage of materials or congestion he/she will be still considered productive. This is due to the fact that labour productivity is monitored only based on the presence of workers at work zones. The tagged studies do not provide any extra information about what really happens within these work zones. In addition, the purchase and maintenance of multiple

tags impose a regular cost in the long term (Nasr et al., 2013). Last but not least, the physical attachment of tags creates a feeling of discomfort to workers (Juels, 2006).

2.1.2 Tag-less studies

The tag-less studies rely on computer vision-based 2D tracking methods in order to calculate the location of workers. This location is 2D instead of 3D. Therefore, entities are tracked only within the range of a camera's view. This type of tracking is non-obtrusive as it processes video data collected through surveillance cameras used for security purposes.

The studies in this category convert the location data into labour productivity through two approaches. The first, links the presence of tracked entities (workers, earthmoving equipment) to specific work zones similar to tagged studies (see section 2.1.1) (Bügler et al., 2014). For this reason, the ambiguity about what really happens within these zones arises again. The second, fits the monitored entities to operation process models (Gong & Caldas, 2010, 2011; Yang et al., 2014). Such models (Halpin & Riggs, 1992; Martinez & Ioannou, 1994): a) break down the construction tasks into sub-tasks (semantic context), b) describe how the sub-tasks relate to specific work zones across the jobsite (spatial context), and c) define the sequential order (i.e. workflow) between the sub-tasks (temporal context). Figure 2-3 illustrates how a concrete pouring task is monitored automatically through such an operation process model.

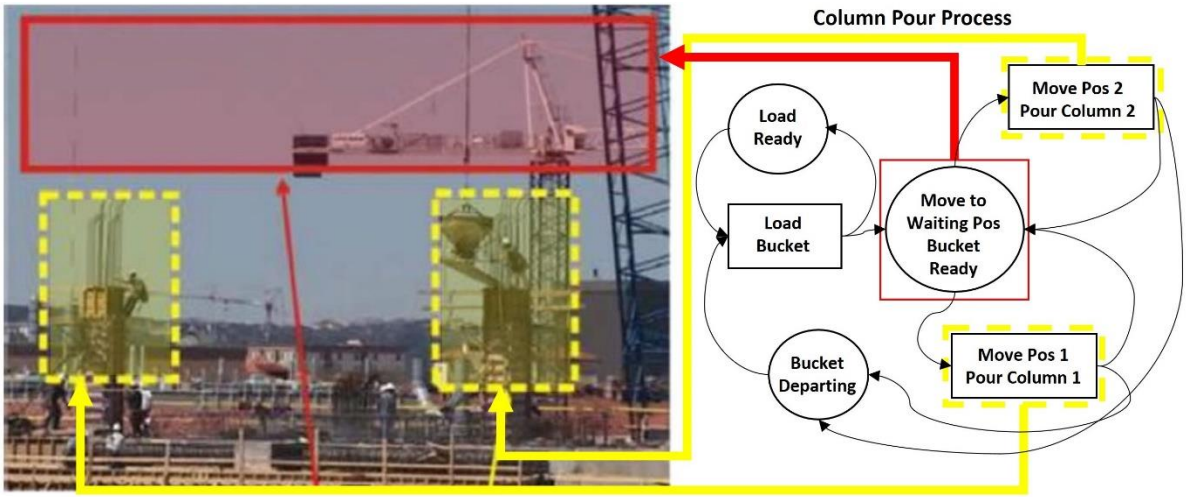

Figure 2-3: Operation process model of a concrete pouring task (Gong & Caldas, 2010, 2011).

In the above Figure 2-3, the work zones (yellow and red rectangles) are manually marked and linked to the operation process model by the user (Gong & Caldas, 2011). In the example of this figure, if the tracked concrete bucket passes through the marked work zones in a sequence that agrees to the operation process model, then the total time the bucket spends in these zones is equal to the labour input of this task. The total output is equal to the number of buckets poured. This study translates successfully and

with high accuracy the video data into labour productivity. However, it relies on human intervention in order to adjust the appropriate process model to each entity. It takes 5-10minutes for an operator to achieve this. Such adjustments should be repeated for every entity on a daily basis. The large numbers of workers and earthmoving equipment entails that such type of studies will be labour intensive if applied in practice.

2.2. Activity-based studies

The studies of this category firstly detect and secondly link activities to specific construction tasks in order to monitor labour productivity. These activities are the physical description of tasks. For example, a brick layer bends to pick up bricks and stretches his arms to place them. Bending and stretching are both activities that describe the brick laying task. This type of studies exploit posture, physiological (e.g. heart, breathing rate) and audio data.

The posture-based studies have been used for monitoring both the labour productivity of earthmoving equipment (Golparvar-Fard, Heydarian, & Niebles, 2013; Zou & Kim, 2007) and construction workers (Bai et al., 2008; Golparvar-Fard et al., 2013; Khosrowpour et al., 2014; Yang et al., 2016; Zou & Kim, 2007). Posture data are detected via feature descriptors such as the Histogram of Oriented Gradients (Dalal & Triggs, 2005) and skeletisation algorithms (Abu-Ain et al., 2013). Figure 2-4 illustrates with coloured rectangles the detected posture data of a truck in (a-c), and an excavator in (d-f), while filling dumping, moving, hauling/swinging, digging, and dumping.

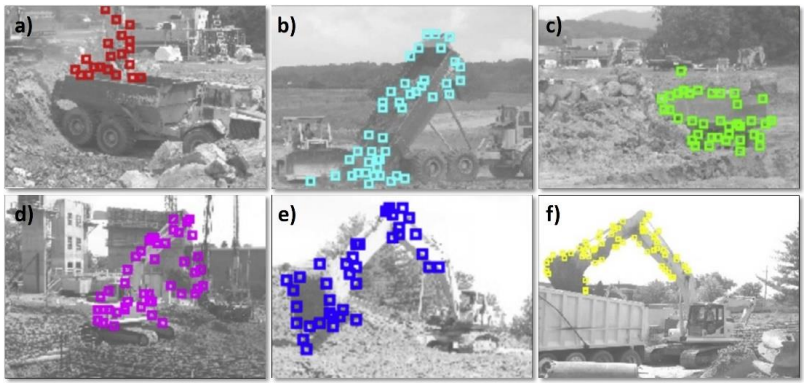

Figure 2-4: Productivity monitoring of earthmoving equipment (Golparvar-Fard et al., 2013).
(a: filling, b: dumping, c: moving, d: hauling/swinging, e: digging and f: dumping)

Machine learning-based algorithms such as Support Vector Machine Classifiers (SVMs) (Brereton & Lloyd, 2010) and Artificial Neural Networks (ANNs) (da Silva et al., 2017) are then trained to link (label) the detected activities to construction tasks. The highest achieved accuracy so far is equal to 59% (Jun Yang et al., 2016). In particular, this study was tested on workers while performing 11 types of tasks i.e. brick laying, transporting, plate cutting, drilling, re-bars fixing, nailing, plastering, shovelling,

bolting, welding, and sawing. The authors admitted that this low accuracy was due to the fact that most of these tasks were not distinguishably described by posture data. On the other hand, posture-based studies perform very well (accuracy >80%) for the case of earthmoving equipment, as such entities have a small but well defined range of postures. For example, an excavation task performed by a dump truck is described only by two postures. The first depicts the unloading of materials and the second the transportation of materials. In addition, earthmoving equipment is used for only one type of tasks whereas workers perform a much larger variety.

The second group of studies of this category, exploits physiological data such as heart rate (beats/minute), breathe rate (breaths/minute), body's force and angular rate (Akhavian & Behzadan, 2016; Chen et al., 2017; Gatti et al., 2014). Such data are acquired through physiological status monitoring (PSMs) and inertial measurement unit (IMU) wearable sensors. The physiological data are used for training machine learning-based classification methods similarly to the studies that exploit posture data. In general, physiological-based studies follow the methodology presented in Figure 2-5.

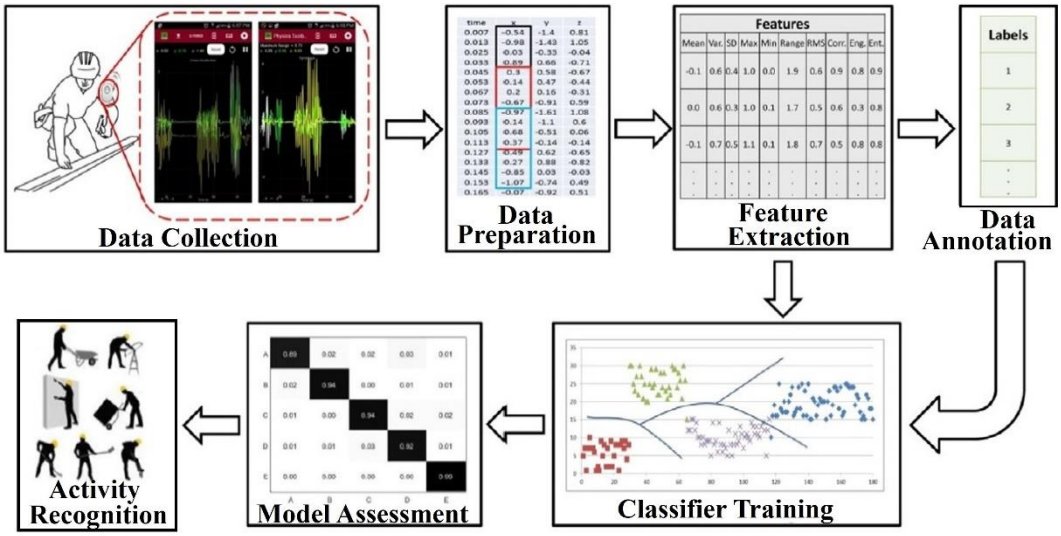

Figure 2-5: Methodology of physiological-based studies for recognizing activities. (Akhavian & Behzadan, 2016)

However, it has been proven that heart and breathe rates cannot establish any relationship with individual's labour productivity (Gatti et al., 2014). On the other hand, body's force and angular rate, extracted with accelerometers and gyroscopes of IMUs sensors, achieved a promising performance (≈80% accuracy) in terms of detecting and labelling activities such as hammering, sawing, turning a wrench, loading/unloading/pushing a wheelbarrow (Akhavian & Behzadan, 2016). Physiological-based studies have been also successfully used for identifying abnormalities in the performance of workers (awkward postures) for health and safety purposes (J. Chen et al., 2017). Their main limitation is that they rely on data collected with wearable sensors that give rise to privacy issues.

Lastly, audio data which are recorded by microphones placed at construction jobsites have also been exploited for monitoring the productivity of construction entities (Cheng et al., 2017; Weerasinghe & Ruwanpura, 2010). These audio-based studies are applicable only to tasks that produce discrete sounds such as nailing, hammering, excavating, and drilling. Although they have managed to successfully remove background noise, they are still not designed to monitor the labour productivity of multiple entities that perform similar tasks simultaneously.

2.3. Summary of current state of research

Monitoring of labour productivity relies on the calculation of the labour input and output. Existing studies focus mainly on the former. This is because the measurement of the latter is quite straightforward through visual inspection (e.g. meters of wall constructed, number of columns poured with concrete). The labour input is equal to the amount of time each construction entity spends on a task. Current state of research in monitoring of labour productivity is grouped in two categories: a) the region-based, and b) the activity-based.

The region-based studies infer the labour input by measuring the time each entity spent in pre-defined work zones. In each of these zones, one or more specific tasks are scheduled to take place such as steel fixing, concrete pouring, bricklaying etc. The region-based studies are subdivided into tagged and tag-less. The former group of studies, calculates the location of workers through sensors (UWB, RFID and GPS), whereas the latter uses computer vision-based tracking methods. These studies are limited by: a) the way the location of entities is calculated, and b) the way the labour input is inferred. With regards to the first limitation, the tagged studies are obtrusive and costly whereas the tag-less studies are restricted to monitoring workers only within the range of a camera's view. The second limitation, relates to the ambiguity that arises when workers are idle instead of productive even if they are located within the correct work zones according to schedule. The tag-less studies link operation process models to work zones to alleviate this ambiguity. In practice, this is not easily applicable as it takes 5 to 10 minutes for an operator to select and adjust the appropriate operation process model to every construction entity.

On the other hand, the activity-based studies estimate the labour input by measuring the duration of the total activities each entity performed. These activities depict sub-tasks of specific tasks. To achieve monitoring, the activity-based studies rely on posture, physiological and audio data. However, none of these has been proven robust. Firstly, the studies that use posture data are accurate only if tasks are depicted by distinguishable postures or involve earthmoving equipment. This is due to the disproportionately small number of human postures compared to the large number of construction tasks. For example, two workers stretch (activity), but one in order to place a pipe and the other to tighten a scaffold. In this case, two similar postures, depict two different tasks. The opposite holds for

the earthmoving equipment. In fact, each posture of such entities depicts a very specific activity. Secondly, the studies that analyse physiological data are restricted mainly by privacy conflicts similar to tagged studies. Thirdly, the activity-based studies that exploit sound data are applicable only if tasks produce dissimilar sounds.

Chapter 1 concluded that all construction entities should be monitored proactively in order for the labour productivity to be improved. Proactively means that multiple workers should be monitored at the same time on a daily basis. Current practices have failed to achieve this, as they are labour intensive and time consuming. Important research has been conducted in order to address these shortcomings. However, none of them has managed to propose a method that can *accurately*, *unobtrusively*, and cost and time *efficiently* monitor the labour productivity of multiple construction workers at the same time. Accurately means that the project manager knows the number of productive hours of each worker with an accuracy of 100%. Unobtrusively means that workers are monitored without any feeling of discomfort. Efficiently in terms of time and cost means that it is practically feasible for a construction company to monitor the labour productivity of multiple workers during their entire work shifts on a daily basis. This gap in knowledge exists because current studies are mainly tailored to entities that perform specific tasks. However, it applies only to workers as activity-based studies have already successfully addressed the monitoring of earthmoving equipment's productivity. Therefore, the main objective of this thesis is to develop a fully automated framework for monitoring labour productivity of construction workers regardless of the type or number of tasks they perform through their work shift. The aims of the researcher are to:

- **Aim 1**: Track construction workers unobtrusively.
- **Aim 2**: Extract the labour productivity of construction workers accurately.
- **Aim 3**: Monitor the labour productivity of construction workers proactively.

2.4. Hypothesis & Proposed framework

The scope of this framework is to monitor the labour productivity of multiple workers at the same time. Figure 2-6 illustrates the overall proposed framework. The skewed parallelogram shapes refer to methods and the circular to inputs/outputs. The framework consists of two main methods illustrated with black coloured skewed parallelogram shapes. The output of the first is the input for the second. In more detail, the cyan coloured skewed parallelogram shapes depict methods of novel contribution, whilst the uncoloured skewed parallelogram shapes are methods taken from literature. The inputs of the framework are video data streamed from multiple cameras, whilst the output of the framework is the total productive and unproductive time spent by each worker. This thesis hypothesizes that task productivity of construction workers can be monitored through their trajectory data.

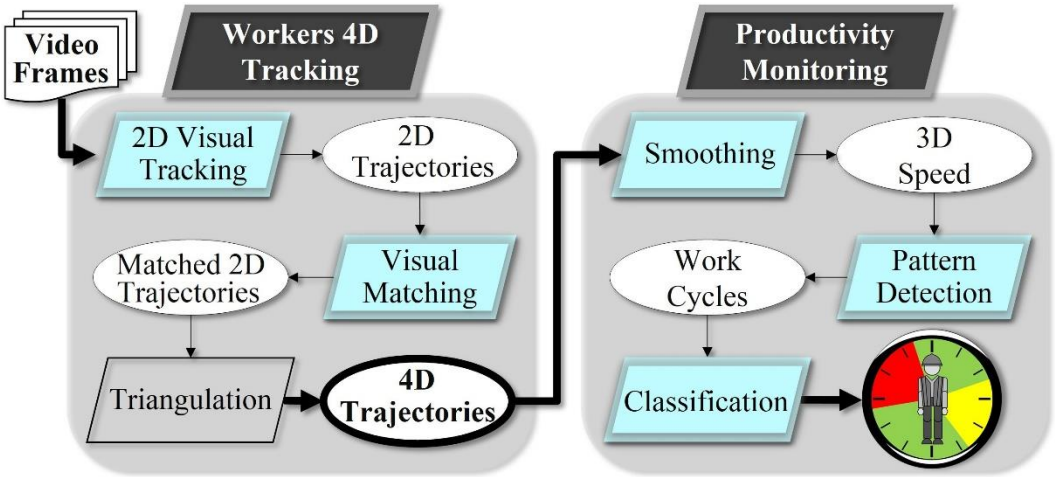

Figure 2-6: Overall framework for automated construction worker task productivity monitoring.

The labour productivity is calculated by diving a worker's total output over the total input (Shehata & El-Gohary, 2011). The determination of output is quite straightforward through visual inspections (e.g. number of pipes installed, number of m^3 being excavated). Hence, this thesis focuses only on the input. The main assumption of this thesis is that all construction-related tasks fit to the same pattern. This pattern dictates that *if a worker's "move" is followed sequentially by one "stop" and a second "move", then these three semantic events define a work cycle*. This assumption is based on the fact that workers "stop" in order to perform a construction-related task and they "move" to start another. In construction, a work cycle is defined as the total time a worker spends on a task (Dozzi & AbourRizk, 1993). Hence, the duration of a work cycle is equal to the duration of the semantic "stop" event. Sequentially, the duration of all work cycles is equal to the labour input of a worker. Therefore, the labour input can be extracted by detecting these work cycles. The four equations in Figure 2-7 form the core of the proposed framework.

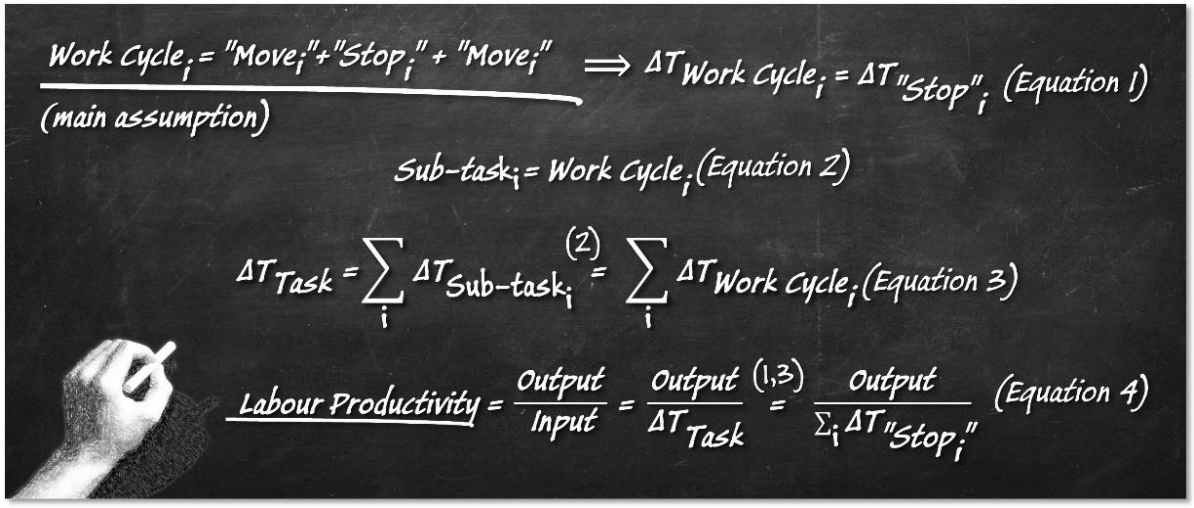

Figure 2-7: Assumptions of proposed framework.

The first method of the proposed framework is a computer vision-based method for 4D tracking of construction workers. This type of tracking is unobtrusive as it is tag-less. The input data are videos collected through the cameras of jobsites' surveillance systems. It returns one 4D trajectory for every worker as output. These 4D trajectories depict the 3D (X, Y, and Z) location of workers across the entire range of a jobsite over time. This 4D localization overcomes the limitation of previous tag-less studies that monitored workers only within a camera's view. An intra and an inter camera tracking are performed sequentially in order to achieve this 4D tracking. The former matches workers under the same unique ID across subsequent frames of a camera, whilst the latter matches workers across multiple cameras. A computer vision-based 2D tracking method is developed in order to perform the intra camera tracking. It returns one 2D trajectory for every worker monitored. A computer vision-based matching method is also devised in order to perform the inter camera tracking. This visual matching method returns the 2D trajectories that belong to the same worker from all cameras. Then, a triangulation method, that is taken from literature (Hartley, 1997), is applied in order to convert the 2D trajectories into 4D. The 4D computer vision-based tracking method addresses the first aim of this thesis.

The second method of the proposed framework is productivity monitoring. It uses the output of the 4D tracking method as input. Initially, a smoothing method removes the noise from the 4D trajectories. Then, the 4D trajectory of each worker is partitioned into smaller 4D sub-trajectories. The 3D speed values of these partitions are exploited to cluster them into work cycles based on the main assumption of this thesis (see Figure 2-7). The accurate detection of these work cycles addresses the second aim of this thesis as their total duration is equal to the labour input of construction workers. The 3D speed values depict the motion of workers along the floor (XZ) and the vertical plane (Y). The detected work cycles are classified as: a) unproductive, b) normal productive, and c) abnormal productive. Initially, they are classified as either productive or unproductive through region-based classification that splits the jobsite into two types of areas, "active" and "inactive". The former contains the areas of the jobsite where tasks such as excavation, brick laying are performed. The latter consists of areas where no construction-related tasks take place. These are the: a) rest areas, b) materials' storage areas, and c) office areas. The work cycles that take place at "active" areas are classified as productive while those that take place at "inactive" areas are classified as unproductive. Then, the productive work cycles are further classified in order to detect potential abnormalities in the pace of the labour input. The durations of the productive work cycles are compared for this purpose. Those with the highest duration are classified as potentially abnormal and the rest as normal. This second classification is used as an indicator. It shows project managers whether something appears to be "wrong" with workers' productivity pace. Such indication can be very beneficial considering that several factors affect labour productivity in a negative way as shown in Chapter 1. Managers can then look into the video data at the time of the day the abnormalities occurred and check whether something was actually incorrect with these work cycles. This way problems are identified and treated fast. The productivity monitoring method does not need any prior knowledge about the type or the number of tasks workers perform.

Therefore, labour productivity of multiple workers can be monitored at the same time. This entails proactivity. With this, the third aim of this thesis is addressed.

Each of the methods of the proposed framework are presented in detail in the following chapters. Chapters 3 and 4 present the visual tracking and matching methods respectively. The proposed framework concludes with Chapter 5 that analyses the pattern recognition method for task productivity monitoring of construction workers.

2.4.1. Experimental set up

The performance of each method of the proposed framework is tested with a C# implementation in Microsoft Visual Studio.Net framework running in a Windows 8.1 operating system. The integrated development environment is Visual Studio 2013, using Windows Forms (WinForms). A desktop PC with the following specs is used: Intel core i7 CPU, 4.0GHz, and 32 GB RAM.

The cameras used in the experiments are two GoPro cameras, black edition 4 with a 1920x1080 frame size, and selected 90^0 narrow field of view to reduce the distortion of camera fish eye effect. Both cameras are mounted in such a way that monitored workers are captured within their overlapping field of views. The 4D trajectories have as reference the local coordinate system of one of the two cameras used (see Figure 2-8).

This thesis validates the proposed framework with data collected both off-site and on-site. This is aligned to construction industry's vision to reduce cost by increasing the off-site production of pre-fabricated construction elements. One of the largest construction companies in the UK stated back in 2012, that the company's aim is to perform 70% of the total construction in off-site facilities (Ferguson, 2012). The main discrepancy between off-site and on-site jobsites is the level of organization in terms of materials' management. In an off-site pre-manufacturing facility workers and materials are pre-allocated within well designated areas. On the other hand, in an on-site environment materials and workers are more difficult to control as work zones change constantly. Therefore, more engineers are required to monitor the workers within each of these work zones and to manage the supply of materials and equipment.

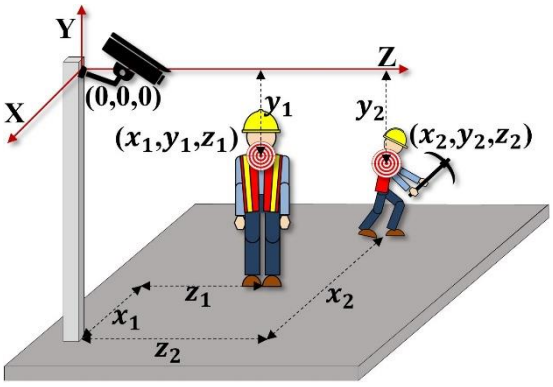

Figure 2-8: Camera centred coordinate system.

3

Adaptive computer vision-based 2D tracking of workers in complex environments

In this chapter, a computer vision-based method for 2D tracking of construction workers is presented. This type of tracking provides the 2D trajectories that are required in order to calculate the 4D trajectories (3D location over time) of workers. The overall proposed framework presented in Chapter 2 uses these 4D trajectories as input data. To date, such trajectories are extracted either by manual observation techniques, the so called spaghetti diagrams (Nyström & Per, 2009), or by tag-based practices. The former is impractical considering the large number of workers that must be monitored on a daily basis. The latter is accurate but it is not welcomed by the personnel as discussed in Chapter 2.

3.1. Introduction

Current vision-based tracking technologies provide automated and tag-less monitoring (Accuware.com, 2017; Projectfine.eu, 2010). These types of technologies are usually applied for tracking targets such as pedestrians and sports players. However, they fail when targets are either partially occluded or very close to each other, and will therefore not provide an efficient way of tracking construction workers. This is due to the challenging nature of construction sites. Such work environments often contain multiple uniformed workers with similar appearance under occlusions/illumination/scale/posture variations, and who may exhibit abrupt changes in movement over the course of their task.

A successful monitoring system for construction workers as proposed by Cheng et al. (2011) has to satisfy the following 5 criteria: a) low cost and maintenance, b) non-disturbing regarding the tasks, c) applicable both indoors and outdoors, d) accurate, e) high data frequency, and f) non-intrusive (privacy issues) for the personnel. Current vision based-tracking complies with 4 out of 5 of these criteria. Accuracy is the only shortcoming. This is mainly because in a construction work environment

workers are excessively congested and occluded. To the best of the authors' knowledge, there is no existing monitoring practice that satisfies all of the above criteria.

In summary, the main problem of existing practices for monitoring construction workers is the lack of accuracy. The aim of this chapter is to address this shortcoming by proposing a novel 2D vision-based method for tracking workers in complex environments. The proposed method uses as input video frames with marked regions, highlighting the positions of workers in view in the first frame. The normal distribution of workers' walking speed values at various depth scales are experimentally measured and used to predict the marked regions' future positions. An offline trained SVM classifier detects the sub-regions of every marked region that contain high visibility worker apparel. A filtering method is then applied to denoise these sub-regions. The boundaries of the apparel's new position are set by implementing a distance based clustering method. Lastly, the proposed vision-based 2D tracking method updates the adaptive model with the current position and appearance in order to be able to predict the future position and remove noise. With such updates, the proposed method adapts to posture/scale/illumination variations, abrupt movements, occlusions and congestion caused by targets similar in appearance (co-workers).

This chapter is structured as follows. Section 3.2 analyses the current state of research on vision-based 2D tracking methods. Then, the proposed solution and methodology are discussed in sections 3.3 and 3.4 respectively, and evaluated in section 3.5. Conclusions are presented in the final section 3.6.

3.2. Background

There has been much research on 2D visual tracking, so that a complete review of the current state of the art is beyond the scope of this thesis. Hence, this section focuses on studies that address the challenges that most commonly appear while tracking workers through the cameras of a construction jobsite's surveillance system. Such challenges are: a) scale variations, b) appearance similarity due to Hi-Vis apparel, c) occlusions caused, for example, by work benches, structural elements and workers, d) posture variations, e) abrupt movement, f) background clutter, g) congestion caused by workers that overlap with each other, and h) illumination variations.

Vision–based tracking methods search to localize the same target across subsequent frames. Each target is represented via a unique appearance model to achieve this. This model describes both the shape of the target and the appearance features within it. Shape can be a rectangle, contour, point(s), skeleton, or ellipse, whilst the appearance features are usually colour, edges, and texture. Both the tracking environment and the type of the target (rigid, non-rigid) are taken into account for the selection of the most appropriate shape and appearance features. Yilmaz et al. 2006 groups tracking methods into three categories based on the appearance model they use. These are: a) the kernel-based methods that use both shape and appearance features (Comaniciu et al., 2003; Ross et al., 2008), b) the point-based

methods that track point(s) shaped targets (Ding and Zhang 2012; Yang and Cao 2013), and c) the silhouette-based methods that rely on edges (Hu et al., 2013).

Kernel-based methods are better for tracking construction resources (equipment & workers) under occlusion, illumination, and scale variations (Park et al., 2011). Such methods mask the appearance features with a kernel in the spatial domain. They track (localize) the target across successive frames via a gradient-based optimization method that returns the position of either the global minimum or maximum of a cost function. In that respect, Yang et al. (2010) used the colour features of workers' legs and torso and their spatial arrangement to track workers through a kernel covariance method. This tracking method was successful under congestion and partial occlusions but failed under illumination variations. In addition, the Kalman filter that was used for predicting workers' future position lost track when workers were severely occluded for a long time. The benefits of an online-based tracking method (Ross et al., 2008) were combined to those of a worker detector (Park & Brilakis, 2012) to track multiple workers in the long term (Park and Brilakis, 2016). However, this method terminates the tracking process when no worker is detected within a close distance to the current position of a worker for at least 3 seconds. This failure will be amplified by occlusions as the proposed worker detector requires a clear view of the worker's full body (Dalal & Triggs, 2005). Zhu et al. (2016) instead employed a particle filter point–based method to track a single worker under severe occlusions and posture variations. This method outperforms those that use Kalman filters or gradient-based optimization methods in terms of accurate prediction of a target's future position. This is because the particle filter based methods, as compared to the Kalman filter based methods, are applicable to both linear and nonlinear problems that follow a non-Gaussian distribution (Yilmaz et al., 2006). In addition they do not get confused by local minima or maxima like the gradient-based optimization methods. However, so far particle filtering based tracking methods focus mainly on single target tracking (Shen et al., & Tao, 2015; Sun et al., 2015; Zhou, Fei et al., 2014). In summary, none of the existing state of the art methods is able to tackle all of the aforementioned challenges that relate to the visual tracking problem of construction workers. Hence, this chapter extends the review on studies that focus on visual tracking of targets which are similar to workers. Such targets are either people or non-rigid objects in general. These studies are grouped into two categories. These are the detection-based and the online-based. The former learns a target's appearance offline through a detector, whilst the latter learns online during tracking.

The detection-based methods improve tracking in two ways. Firstly, they automate the initialization of the tracking process. Secondly, they stabilize the performance of the tracker and reduce drifting over background pixels. Breitenstein et al. (2009), Choi et al. (2015), and Li et al. (2016) used a human detector's output as an observation model for tracking multiple pedestrians through particle filtering. The common limitation of all the detection-based methods is that detectors are not invariant to appearance variations. However, such changes in appearance are very common for workers, due to occlusions, congestion, variations in posture, scale, and illumination.

The online-based methods update the appearance model while tracking, to overcome the shortcomings of the detection-based methods. Ross et al. (2008) proposed an incremental learning method that keeps the latest but discards the oldest appearance data after a predefined number of frames. Their method was successful for single target tracking under posture, scale, and illumination variations. However, the authors admit that it fails when targets get partially occluded. Similarly, Ji et al. (2014) proposed an online updated appearance model that consists of two separate models. The first uses random ferns to localize the target, whilst the second compares the target's previous and current templates to validate the localization of the first model. However, as the authors highlight, their method does not perform well on data that contains a fluttering background. Zhao et al. (2016) enhanced the continuously adaptive mean shift algorithm (CAMShift) (Allen et al., 2006) with an updated online structural local sparse appearance model to alleviate this. Still, their method is designed for single human tracking only. Gaxiola et al. (2016) deployed a dynamically updated correlation filter that contains all past templates of a target. This study uses this filter to compare a search area with all past templates of the target. If the comparison returns a score higher than a threshold, then the filter gets updated and tracking continues. A similarity score smaller than the threshold, terminates tracking, and re-initializes it by searching for the target within a larger search area. However, these search methods are not appropriate for workers, as construction sites are usually congested and workers share similar appearance due to their safety apparel (Hi-Vis). Hence, more than one potential comparison match might be enclosed within a single search area. The main common drawback of all the online-based tracking methods is drifting that appears over time (Liu et al., 2014). The tracker loses the target when this occurs. This is due to background pixels that slowly propagate within the online updated appearance model (Bertinetto et al., 2016). In summary, none of the existing studies can fully tackle the visual tracking problem of construction workers (Kalal et al., 2012).

The remainder of this chapter aims to deal with all the aforementioned shortcomings in order to achieve long term tracking of multiple workers within a construction jobsite. The main objectives are: a) to deploy an appearance model that can adapt to all changes of workers' appearance caused by e.g. occlusions, illumination variations, b) to discard background objects that share similar appearance with workers, c) to retain tracking of multiple targets within a congested environment, and e) to absorb abrupt changes of workers' motion. The key research question that this chapter wishes to answer is: "*Can targets that share similar appearance be tracked under abrupt movements and variations (scale, posture, and illumination) within a complex environment?"*

3.3. Proposed solution

This section addresses the tracking issue of construction workers when occlusions, variations in posture, scale and illumination, background clutter (i.e. objects similar to target, abrupt movement) and congestion occur. Figure 3-1 illustrates the flowchart of the overall tracking method. The input data are

video frames, whilst the output data are bounding boxes (rectangular regions) that enclose each target. In this figure, skewed parallelogram shapes refer to the sequential processes applied to all tracked targets within a frame at time t_n. The shaded areas next to each of these skewed parallelograms represent the output of each process. A bounding box is manually applied around each target in the first frame to initialize tracking. Automatic initialisation could be implemented as shown by Park et al. (2012).

The method outlined in the following section relies on an adaptive model to achieve long term tracking of multiple construction workers. This model predicts at time t_{n-1} the regions that enclose each target's position within a frame at time t_n using a prior. Figure 3-1 illustrates with a yellow outlined rectangle the predicted search region for one target. Then, an appearance model searches for unique features of workers within these regions and it returns only the image patches that most likely belong to tracked targets. These patches are depicted as cyan coloured rectangles within the green shaded region. Then, a filtering model is applied in order to discard the image patches that either belong to background objects or simply depict noise. The remaining patches (pink coloured rectangles) within each search region are clustered into a bounding box, which is the output for each target at time t_n. Finally, the adaptive model updates a feature vector about all possible variations (e.g. posture, occlusions) that occurred at time t_n. This allows us to accurately predict each target's future position at time t_{n+1}. The rest of this section describes in detail each of these processes.

The method presented in this chapter makes the following key assumptions: a) tracked workers cannot disappear from frame to frame unless they are within a specific distance from the borders of the frame which is possible to be covered given the camera's frame rate (fps), b) construction workers do not run across construction sites, they walk instead for safety reasons, c) colour features are robust for describing workers under appearance variations caused by occlusions, scene illuminations, posture and scale variations, d) a worker's appearance does not change during congestion due to its short duration, as in practice workers need enough free space to perform their tasks, and e) noise appears at random positions from frame to frame e.g. noise that appears at position A at time t_{n+1} does not re-appear at the same position A at time t_{n+2}.

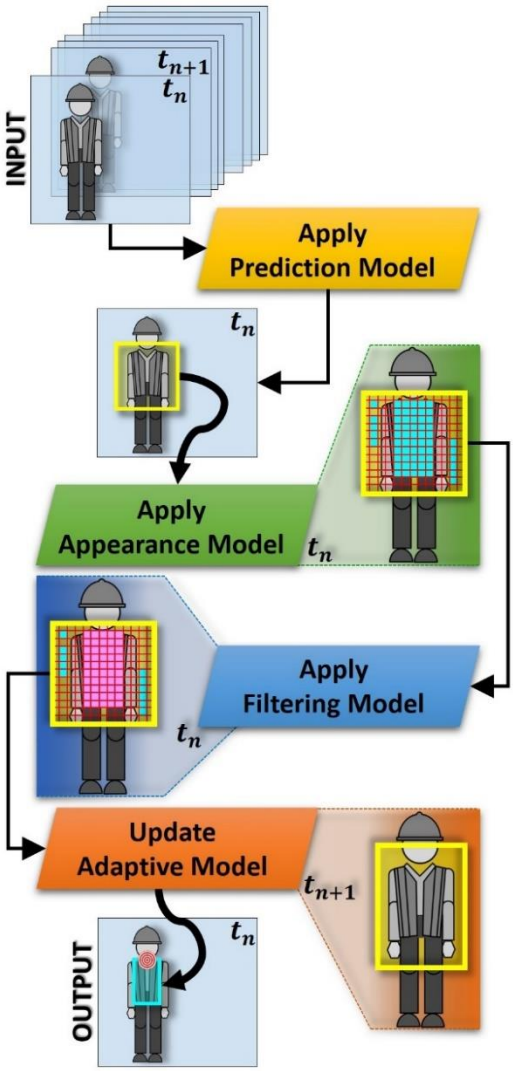

Figure 3-1: Flowchart of proposed computer vision-based 2D tracking method.

3.4. Proposed methodology

This section presents a computer vision-based 2D tracking method of construction workers.

3.4.1. Prediction model

The proposed prediction model relies on a prior to predict a search region R_s for each tracked target within the following frame. This prior is not a threshold. It is derived from the mathematical interpretation of assumptions (a) and (b) of section 3.3, that workers do not run and cannot disappear from frame to frame. Given these assumptions, each tracked worker should be located within a specific walking distance from his/her previous position in the image. The average human walking speed is known and equal to1.45m/s (Boonstra et al., 1993). This needs to be converted into speed in the image, noting that targets which are close to the camera (large scale) cover a larger distance between successive frames compared to targets that are further away (small scale). The camera frame rate (fps) and the scale of the targets are used in order to achieve this conversion.

Firstly, the proposed tracking method groups targets into three scale sizes $\{scale_i\}_{i=1,2,3}$: a) large (i=1), b) medium (i=2), and c) small (i=3). This prediction model exploits the width of targets instead of the height in order to predict the scale of workers, as the latter fluctuates significantly especially under occlusions (e.g. work benches, equipment). A Gaussian distribution is fitted to the data sample of each scale, in order to calculate the probability of a tracked target at time t_n belonging to one of the three scales given its width at time t_{n-1}:

$$P\left(scale_i | width_{t_{n-1}}\right) = \frac{1}{\sigma_{i_{width}}\sqrt{2\pi}} e^{\frac{-\left(width_{t_{n-1}} - \bar{x}_{i_{width}}\right)^2}{2\sigma_{i_{width}}^2}} \quad (3\text{-}1)$$

where $\bar{x}_{i_{width}}, \sigma_{i_{width}}$ are the average (mean) and standard deviation respectively of the width of workers in each scale. These parameters are defined in the following section. Every tracked target is assigned at time t_n to the scale that scored the highest probability. For each target, this is expressed as:

$$scale_{t_n} = \max_{i=1,2,3} P\left(scale_i | width_{t_{n-1}}\right) \quad (3\text{-}2)$$

After all tracked targets are grouped into scales, their walking speed values $\left\|\vec{v}_{(walk)}\right\|_{t_n}$ at time t_n are calculated by linearly interpolating the representative walking speed $\left\|\vec{v}_{(walk)}\right\|_i$ and width $width_i$ values of each scale $\{scale_i\}_{i=1,2,3}$. This interpolation is expressed as:

$$\left\|\vec{v}_{(walk)}\right\|_{t_n} = \left\|\vec{v}_{(walk)}\right\|_i + \frac{\left|\left(width_{t_{n-1}} - width_i\right)\left(\left\|\vec{v}_{(walk)}\right\|_{i+1} - \left\|\vec{v}_{(walk)}\right\|_i\right)\right|}{\left|width_{i+1} - width_i\right|} \quad (3\text{-}3)$$

The representative $\left\|\vec{v}_{(walk)}\right\|_i$ and $width_i$ values correspond to the highest cumulative probability. Secondly, the proposed prediction model uses $\left\|\vec{v}_{(walk)}\right\|_{t_n}$ to predict the maximum distance d_{xy} that every worker covers between two successive frames $\{t_{n-1}, t_n\}$ with time difference $dt = \frac{1}{\text{fps}}$ through the equation:

$$d_{xy} = \left\|\vec{v}_{(walk)}\right\|_{t_n} dt \quad (3\text{-}4)$$

This distance d_{xy} is the prior that is used in order to define the dimensions $\{W_S \times H_S\}$ of the search region R_s for each target. In the example presented in Figure 3-2 all corners of a target's tracking region R_T (cyan rectangle) with dimensions $\{W_T \times H_T\}$ at time t_n are expanded by a radius equal to d_{xy}. This way the proposed method captures all possible regions that a target might move towards in the following

frame at time t_n: a) left, b) right, c) upwards, d) downwards, and e) diagonally. The search region R_s of a target at time t_n given its previous position at time t_{n-1} is expressed by the following equations:

$$W_{S_{t_n}} = W_{T_{t_{n-1}}} + 2d_{xy} \quad (3\text{-}5)$$

$$H_{S_{t_n}} = H_{T_{t_{n-1}}} + 2d_{xy} \quad (3\text{-}6)$$

where $W_{S_{t_n}}, H_{S_{t_n}}$ is the width and height respectively of the search region R_S in current frame at time t_n, and $W_{T_{t_{n-1}}}, H_{T_{t_{n-1}}}$ is the width and height respectively of tracking region R_T in previous frame at time t_{n-1}.

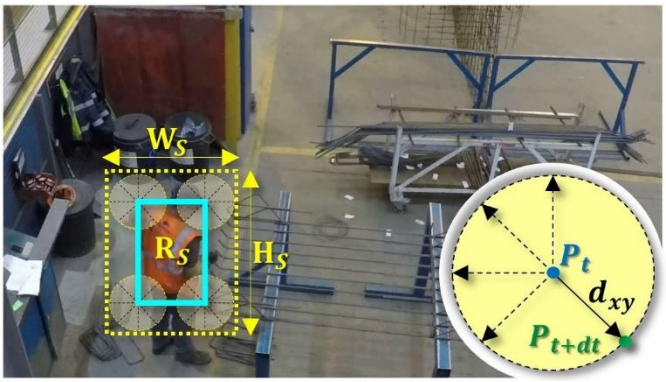

Figure 3-2**:** Prediction of the position of a tracked target in the following frame.

3.4.2. Appearance model

The proposed method applies the appearance model once the search region R_S is predicted for each tracked target. This model exploits the colour features of workers' high visibility apparel (Hi-Vis). Hence, only the part of a worker's body that is covered by Hi-Vis is tracked. The proposed appearance model uses the HSV colour space to extract these colour features. This is because it describes better both orange (a) and yellow (b) colours of Hi-Vis apparel as compared to the RGB colour space, as seen in Figure 3-3.

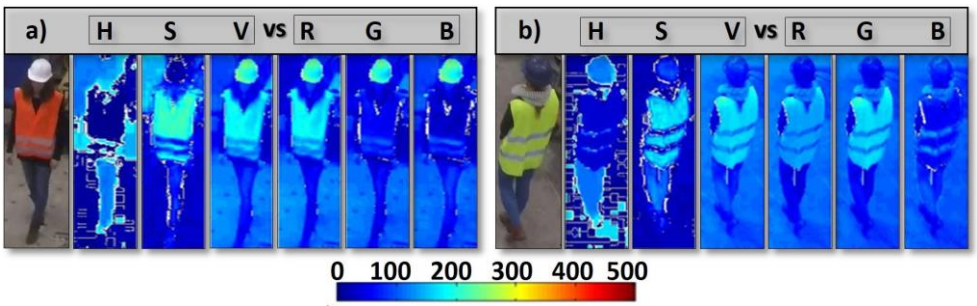

Figure 3-3: HSV vs RGB channels for (a) orange & (b) yellow Hi-Vis apparel.

The Hi-Vis apparel of workers commonly varies in terms of: a) colour combinations i.e. orange, yellow, b) type of apparel i.e. jacket, vest, uniform, c) design pattern of reflecting grey stripes, and d) faded colours i.e. dirty or worn out. Each target's body is segmented in order to alleviate the effect of such appearance variations. A grid divides the search region R_S of every target with dimensions $\{W_S \times H_S\}$ into smaller equally sized rectangular patches p_i with dimensions $\{\delta_x \times \delta_y\}$. The size of these patches p_i is experimentally defined in the following section 3.5.1. A support vector machine classifier (SVM) identifies the segmented patches p_i that belong to workers' Hi-Vis apparel. Figure 3-4 shows in (a) the segmentation and in (b) the classification processes.

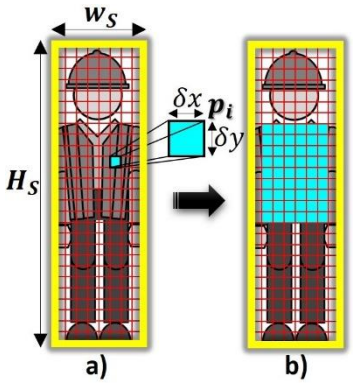

Figure 3-4: Proposed appearance model. (a) Segmentation of the search region R_S into rectangular patches $\{p_i\}_{i=1..N}$. (b) Classification of the segmented patches $\{p_i\}_{i=1..N}$ as part of tracked worker's Hi-Vis apparel.

Images of both colours (orange & yellow) of workers' Hi-Vis apparel are used as positive, whilst random background images are used as negative to train the SVM. In total, 688 images are used as training data. From them, 491 are positive and 197 are negative. The positive training data are labelled with +1 and the negative with -1. Figure 3-5 illustrates the training data.

Figure 3-5: Training data of a multi-colour SVM classifier.

The histograms of hue (actual colour) and saturation (purity of colour) colour channels are extracted for each of these images. The value (brightness) colour channel is discarded as it is sensitive to illumination variations. The histograms of hue and saturation have 360 and 100 values respectively. These values

are normalized since the SVM is not scale invariant. Hence, the feature vector $\vec{x}$ that describes each worker belongs to the two dimensional feature space ($\vec{x} \in R^2$) and is expressed as follows:

$$\vec{x} = \left\{ \begin{bmatrix} h_1 \\ \cdot \\ \cdot \\ h_{360} \end{bmatrix}^T , \begin{bmatrix} s_1 \\ \cdot \\ \cdot \\ s_{100} \end{bmatrix}^T \right\} \tag{3-7}$$

After the SVM is trained, it returns a hyperplane $f(\vec{x})$ that maps the segmented patches p_i from the two dimensional feature space (R^2) into a higher one through a Gaussian kernel:

$$f(\vec{x}) = \sum_1^N \alpha_i \, \vec{y}_i \exp\left(-\frac{1}{2\sigma^2} \|\vec{x} - \vec{x}_i\|^2\right) + b \tag{3-8}$$

where $\vec{x}_i$ are the values of the training input (support vectors), $\vec{y}_i$ are the binary labels {-1, +1} of the training input, $\vec{x}$ are the values of the input data, N is the size of the training data, σ is the standard deviation of the kernel, $f(\vec{x})$ are the binary labels {-1, +1} of the input data, and α_i are weights of the training input. The SVM performs cross-validation to calculate the parameters of the hyperplane $f(\vec{x})$ (i.e. α_i, b, σ). The training data are split into a total of K subsets for this purpose. One of these subsets is repetitively used for validation and the others for training. This repetitive validation minimizes the estimation error of the parameters. For our training data, cross validation converges to the same gamma variable ($\gamma = \frac{1}{2\sigma^2} = 0.5063$) for random values of subsets K (2, 5, 10). Therefore, the total number of subsets K is taken equal to the minimum (i.e. K=2) for computational efficiency. This hyperplane $f(\vec{x})$ linearly separates the segmented patches $\{p_i\}_{i=1..N}$ into two groups (see Figure 3-6). In this figure, any segmented patch p_i that falls above the hyperplane $f(\vec{x})$ is positively classified as part of a worker's Hi-Vis apparel (green points), whilst any pixel that falls below is negatively classified as part of the background (blue points). Finally, all positively classified segmented patches $\{p_i\}_{i=1..N}$ of a search region R_S are clustered within a bounding box that depicts the final tracking output R_T of a target.

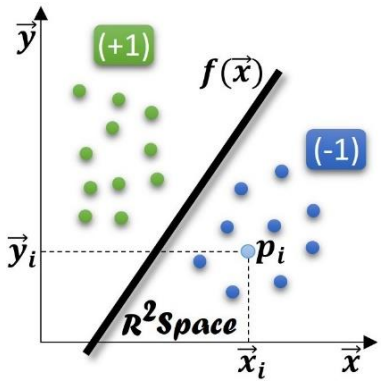

Figure 3-6: Linear classification of colour image patches with an SVM.

Every search region R_S should contain only one worker's positive classified segmented patches p_i as mentioned above. However, this is difficult in congested working environments such as construction sites, and the search regions R_S of close by workers may overlap as a result. The proposed appearance model uses templates instead of the colour features of Hi-Vis apparel when this overlapping occurs to track workers. This is because the spatial arrangement of colours as provided by templates is more representative than colour histograms especially when congested targets share similar appearance e.g. workers in the same colour of Hi-Vis apparel. The normalized cross correlation (NCC) is implemented in order to compare the templates of congested workers between consecutive frames. The main disadvantage of such template comparison methods is that they are not invariant to posture, scale and appearance variations. The proposed method re-activates tracking via colour features when targets are not occluded any more. The NCC compares the search regions R_S of the congested workers within a frame at time t_n with all possible matching candidates from a previous frame at time t_{n-k}. These candidates are the tracking regions R_T of the workers at the time instance that congestion appeared. They depict each worker's appearance as long as congestion lasts. This relies on the previous (b) assumption of section 3.3 that a worker's appearance does not change during congestion. The implementation of the NCC is expressed as follows:

$$P_{x,y} = \frac{\sum_{x\prime,y\prime}(T(x\prime,y\prime)*I(x+x\prime,y+y\prime))}{\sqrt{\sum_{x\prime,y\prime} T(x\prime,y\prime)^2 * \sum_{x\prime,y\prime} I\,(x+x\prime,y+y\prime)^2}} \tag{3-9}$$

where the regions R_S are the source images $I(x)$, and the tracking regions R_T are the template images $T(x)$. The above equation is used to scan each template image $T(x)$ along all available source images $I(x)$. It finally returns the position $P_{x,y}$ of each congested worker in the frame at time t_n when a positive match is confirmed between a source $I(x)$ and a template $T(x)$ image (see Figure 3-7).

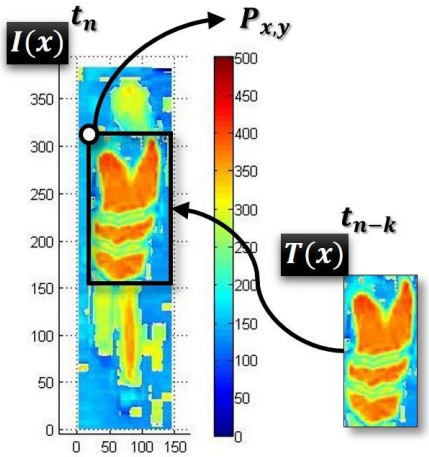

Figure 3-7: Template matching of the same worker between non-successive frames.

3.4.3. Filtering model

This section presents how the proposed filtering model filters the segmented patches $\{p_i\}_{i=1..N}$ that are wrongly classified as positive by the appearance model of section 3.4.2. Then, the filtered patches $\{p'_i\}_{i=1..N}$ are clustered into a bounding box to return the corrected tracked region R_T. We use the term outliers for these patches for the remainder of this chapter for simplicity. These outliers depict either noise or objects that belong to the background and share a similar colour to the Hi-Vis apparel. Figure 3-8 summarizes the proposed filtering model. Firstly, the outliers are detected. Secondly, the filtering model discards them and thirdly, the corrected tracked region is extracted R_T.

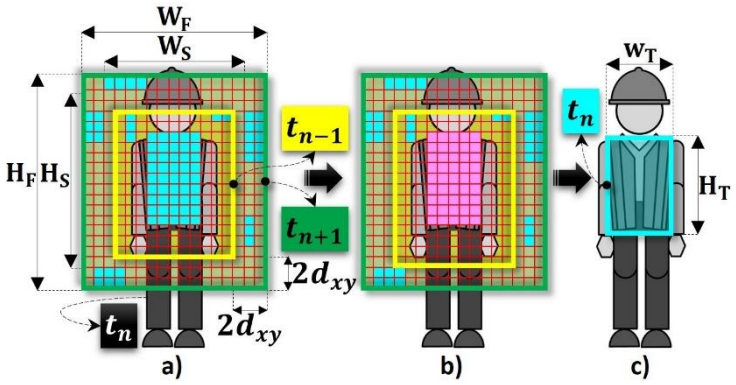

Figure 3-8: Proposed filtering method. (a) Detection of outliers within a filter region R_F (green large rectangle). (b) Filtering of outliers (cyan rectangles). (c) Clustering of positively classified patches (pink rectangles) into the final tracking region R_T (cyan rectangle).

The filtering model enhances the performance of the appearance model. This is because the latter calculates the tracking region R_T by fitting all positively classified patches $\{p_i\}_{i=1..N}$ of a search region R_S into a bounding box. However, if some of these patches are either wrongly classified as part of a worker's body or correctly classified but belong to the background, then the prediction model will fail to track the same worker across successive frames. Figure 3-9 illustrates such an example. In this figure, the proposed prediction model fails to track a worker while passing in front of an object of a similar colour (Hi-Vis jacket hanging on the wall).

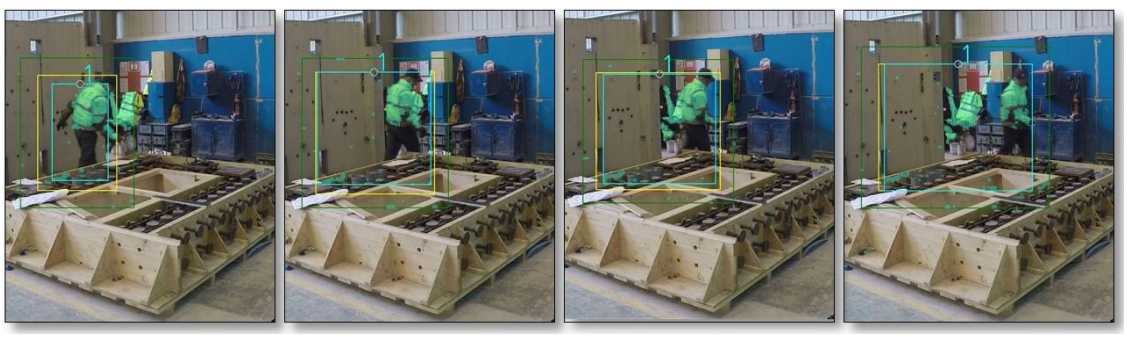

Figure 3-9: Failure of the proposed prediction model under the appearance of outliers.

The proposed filtering model exploits target motion contours to filter outliers. These contours are extracted by frame differencing each search region R_S between two successive frames (t_n, t_{n-1}):

$$\Delta RGB_{(x,y)_{t_n}} = RGB_{(x,y)_{t_n}} - RGB_{(x,y)_{t_{n-1}}} = \frac{R_{(x,y)_{t_n}} + G_{(x,y)_{t_n}} + B_{(x,y)_{t_n}} - \left(R_{(x,y)_{t_{n-1}}} + G_{(x,y)_{t_{n-1}}} + B_{(x,y)_{t_{n-1}}}\right)}{3} \quad (3\text{-}10)$$

where $RGB_{(x,y)_{t_n}}$ is the RGB value of a pixel at position (x, y) in the image. The pixels that belong to the foreground are labelled as "1" value and the pixels that belong to the background as "0". Then, a matrix $M_{(m,n)}$ is populated with the binary {0, 1} results of frame differencing (see Figure 3-10). This matrix $M_{(m,n)}$ has dimensions $\{m, n\}$ equal to the height and width respectively of each target's search region R_S. Hence, each cell within the matrix $M_{(m,n)}$ corresponds to a pixel within R_S.

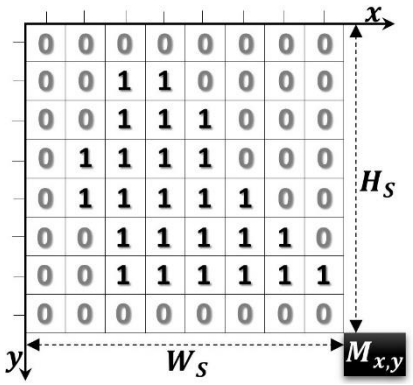

Figure 3-10: Motion-based classification of pixels.

This binary labelling depends on a threshold value ($thres$) that filters out random noise and is expressed as follows:

$$M_{x,y} = \begin{cases} 1, & \Delta RGB_{x,y_{t_n}} > thres \\ 0, & \Delta RGB_{x,y_{t_n}} \leq thres \end{cases} \quad (3\text{-}11)$$

Noise appears due to systematics such as lighting conditions, hardware, and camera mounting. A range of values has been tested in order to experimentally define this parameter given the specific experimental set up conditions described in section 2.4. Figure 3-11 shows how the proposed method removes noise.

Figure 3-11: Example of frame differencing between successive frames for motion extraction. (a) Noisy outcome. (b) Denoised outcome.

The foreground pixels are categorised into four groups. The C_L group that contains the far left foreground pixels:

$$C_L = \left\{ \forall\, (x_i, y_i)_{p_i} \in R_S : M_{x_i,y_i} = 1 | \operatorname*{argmin}_x \right\} \tag{3-12}$$

the C_R group that contains the far right foreground pixels:

$$C_R = \left\{ \forall\, (x_i, y_i)_{p_i} \in R_S : M_{x_i,y_i} = 1 | \operatorname*{argmax}_x \right\} \tag{3-13}$$

the C_T group that contains the top foreground pixels:

$$C_T = \left\{ \forall\, (x_i, y_i)_{p_i} \in R_S : M_{x_i,y_i} = 1 | \operatorname*{argmin}_y \right\} \tag{3-14}$$

and lastly the C_B that contains the bottom foreground pixels:

$$C_B = \left\{ \forall\, (x_i, y_i)_{p_i} \in R_S : M_{x_i,y_i} = 1 | \operatorname*{argmax}_y \right\} \tag{3-15}$$

where x_i, y_i = image coordinates of a pixel i, $(x_i, y_i)_{p_i}$ = position of a patch p_i in an image, and M_{x_i,y_i} = motion label of each patch (p_i).

Only the foreground pixels that overlap with the image coordinates of positively classified segmented patches $\{p_i\}_{i=1..N}$ are returned as pixels of the motion contour R_C. This pairing (see (a) in Figure 3-12) ensures that all pixels of the motion contour R_C positively belong to a tracked target rather to random noise. Therefore, the motion contour R_C of each target defines a region which is expressed as follows:

$$R_C = \{\forall\,(x_i, y_i)_{p_i} \geq C_L\} \cap \{\forall\,(x_i, y_i)_{p_i} \leq C_R\} \cap \{\forall\,(x_i, y_i)_{p_i} \leq C_T\} \cap \{\forall\,(x_i, y_i)_{p_i} \geq C_B\} \quad (3\text{-}16)$$

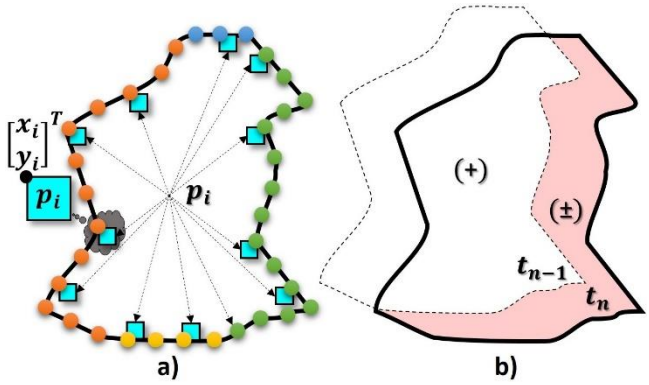

Figure 3-12: Motion contour for filtering outliers. (a) Categorisation of motion contour's pixels into four groups (orange: left, green: right, blue: top, yellow: bottom). (b) Calculation of potential outlier regions (red shaded) between successive frames (t_{n-1} = dotted shaped contour, t_n = continuous shaped contour).

Figure 3-13 illustrates with white lines four examples of motion contours of construction workers as they walk (a-b), stretch (c) or bend (d). These contours correspond only to the part of a worker's body that is covered by Hi-Vis apparel.

Figure 3-13: Motion contour while target walks (a-b), partially moves (c), and bends (d).

Then, the motion contours R_C of the same target across successive frames are compared in order to locate the regions that most likely (±) enclose potential outliers. In (b) of Figure 3-12, the continuous shaped motion contour of a target at time t_n is compared spatially to the dotted shaped motion contour of the same target at previous time t_{n-1}. The searching for outliers takes place only within the non-overlapping region R_N of the motion contour at t_n (red shaded region in (b) of above Figure 3-12). The R_N is equal to the image area that a moving target occupies within a period of time $dt = t_n - t_{n-1}$ and is expressed as follows:

$$R_N = R_C | t_n - CC \quad (3\text{-}17)$$

where CC is the overlapping region of the motion contours of a target between two successive frames (t_n, t_{n-1}) and is expressed as follows:

$$CC = \{\forall\ (x_i, y_i)_{p_i} \in R_C | t_n\} \cap \{\forall\ (x_i, y_i)_{p_i} \in R_C | t_{n-1}\} \tag{3-18}$$

The region CC is positively (+) classified as part of a tracked target's body as it only encloses segmented patches p_i that have already been classified as positive in the previous frame at time t_{n-1}. Finally, the filtering model discards all outliers within region R_N and all positively patches p_i that are not enclosed by the motion contour at time t_n:

$$\{p'_i\}_{i=1\ldots N} = \left\{\forall\ p_i \in\ R_N : \Delta RGB_{(x_i,y_i)_{p_i}} \geq thres\right\} \cup\ \{\ AND\ \forall\ p_i \in CC\} \tag{3-19}$$

This way the proposed filtering method keeps only the patches p_i that are enclosed within the motion contour of a target. The filtering method is activated only if potential outliers are detected within a small distance from the target. This selective activation ensures computational efficiency. The prediction model calculates at time t_{n-1} the search region R_S of a tracked target at time t_n (see section 3.4.1). Therefore, all outliers that are located outside R_S are at least one time interval away from the target at time t_n and two time intervals away from the target at time t_{n-1}. These are not assumptions. They are derived from the proposed prediction model. Given this, all corners of each search region R_S at time t_{n-1} are expanded by a distance equal to two times d_{xy} (see section 3.4.1) to calculate the dimensions of a filter region R_F $\{W_{F_{t_n}}, H_{F_{t_n}}\}$. The filter region R_F of a target at time t_n given its previous position at time t_{n-1} is expressed by the following equations:

$$W_{F_{t_n}} =\ W_{S_{t_{n-1}}} + 4d_{xy} \tag{3-20}$$

$$H_{F_{t_n}} =\ H_{S_{t_{n-1}}} + 4d_{xy} \tag{3-21}$$

The role of R_F is to activate and deactivate the proposed filtering model. It activates the filtering model only if the target moves towards the outliers. This section proposes a method that classifies a tracked target as either moving or stationary to achieve this. Motion is detected by comparing the RGB values within a target's search region R_S. This relies on the fact that the sub-regions towards which the target does not move retain their RGB values across successive frames. A moving target might move towards at least eight possible sub-regions $\{j\}_{1\ldots 8}$ between two successive frames as seen in (a) of Figure 3-14.

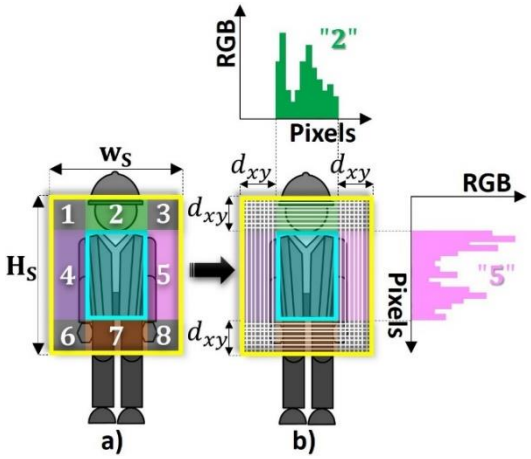

Figure 3-14: Classification of tracked target as moving or stationary. (a) Potential future directions of tracked target. (b) Layers of RGB intensity values.

Each of these sub-regions $\{j\}_{1\dots8}$ is split into layers of RGB intensity values to detect motion. These layers are set parallel to the biggest direction of each sub-region so that their total number is the same and equal to d_{xy} (see (b) in Figure 3-14). The Pearson's correlation coefficient is used to compare the RGB intensity similarity of the same layer l_i across two successive frames t_{n-1} and t_n. The coefficient of each layer l_i at time t_n within a sub-region j is expressed as follows:

$$r_{l_{i_{t_n}}} = \frac{\sum_{k=1}^{N}(RGB_{(x_k,y_k)_{t_{n-1}}} - \overline{RGB}_{(x,y)_{t_{n-1}}})(RGB_{(x_k,y_k)_{t_n}} - \overline{RGB}_{(x,y)_{t_n}})}{\sqrt{\sum_{k=1}^{N}(RGB_{(x_k,y_k)_{t_{n-1}}} - \overline{RGB}_{(x,y)_{t_{n-1}}})^2}\sqrt{\sum_{k=1}^{N}(RGB_{(x_k,y_k)_{t_n}} - \overline{RGB}_{(x,y)_{t_n}})^2}} \quad (3\text{-}22)$$

where $RGB_{(x_k,y_k)_{t_n}}$ is the RGB intensity value of a pixel "k" at position $(x_k,\ y_k)$ along a layer l_i at time t_n, $\overline{RGB}_{(x,y)_{t_n}}$ is the average intensity RGB value of all pixels along a layer l_i at time t_n, and N is the total number of pixels along a layer.

The Pearson's correlation coefficient $r_{l_{i_{t_n}}}$ is equal to 1 only when all compared layers of a sub-region j at time t_n have not changed with respect to the previous frame at time t_{n-1}. Therefore, a target most likely moves towards a sub-region j if all layers return $r_{l_{i_{t_n}}} < 1$. Given this, the probability that a target moves towards a sub-region j is expressed as follows:

$$P(j) = \frac{\sum_{i=1}^{d_{xy}} r_{l_{i_{t_n}}}}{d_{xy}} \quad (3\text{-}23)$$

The proposed method classifies a target as moving if $P(j)$ is higher than 99% for any of the sub-regions $\{j\}_{1\dots8}$. The filtering model combines the motion of targets and the appearance of outliers to activate the filtering model. These form two criteria which are combined under a scoring system. The

motion of a target is the first criterion and the appearance of outliers is the second. One point is added (+) to the total score if both criteria are true whilst one point is removed (-) when the first criterion is false. A total score equal to two activates the filtering model. This allows the proposed tracking method to identify all outliers that lie within two frame time $dt = t_{n+1} - t_{n-1}$ distance. Equally, two sequential false results of the first criterion deactivate the filtering model and set the total score to zero. Figure 3-15 illustrates how the proposed filtering model improved the performance of the prediction model under the appearance of outliers of previous Figure 3-9. In Figure 3-15, the filter is activated after (a) and deactivated after (c) as the worker moves away from the outliers.

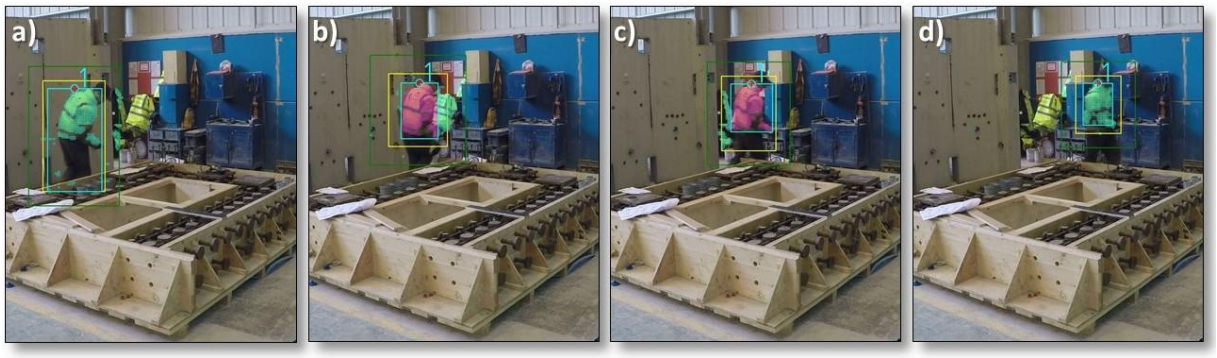

Figure 3-15: Activation of proposed filtering model in (a) and deactivation in (d). The filtered patches $\{p'_i\}_{i=1..N}$ are coloured pink in (b-c).

3.4.4. Adaptive Model

The proposed tracking method continuously updates a feature vector $\vec{X}$ to achieve long term tracking of multiple workers under all of the most common tracking challenges within a construction jobsite. Such challenges are background clutter, posture, appearance and scale variations, and changes in scene illumination, abrupt movement, occlusions, and congestion. Figure 3-16 illustrates an example of how the proposed tracking output (cyan rectangle) adapts to some of these challenges

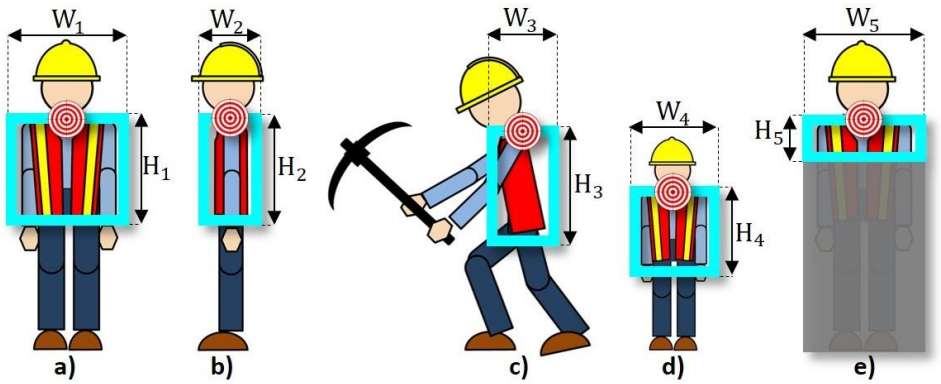

Figure 3-16: Workers' appearance variations due to posture (a-c), scale (d) and occlusion (e).

The feature vector $\vec{X}$ contains for each target at time t_n the following features: a) dimensions $\vec{x}_{R_T}$ of tracking region R_T, b) position $\vec{x}_P$ in the frame, c) RGB values $\vec{x}_{RGB_{R_F}}$ of all pixels within the filter region R_F, d) motion contour $\vec{x}_{R_C}$, e) template $\vec{x}_{Tem_{R_T}}$ of tracking region R_T, and lastly f) vector $\vec{x}_{HS}$ that contains the hue and saturation values of patches p_i within the search region R_S. Table 3-1 summarizes which features are used to overcome each of the aforementioned tracking issues.

Table 3-1: Proposed features for overcoming most common tracking challenges within the complex environment of a construction jobsite.

	$\vec{x}_{R_T}$	$\vec{x}_P$	$\vec{x}_{RGB_{R_F}}$	$\vec{x}_{R_C}$	$\vec{x}_{Tem_{R_T}}$	$\vec{x}_{HS}$
1.Background clutter	✓	✓	✓	✓	✘	✓
2.Occlusions	✓	✓	✘	✘	✘	✓
3.Posture variations	✓	✓	✘	✘	✘	✓
4.Scale variations	✓	✓	✘	✘	✘	✓
5.Abrupt movement	✓	✓	✘	✘	✘	✓
6.Congestion	✓	✓	✘	✘	✓	✘
7.Scene illuminations	✓	✓	✘	✘	✘	✓
8.Appearance Similarity	✓	✓	✘	✘	✘	✘

✓*:satisfied,* ✘*: unsatisfied*

3.5. Experiments and results

The evaluation video samples were collected from real construction sites both indoors and outdoors. The indoor video samples depict workers performing pipe installation and steel fixing. The outdoor video samples capture workers involved in a concrete pouring task. The indoor sample was recorded at 30 fps with a GoPro camera (see Section 2.4.1), whilst the outdoor sample was captured at 5fps with a Logitech web camera. Both cameras were fixed at a large height in order to simulate the setup of a jobsite's surveillance camera system.

Two measurement metrics as defined by Čehovin et al. (2015) are used in order to evaluate the performance of the proposed tracking method. The first is a distance error d_{error} metric. It is equal to the Euclidean distance (see Equation 3-24) between the reference points that represent the tracker's output P_T and the ground truth P_G (see Figure 3-17). The proposed method uses as a reference tracking point, the upper middle point of the tracker's output region R_T. Such a reference point better represents a moving target in comparison with the centre of the bounding box that is commonly used by most tracking methods. This is because the latter depends on the height of the tracked object which however resizes abruptly when occlusions appear. In construction sites, the lower part of workers' bodies is often occluded by work benches, structural elements and equipment.

$$d_{error} = \sqrt{(x_{P_T} - x_{P_G})^2 + (y_{P_T} - y_{P_G})^2} \quad (3\text{-}24)$$

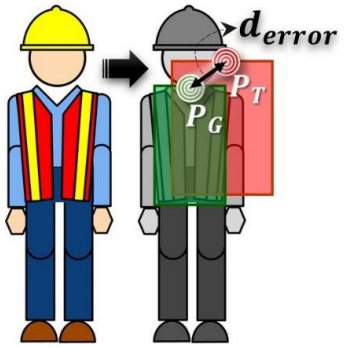

Figure 3-17: Distance error performance metric between the tracking output (red shaded region) and the ground truth (greed shaded region).

The second metric is the F-measure. It returns the spatial overlap between the region of the tracker R_T and the region of the ground truth R_{GT} :

$$F = \frac{R_{GT} \cap R_T}{R_{GT} \cup R_T} = \frac{TP}{TP+FN+FP} \quad (3\text{-}25)$$

where true positive TP depicts the correctly tracked region:

$$TP = R_{GT} \cap R_T \quad (3\text{-}26)$$

false positive FP depicts the incorrectly tracked region:

$$FP = R_T - (R_{GT} \cap R_T) \quad (3\text{-}27)$$

and false negative FN depicts the incorrectly not tracked region:

$$FN = R_{GT} - (R_{GT} \cap R_T) \quad (3\text{-}28)$$

Figure 3-18 illustrates two examples of the F metric. In the first example (a) in Figure 3-18, the bounding boxes of the tracker R_T (red shaded) and the ground truth R_{GT} (green shaded) overlap partially. The true positive TP value is equal to this overlapping region. In the second example (b) in Figure 3-18, where no overlapping region ($TP = \emptyset$) exists between the tracking output R_T and the ground truth R_{GT}, the F-measure metric returns zero. The ground truth for the evaluation of the proposed vision-based 2D tracking method is manually extracted. A bounding box that captures the part of each

worker's body that is covered by safety apparel (Hi-Vis) is fitted in each frame (recall the appearance model relies only on such colour features).

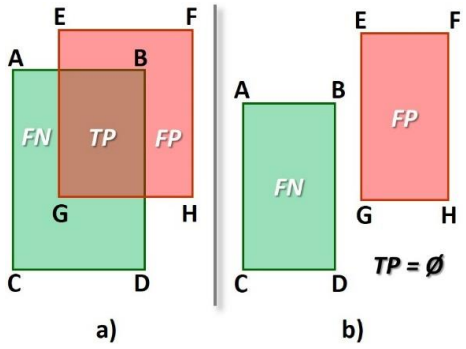

Figure 3-18: F-measure performance metric of partial (a) and zero (b) overlapping between the tracked output (red shaded region) and the ground truth (green shaded region).

3.5.1. Definition of parameters

This section determines experimentally the parameters of the proposed computer vision-based 2D tracking method of construction workers.

2.5.1.1 Definition of scale parameters

The proposed prediction model must be calibrated to the scale of the tracked targets. As mentioned in section 5.4.1 three scales are used to achieve this. A data sample of width and walking speed values is collected for every scale by manually tracking 3 workers while walking across 40 successive frames. The goodness of fit of a Gaussian cumulative distribution function to the manually collected width and walking speed values of each scale is illustrated in Figure 3-19. The walking speed and width values with a cumulative probability equal to 1.00 are selected as representative of each scale as this entails that the probability for a worker to have either a width or a walking speed higher than the representative is 0%.

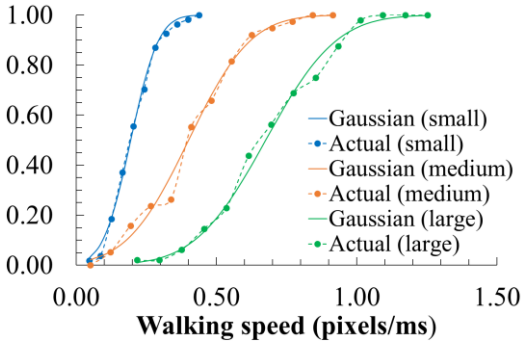

Figure 3-19: Cumulative probability of tracked targets' walking speed per scale.

The quality of the manually collected width and speed values is evaluated through true positives TP and false positives FP. If a worker is correctly captured within the predicted region R_s, then it is a TP prediction. If not, then it is a FP prediction. The interpolation Equation 3-3 as described in section 3.4.1 is not used during this evaluation. The regions of workers in following frames are predicted only through the speed values of the scale they are manually assigned during this evaluation. The sample size is defined by (Eng, 2003):

$$n = \frac{4{Z_{crit}}^2 p(1-p)}{D^2} \tag{3-29}$$

where Z_{crit} is the confidence level based on normal distribution, D is the limit of error and p is an estimate of the accuracy of the test. For a confidence level of 95%, a limit of error (D) equal to ±5% and a pessimistic estimation of the expected accuracy ($p = 50\%$), the minimum sample size to be tested for each scale is equal to 384. For the large scale 393 frames were tested, for the medium scale 388, and for the small scale 389 (see Figure 3-20). This evaluation returns 0 FP results for all three scales. Therefore, the proposed method uses the values of Table 3-2 to predict the regions that contain workers in the following frames.

Figure 3-20: Screenshots of the evaluation of the proposed prediction model (yellows rectangles) given the representative speed values of scale small (a), scale medium (b), and scale large (c).

Table 3-2: Average width and walking speed values of workers per scale.

Scale (i)	Width (pixels)		Walking speed (pixels/ms)	
	$\bar{x}_{i_{width}}$	$\sigma_{i_{width}}$	$\bar{x}_{i_{walk}}$	$\sigma_{i_{walk}}$

1. Small	55.51	7.11	0.19	0.08
2. Medium	133.26	6.72	0.40	0.17
3. Large	243.22	20.52	0.68	0.20

2.5.1.2 Definition of segmentation parameters

The optimum values for the size (width, height) of the segmented patches p_i are experimentally defined. The proposed method is tested under the most challenging appearance variations to achieve this. Such variations in workers' appearance are: a) small scale, b) open vest and c) worn out or dirty apparel (see Figure 3-21). This figure illustrates the ground truth GT with dark blue and the classified patches, p_i with cyan, for three different sizes of the segmentation grid (1x1, 5x5, and 10x10).

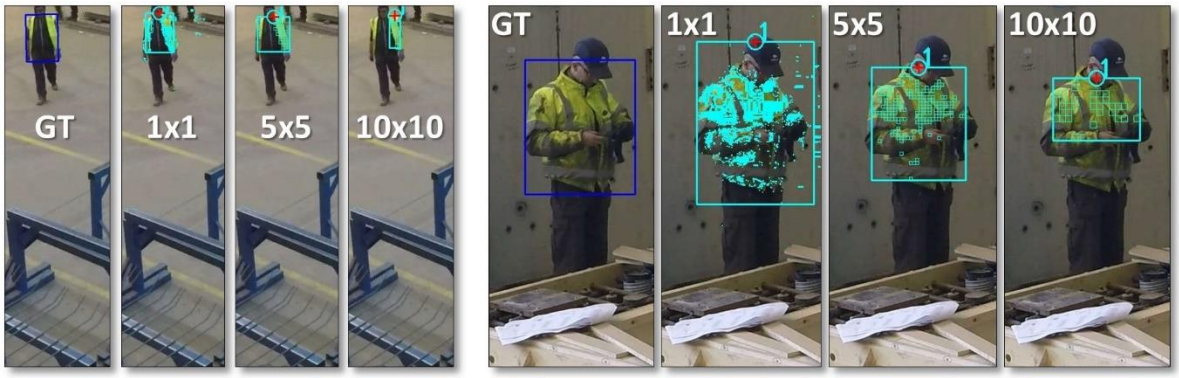

Figure 3-21: Examples of tracking performance for different sizes of the segmentation grid under small scale (left) and dirty apparel (right).

Table 3-3 illustrates the tracking performance of the proposed method based on the F-measure metric for different sizes of patches p_i that range from 1x1 to 10x10 pixels values for a small scale worker with an open vest and a worker with dirty Hi-Vis apparel.

Table 3-3: Proposed method's tracking performance under different segmentation

	Small scale & Open Vest			**Worn out Colours**		
Size p_i	**Region R_T**	**F-measure**	**Processing Time (msec)**	**Region R_T**	**F-measure**	**Processing Time (msec)**
1x1	52x73	0.737	1623	186x258	0.774	8254
2x2	50x72	0.719	494	186x214	0.839	1718
3x3	66x81	0.699	264	159x192	0.806	724
4x4	60x64	0.635	202	160x204	**0.808**	455
5x5	60x65	**0.757**	**134**	145x175	0.713	**286**
6x6	60x66	0.741	115	138x114	0.415	204
7x7	56x63	0.708	107	133x105	0.369	174
8x8	56x56	0.709	97	128x112	0.378	125
9x9	27x63	0.323	91	117x99	0.306	111
10x10	20x50	0.280	93	140x100	0.37	96

The smallest size (1x1) is the worst performer in both datasets in terms of processing time. It also returns false positive classified patches p_i. On the other hand, the largest (10x10) is the fastest but exhibits the worst performance. It appears that for a small scale (GT = 65x66) and an open jacket the best performance is achieved for a segmentation grid of 5x5, whereas for a larger scale (GT = 174x214) but worn out apparel a 4x4 grid gives the best performance. Ultimately, the filter region R_F is segmented into patches p_i of 5x5 in order to increase the proposed method's computational efficiency considering that multiple workers will be tracked simultaneously.

3.5.2. Quantitative Evaluation

In this section, the tracking method outlined in this chapter is compared to the method in Park and Brilakis (2016). The Park and Brilakis (2016) study (here after referred to as Park2016) is selected for the following two reasons: a) it combines the advantages of both a detector-based and an online-based tracking method, and b) is the most recent study in the literature on visual tracking of multiple construction workers. Both methods are tested under all of the aforementioned challenges that most commonly appear while tracking workers. Seven video samples were collected for each of these challenges separately (1-7) and one video sample (8) that combines several challenges together. These video samples are described in detail in Table 3-4.

Table 3-4: Evaluation video samples.

Video Sample	#Frames	Challenges
1.Abrupt movement	180	1 worker jumps in formwork
2.Background clutter	179	1 worker passes through hanging Hi-Vis jackets
3.Scene illuminations	149	1 worker in a sunny jobsite
4.Posture variations	151	2 workers bend and rotate
5.Scale variations	96	1 worker captured at small scale
6.Occlusions		
i) by wall	466	1 worker occluded almost totally ≥ 86%
ii) by steel bars (A)	114	1 worker ("0-2") occluded totally and 1 worker ("1-3") partially by 35%
iii) by ladder	212	1 worker occluded almost totally ≥ 90%
iv) by steel bars (B)	127	1 worker occluded ("0-2") by 67% and 1 worker ("1-3") by 62%
7.Congestion	186	3 workers share similar appearance, 4 overlap
8. Combined	219	9 workers exhibit congestion, occlusions, background clutter, similar appearance and scale variations

The proposed visual tracking method is quantitatively evaluated based on both the distance error d_{error} metric and the spatial overlap F-measure metric. The higher the values of the F-measure the better the tracking performance, and the smaller the distance error d_{error} the better the accuracy of the extracted trajectories. The former reflects the tracking error of the extracted trajectories whilst the latter reflects

how efficiently the tracking method's output encloses the figure of the target. The graphs in Figure 3-22 illustrate the performance per frame of our method (red) and Park2016 (green) under: a) abrupt movement, b) background clutter, c) illumination, d) posture variations, and e) scale variations. The results for both metrics are normalized to 1.

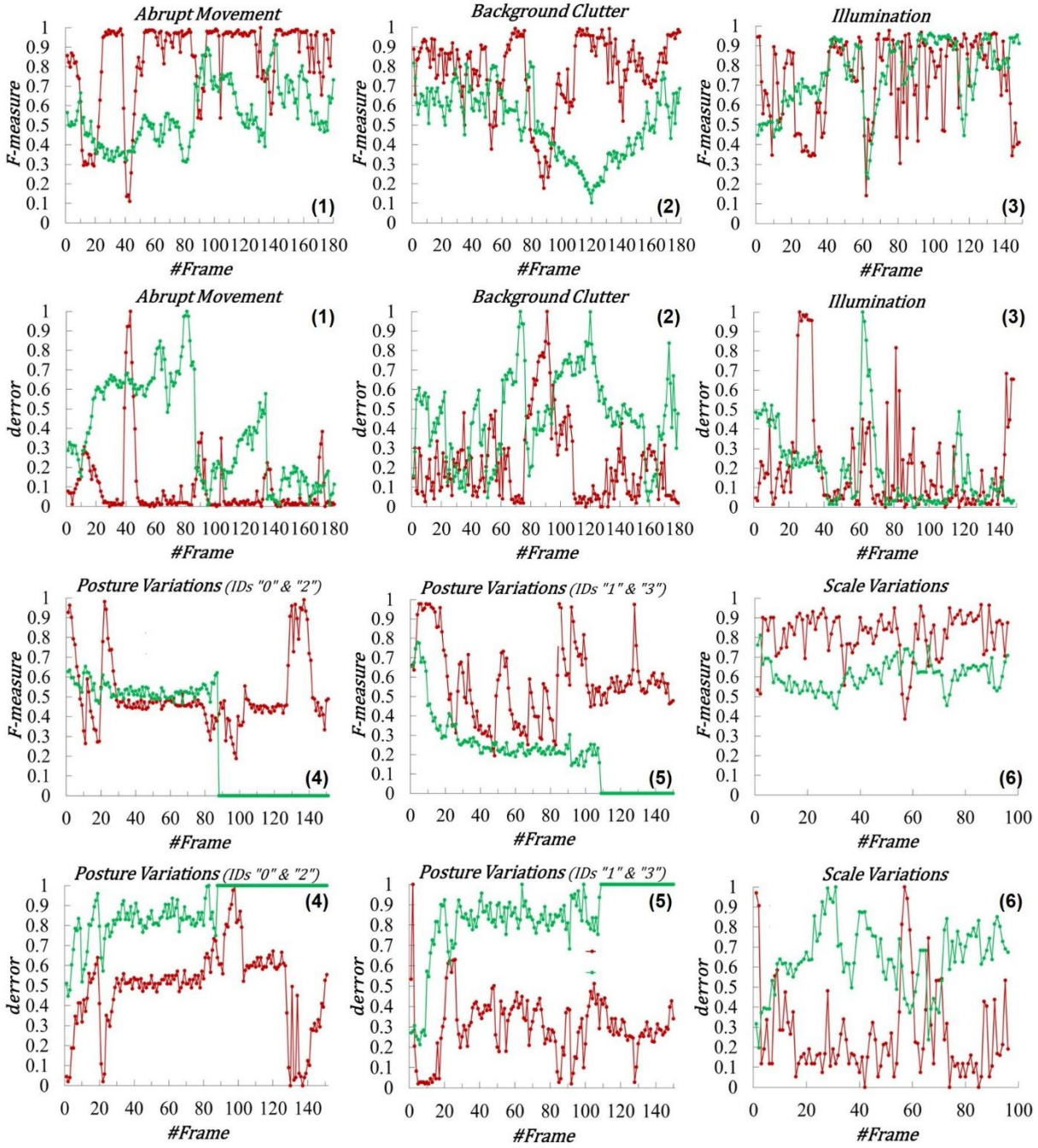

Figure 3-22: Comparison results of this chapter's proposed visual tracking method (red) vs Park2016 (green) under: a) abrupt movement, b) background clutter, c) illumination, d) posture variations, and e) scale variations.

Our method outperforms Park2016 in all cases except that of illumination variation. This is because the former resizes in a more abrupt way as compared to the latter. For the posture variation case, our method

is better than Park2016 in terms of both performance and long term tracking. This is due to the termination criteria of the latter that are activated when the detector loses the tracker. The bending posture of both workers in this video sample caused the termination of tracking of Park2016 as no detector appeared for a period of 3 seconds. Figure 3-23 contains screenshots of the performance of both methods. In this figure, the cyan coloured rectangles depict our method's output whilst the dark blue coloured rectangles the output of Park2016.

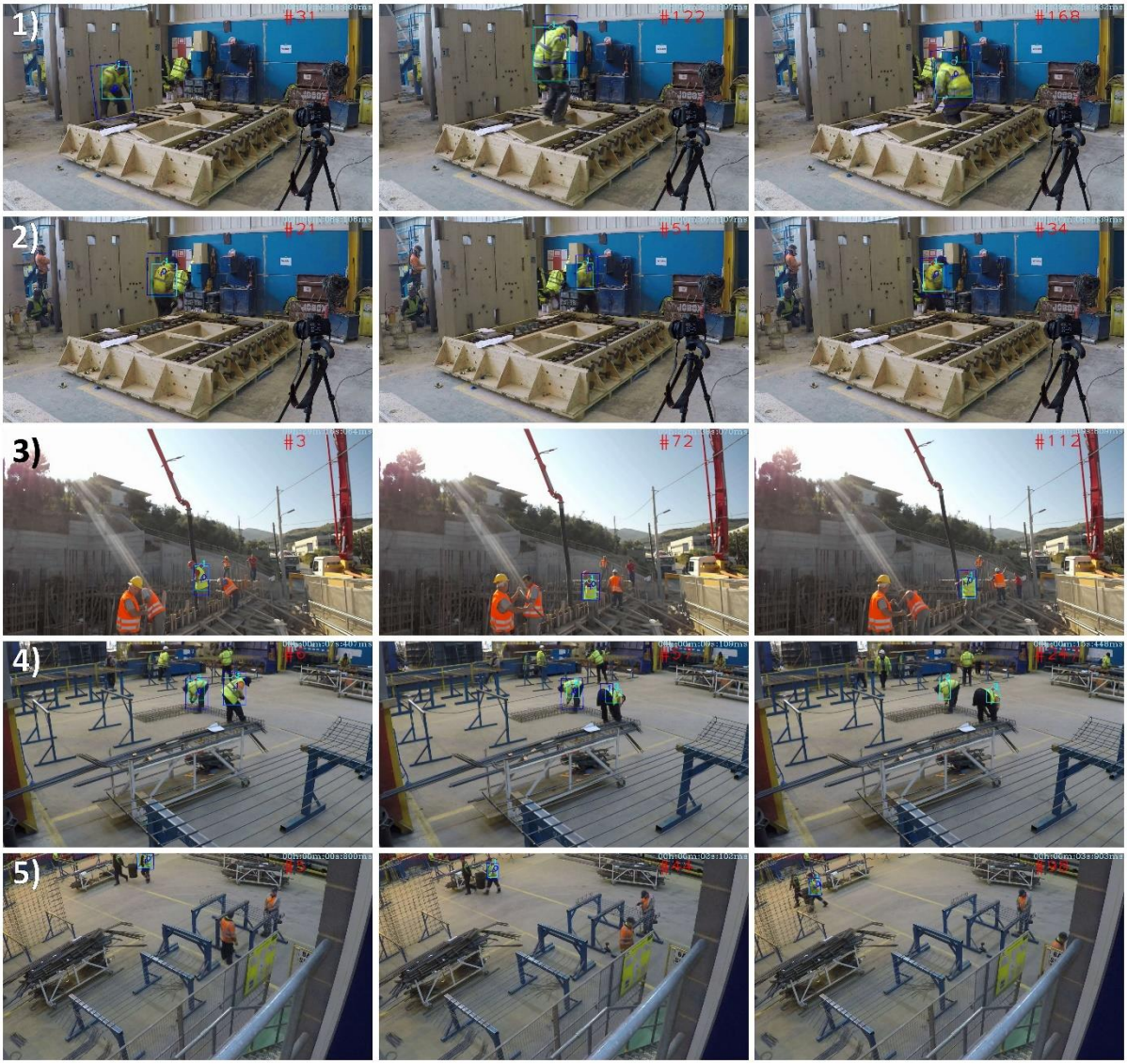

Figure 3-23: Screenshots of the performance of the proposed tracking method (cyan coloured rectangle) and Park2016 (dark blue coloured rectangle) under: 1) abrupt movement, 2) background clutter, 3) illumination, 4) posture variations, and 5) scale variations.

The graphs in Figure 3-24 depict the performances of the proposed visual tracking method and Park2016 under partial and severe occlusions. An occlusion is defined as partial when a worker's body is hidden by less than 70% and severe when it exceeds this percentage. The F-measure graphs show that this

chapter's proposed method resizes in a more abrupt way as compared to Park2016. However, it still manages to maintain tracking in all tested occlusion cases in contrast to Park2016 that terminates, due to failure to identify workers under occlusion. Park2016 appears to perform better only in one case. This is when the worker "1-3" in 6.ii video sample is partially occluded (35%) by steel bars.

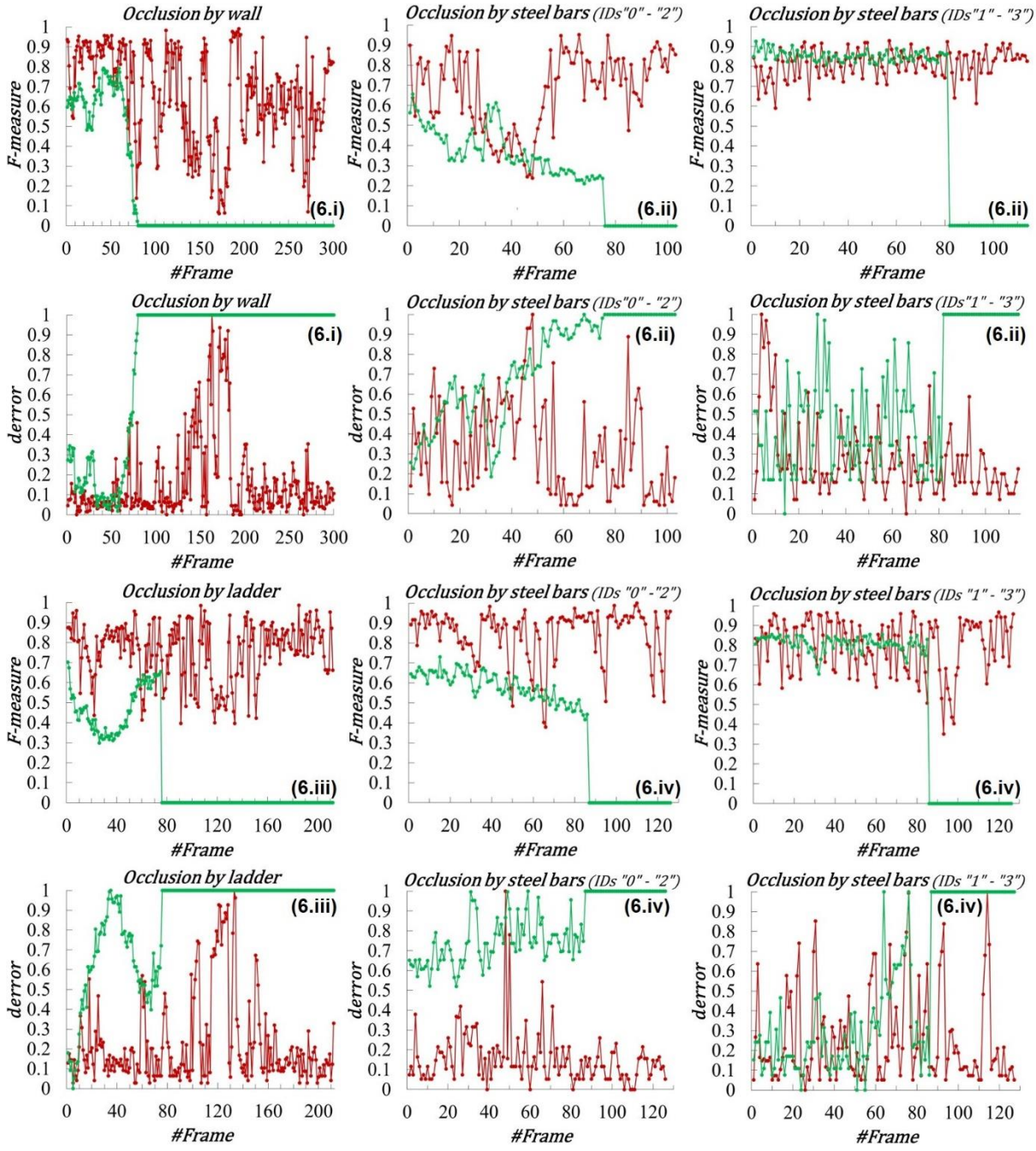

Figure 3-24: Comparison results of this chapter's visual tracking method (red) vs Park2016 (green) under: a) total occlusion by a wall, b) total occlusion by steel bars for worker with IDs "0"-"2", b) partial occlusion of worker with IDs "1"-"3" by steel bars, and c) total occlusion of worker by ladder.

Figure 3-25 illustrates representative screenshots of the performance of this chapter's method and Park2016 under different occlusion levels. The video sample 6.i contains the most challenging case of

occlusion as a worker gets almost totally hidden by a wall. Still, our method preserves tracking as the worker moves two times through this wall in order to return back to his work bench. Park2016 loses the target in an early stage as seen in 6.i of Figure 3-25. As a result it terminates after 3 seconds.

Figure 3-25: Screenshots of the performance of the proposed tracking method (cyan coloured rectangle) and Park2016 (dark blue coloured rectangle) under almost total occlusions (worker in 6.i, worker "0-2" with orange Hi-Vis in 6.ii, and worker in 6.iii) or by partial occlusions (worker "1-3" with yellow Hi-Vis in 6.ii, and both workers in 6.iv).

Table 3-5 summarizes the average F-measure and average distance error d_{error} metrics of all the above challenges. It highlights in bold the best performances achieved. Our visual tracking method outperforms Park2016 in terms of spatial overlap as described by the F-measure metric under occlusions, abrupt movement, background clutter, posture and scale variations. It performs worse by 2.86% only under illumination. The average distance error d_{error} of our method is, however, worse under both scene illumination changes and under partial occlusions (worker "0-2" in 6.ii and worker "1-3" in 6.iv).0

We can also compare our method with the results of the latest method on single worker tracking under occlusions (Zhu et al., 2016). The authors of this study achieved an average F-measure metric of 65.2% and a distance error that ranged from 8.6 pixels to 19.7 pixels. Our method performs better for

similar and higher (>80%) occlusion levels in both metrics as it returns an average distance error d_{error} of 6.34 pixels and an average F-measure metric equal to 75.74%.

Table 3-5: Average quantitative evaluation results *(best values are highlighted bold).*

	Video samples	F-measure (%)		d_{error} (pixels)	
		Chapter 3	Park2016	Chapter 3	Park2016
1.	Abrupt movement	**83.68**	55.45	**10.64**	50.78
2.	Background clutter	**77.31**	50.84	**11.97**	30.04
3.	Scene illuminations	72.95	**75.81**	13.34	**12.74**
4.	Posture variations				
	Worker ("0-2")	**50.89**	30.88	**23.89**	25.36
	Worker ("1-3")	**57.12**	20.62	**14.96**	42.89
5.	Scale variations	**81.62**	61.03	**4.92**	16.31
6.i	Occlusion by wall	**65.31**	15.72	**11.00**	18.61
6.ii	Occlusion by steel bars (A)				
	Worker ("0-2")	**67.48**	27.09	**7.51**	32.09
	Worker ("1-3")	**81.40**	60.64	3.78	**2.46**
6.iii	Occlusion by ladder	**75.62**	16.63	**7.91**	23.74
6.iv	Occlusion by steel bars (B)				
	Worker ("0-2")	**84.71**	39.74	**3.00**	17.01
	Worker ("1-3")	**79.90**	54.05	4.85	**3.64**
	Occlusion Average	**75.74**	35.65	**6.34**	16.26
	Total Average	**72.17**	42.38	**9.81**	22.97

3.5.3. Qualitative evaluation

This section evaluates the performance of the method proposed in this chapter in a qualitative way to show that it is able to track multiple workers under appearance similarity. Firstly, the proposed visual tracking method is tested on a video sample that contains congested workers as illustrated in Figure 3-26. In this video sample workers "4" and "2" are overlapped by workers "3" and "1" respectively. Figure 3-26 shows that the method maintains tracking in both cases of congestion even when workers are almost identical ("2" vs "1"). Secondly, the method is evaluated under several challenges that occur simultaneously (see Figure 3-27). In this figure: a) worker "4" is initially occluded by a wall as seen in frame #2 while worker "9" is hidden by limitations of the field of view, b) workers "3", "6", "7," and "8" are continuously occluded by their work benches, c) worker "1" is severely overlapped by worker "2" in frames #17, #19, d) worker "7" is overlapped by worker "6" in frame #85, e) workers "6", "7", and "8" are of a very small scale as compared to worker "9" from frame #2 to frame #85, and f) workers "2", "4" and "9" have background clutter caused by floor stripes and a steel separation wall which are both coloured with the same Hi-Vis apparel's yellow colour. As seen in this figure, our method is able to preserve tracking for all nine workers, and terminates tracking for workers "6" and "9" that left the field of view.

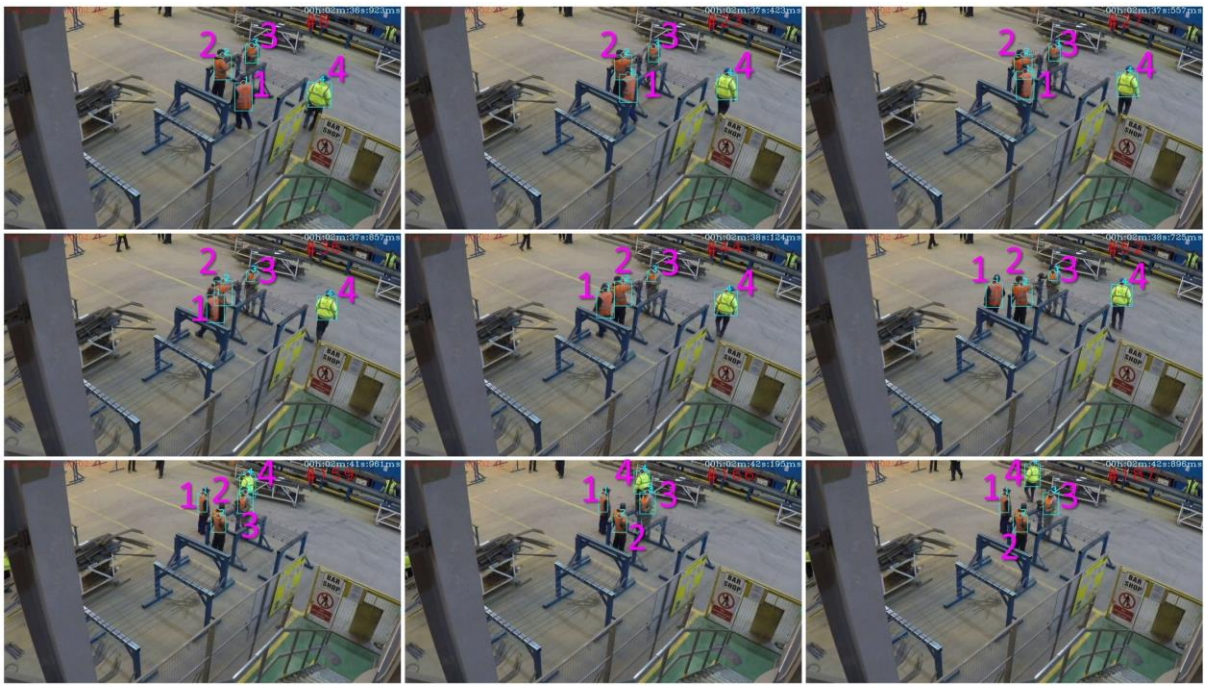

Figure 3-26: Screenshots of the performance of the proposed method while tracking workers under congestion (the IDs of the tracked workers are manually highlighted with pink).

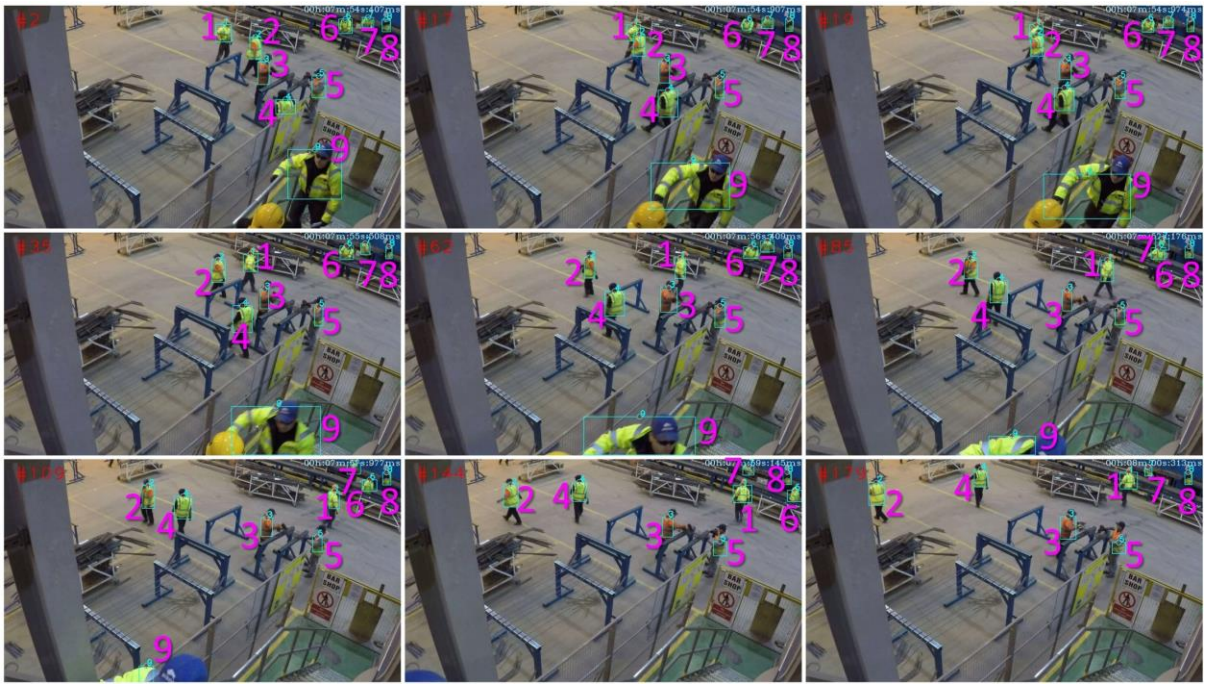

Figure 3-27: Screenshots of the performance of the proposed method while tracking multiple workers under combined challenges: a) congestion, b) scale variations, c) occlusions and d) background clutter (floor with yellow coloured stripes, wall coloured yellow), (the IDs of the tracked workers are manually highlighted with pink).

3.6. Chapter overview

This chapter proposes a computer vision-based method that exploits targets' colour features to track multiple workers that share similar appearance under several challenges. These challenges are: a) abrupt movement, b) background clutter, c) scene illuminations, d) posture variations, e) scale variations, f) occlusions and g) congestions. The proposed method outperforms the latest visual methods that focus on tracking of workers in terms of spatial overall accuracy and distance error. It scores an average F-measure metric equal to 72.17% and an average distance error d_{error} of 9.81 pixels and succeeds in tracking multiple workers when several challenges occur at the same time.

This type of worker monitoring improves safety in construction sites and labour productivity. Existing monitoring practices either rely on work sampling and observation techniques or exploit tags. The former is extremely labour intensive and time consuming considering it must be repeated for all workers on a daily basis. The latter is not usually welcome by the personnel and implies high cost in the long term due to maintenance and purchase of hundreds of tags. Existing visual based tracking methods are less obtrusive and more cost efficient since they use the cameras of the surveillance system. However, they lack applicability as they require human operators to correct tracking under congestion.

The main restriction of the method proposed in this chapter is that it is limited to construction workers that wear Hi-Vis apparel. This is due to the offline training of the appearance model with the according colour features. This method proves that it successfully overcomes the most common challenges that arise while tracking workers within a construction jobsite in order to achieve long term monitoring.

4

Matching of construction workers across views for automated 4D vision tracking

In this chapter, a computer vision-based method for matching the same workers across multiple cameras is introduced. This method uses as inputs the outputs of the 2D tracking method presented in Chapter 3. This target matching is essential to computer vision-based 3D tracking. This type of tracking requires that targets are matched under the same unique ID not only across subsequent frames of a single camera (intra tracking) but also across multiple cameras (inter tracking). This inter camera matching is straightforward when it involves easily distinguishable targets in uncluttered scenes. However, it can be challenging in industrial scenes such as construction sites due to congestion, occlusions, and workers in greatly similar high visibility apparel.

4.1. Introduction

To date, target matching is performed mostly manually. In sports particularly, two to three human operators are required in order to constantly label the twenty two players across all cameras that spread along the sports field (Chyronhego.com, 2016). However, this manual labelling gets labour intensive and time consuming as the operators repeat the process every time tracking is initialized per target per view. This occurs when a target enters a view or re-appears after being occluded. This process is even more labour intensive in construction for the following reasons: a) the total number of workers exceeds by far the one of players, b) work shifts last longer than a sports match, and c) more cameras in general are required in order to capture the entire range of a job site due to its size. Especially the latter makes matching workers manually a nearly impossible task. Figure 4-1 illustrates such an example. Typically, construction workers traverse the jobsite several times per day for either: a) picking up material, b) visiting rest areas, and c) getting instructions by engineers or foremen. A worker in area 1 moves only within the field of views of cameras 1, 3 & 4. This worker is also captured by camera 2 when he/she moves to area 2, and likewise in area 5 and camera 4. Each time the operator must match him/her manually.

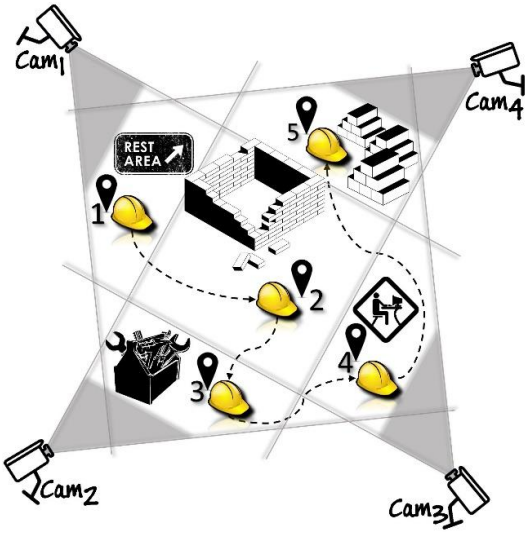

Figure 4-1: Worker's motion through the surveillance cameras of a jobsite.

Tracking errors is another cause that increases the labour intensiveness of manual matching. This is mainly due to illumination and posture variations, occlusions, and congestion issues that challenge the performance of vision-based tracking methods (Park et al., 2011; Teizer & Vela, 2009). As a result, tracking fails frequently and terminates (FINE, 2017; Y. Liu, Wang, & Chen, 2014). Targets lose their previously assigned identification label when this occurs. Hence, matching has to be repeated every time tracking re-initializes.

Existing software solutions have tried to address this matching problem using facial appearance features (Axis.com, 2016; Business.panasonic.co.uk, 2016; Facefirst.com, 2016). These solutions automatically recognize the same person across several either disjoint or overlapping cameras. This approach requires a clear line of sight to each target's face. However, the nature of construction jobsites makes this impossible due to several occlusions, caused by construction components and equipment (e.g. formwork), frequent bending (e.g. steel fixing, brick laying), and the hard hat and goggles. These issues are clearly illustrated in Figure 4-2. In this example, out of a total of sixteen workers only the worker with ID "13" is clearly captured by the camera field of view.

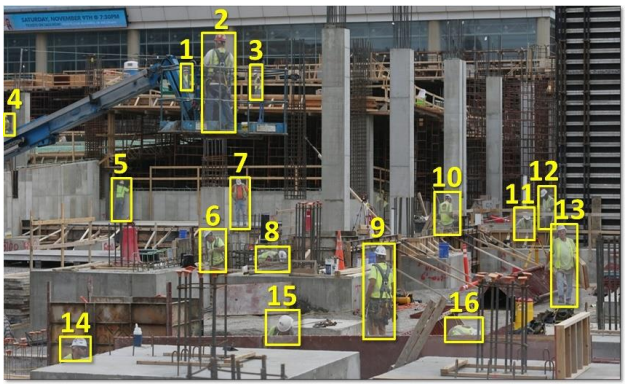

Figure 4-2: Typical workforce in a construction site.

In summary, the main issue of the existing inter camera matching practices is the lack of automation. The large number of workers in combination with the large scale of jobsites turns manual matching into a time consuming and labour intensive practice. Hence, the aim of this chapter is to propose a method that automatically matches workers across multiple cameras. The proposed method uses as input the output of the computer vision-based 2D tracking method of Chapter 3 and searches for potential matches in three sequential steps. It terminates only when a positive match is found. The first step returns the strongest candidate by correlating a segment of workers' past 2D trajectories. The second employs geometric restrictions, whilst the third correlates colour intensity values. The chapter concludes with the performance evaluation of the proposed matching method.

The remainder of this chapter continues with section 4.2 that provides a discussion of the current state of research in computer vision-based matching methods. Then, the proposed method is presented in sections 4.3, 4.4 and evaluated in section 4.5. Lastly, a chapter overview is provided in the final 4.6 section.

4.2. Previous related work

The studies that focus on target inter camera matching can be roughly grouped into two categories based on the features used. The first category relies on appearance based features, and specifically colour, texture and shape. The second uses spatiotemporal features, such as speed, trajectory position and direction, entry/exit zones of monitored areas and camera topology. The author reviews the latest matching methods mainly for people, since the objective of this chapter is to automate matching in order to achieve 3D tracking of construction workers.

Studies that rely on appearance features only use vectors to efficiently describe the neighbourhood around each pixel point (Bay et al., 2008; Dalal & Triggs, 2005; Lowe, 2004). These feature vectors depict either corners and edges or entire blobs (Kong et al., 2013). Such methods perform matching by comparing the feature vectors based on experimentally defined thresholds (Joglekar et al., 2010). Blobs are more efficient in matching people because the human body lacks distinguishable edges or corners. Such blobs are predominately described via colour histograms that are robust under occlusions and scale variations. However, the same people might appear differently from camera to camera due to discrepancies in the viewing angle. The same target can be captured frontside in one camera and sideways in the second if the cameras have a large relative rotation. Such changes in posture variations often change the colours seen, and hence can be very challenging for colour based matching methods.

Unique soft-biometric trails, such as hair/sleeves length and clothes colour/pattern were extracted separately for the head, torso and legs to overcome appearance restrictions (Zhou et al., 2016). Other researchers employed instead machine learning-based classification algorithms trained online

with both colour and texture features (An et al., 2016; Gray & Tao, 2008; Lin et al., 2016; Shah et al., 2016; Teixeira & Corte-Real, 2009; Wang et al., 2014). Such type of training improves matching efficiency as the appearance model is incrementally updated with the latest changes. The outcome depends only on the quality of the dynamically available data in these studies. This lack of offline training can be a downside only if not all appearance features are available (Wang, 2013). Such probability is quite common in highly occluded environments like construction jobsites. Different types of obstacles (e.g. equipment, co-workers, and benches) may visually block the workers' distinguishable training data for long durations in these cases (see (a-d) in Figure 4-3).

Illumination and lighting variations can also make colour histograms appear differently from camera to camera. Han & Kim (2013) proposed the use of Hue values of the HSV colour space to overcome this. Mazzeo & Spagnolo (2008) employed supervised training for learning the Brightness Transfer Model (BTF) among several non-overlapping cameras. However, such training requires manual labelling of the same tracked people that has to be repeated every time the position of a camera is changed i.e. tilted. Chen et al. (2008) and Gilbert & Bowden (2006) extracted the colour correspondences through unsupervised methods to avoid this computational complexity. Overall, as noted by Cheikh et al. (2012), existing appearance based matching methods are not efficient when people have relatively similar appearance. This is the case of construction workers because of their high visibility (Hi Vis) apparel as illustrated in (e) of Figure 4-3.

Figure 4-3: Examples of occluded (a-d) and similarly dressed (e) construction workers.

Researchers incorporated both appearance and spatiotemporal features to overcome the limitations of the above mentioned appearance-based inter camera matching methods. In this respect, Choi & Seo (2011) trained a support vector machine classifier in order to match the same football players across 4 cameras. The training data comprised of players' appearance and position across the sports field. The authors used the former to group all players into 5 categories (team A/B, goal-keeper A/B, and referees). This way they managed to reduce the number of candidates for each player. Then they used position in order to match the remaining players within each classified group. Position was calculated through planar homography. To achieve this, the authors set the height Y of the tracked players equal to zero and used only their bottom coordinates. However, such method is not efficient for construction workers

because planar homography is not applicable due to occlusions such as benches and ladders that block the bottom half of workers. The projection on the ground plane of a point that corresponds to the upper half of a target's body ($Y \neq 0$) could potentially overcome this restriction. However, this type of projection results to an error that leads to mismatching close by targets (Lee et al., 2000). Peng et al. (2015) reduced this error by combining multiple views of the same target. Still, the performance of their method is not promising under congestion and occlusions. In particular, they achieved a high precision of 94% and a lower recall of 87% for basketball players. But recall dropped to 81% when tested in a data set that comprised of more occluded and congested walking pedestrians. Suolan et al. (2016) and Yang et al. (2014) combined appearance features along with prior knowledge of the camera topology and temporal cues. However, such features are efficient only in distinguishing tracked targets across disjoint camera views. This is due to the fact that all targets that move across overlapping camera views share the same camera topology and hence temporal features. In this case geometric restrictions can be applied. Target matching can be achieved through epipolar geometry. According to this, the same point from one view lies along specific lines (epipolar) across all available overlapping camera views. Lee et al. (2016) applied this principle based on an arbitrary defined threshold in order to match workers. Such approach can be seriously challenged under congestion as multiple candidates might be equally close to the same epipolar line. In that respect, Liu et al. (2014) used an expanded epipolar line and managed to reduce by 50% the total number of potential matching candidates.

In summary, the currently available methods have not been able to provide a robust solution to the matching problem. This is mainly due to similarities in workers' appearance, congestion, posture/scale variations, and heavy occlusions. This chapter focuses on construction workers for the following reasons: a) their performance is directly related to labour productivity, b) their number exceeds by far any other type of construction resource, and c) the rest of the resources (e.g. earthmoving equipment, cranes) are in practice easily trackable through tags without arising any privacy conflicts.

The main objectives of this chapter are: a) to identify which features are able to uniquely describe construction workers, b) to develop a method that incorporates these features for the comparison of workers, and c) to devise a method that will be able to efficiently adopt to all matching challenges that arise in a construction site. The key research question that this chapter aims to answer is: "*how can similar in appearance entities be matched among multiple camera views under occlusions, congestion, and posture/scale variations?*"

4.3. Proposed solution

In this section, the matching problem is addressed by proposing a multi-step method for construction workers. Figure 4-4 illustrates the flowchart of the overall proposed method. The skewed parallelogram shapes refer to processes and the circular shapes to inputs/outputs. The grey coloured processes depict the parts of the method with novel contribution. A 2D vision tracking method is used in order to obtain:

a) the 2D trajectory over a period of time ΔT (see (a) in Figure 4-4), b) the position (x, y) in a frame at time t_n (see (b) in Figure 4-4), and c) the appearance at time t_n as enclosed within the bounding box that depicts the output of tracking (see (c) in Figure 4-4).

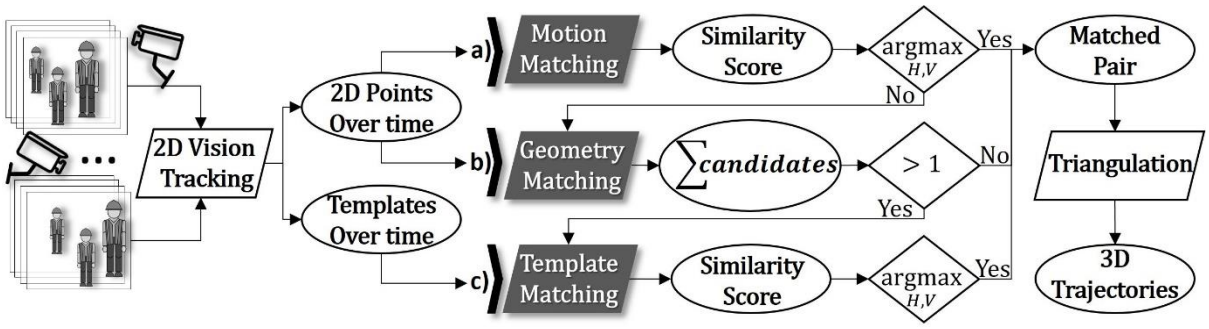

Figure 4-4: Flowchart of computer vision-based method for matching construction workers across multiple camera views.

The overall proposed method comprises of three steps. Each of them depicts a unique matching method that tackles only a few of the matching challenges. They exploit 2D trajectories, geometric restrictions, colour data and are sequentially activated. Such type of processing allows computational efficiency since the searching for a match terminates after one step returns a positive result. The processing sequence is chosen given the expected effectiveness of each method (see Table 4-1).

Table 4-1: Expected efficiency of geometry, motion and template-based methods under the most common matching challenges met on a construction site.

	Geometry	Motion	Template
Congestion	✗	✓	✓
Occlusions	✓	✓	✗
Posture Variations	✓	✓	✗
Scale Variations	✓	✓	✓
Similar Appearance	✓	✓	✗
Lack of Motion	✓	✗	✓

The proposed method employs the motion-based method in the first step even if it is expected to perform equally well with the geometry-based method. This choice relies on the hypothesis that the motion-based method will be more efficient in a congested environment since workers have motion while walking in and out from camera views. In the second and third steps, the geometry-based and template-based matching methods are implemented respectively. The latter is expected to be quite negatively affected by issues such as posture variations, occlusions, and appearance similarity since it relies on colour features. However, its addition assists in reducing potential missed matches caused by the assumed ineffectiveness of the geometry-based method under congestion.

The 2D image coordinates of the same workers are triangulated after a positive match is confirmed in order to turn them into 3D world coordinates. This output is illustrated with "P" in Figure 4-5 and represents each worker's position across the range of a jobsite. The proposed method uses as reference point the upper middle point of the tracker's bounding box for each target in contrast to Konstantinou & Brilakis (2016), and Y. Lee et al. (2016) that used the centroid. This choice was made after noticing large triangulation errors that are caused by the use of centroids as reference points under occlusions. For example if one worker's body is fully captured in camera A but half captured in camera B then the y coordinate of the centroid in camera A will be $\frac{H}{2}$ whilst in camera B $\frac{H}{4}$.

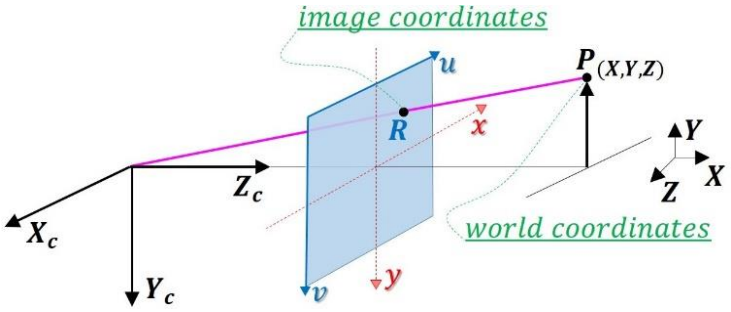

Figure 4-5: Image and world coordinate systems of a pinhole camera model.

The proposed method makes the following assumptions: a) workers will not move simultaneously out of all the field of views in order to enter new ones Cam_{new}, b) all areas of the monitored jobsite are covered by at least two cameras, and c) no "blind" areas exist. A camera set-up that complies with these assumptions, allows to: a) retain workers' previous IDs in at least one camera's view Cam_{pr}, and b) keep an equal number of unlabelled candidates and workers with known IDs. This works as prior knowledge that allows to re-assign to all unlabelled workers that enter new camera views Cam_{new} their previous IDs. This way all workers are uniquely represented by the same labels during large periods of time. Such label stability will assist project managers to review easier and faster several workers' movement diagrams for improving construction practices. In a different case, where it is impossible to arrange cameras in such way that allows to keep the IDs in at least one view, the IDs of workers in Cam_{pr} will simply represent newly assigned ones.

4.4. Proposed methodology

This section presents a computer vision-based method for matching construction workers across multiple camera views.

4.4.1. Motion-based matching method

This method compares one worker from Cam_{pr} with candidates from all Cam_{new} based on a feature motion vector M. This vector contains workers' 2D coordinates (x, y) of past positions P_{t_i} over a period of time ΔT:

$$M_{(nx1)} = \begin{bmatrix} P_{\mathrm{t}} \\ \cdot \\ \cdot \\ \cdot \\ P_{\mathrm{t}+\Delta T} \end{bmatrix} = \begin{bmatrix} \{x_t, y_t\} \\ \cdot \\ \cdot \\ \cdot \\ \{x_{t+\Delta T}, y_{t+\Delta T}\} \end{bmatrix} \quad (4\text{-}1)$$

where t is a timestamp, and x_t, y_t are a worker's coordinates at time t. It is assumed that each worker has the same uniquely distinguishing motion pattern over time across all camera views, in terms of direction and magnitude (see Figure 4-6). Hence, only workers with identical motion pattern depict the same entity.

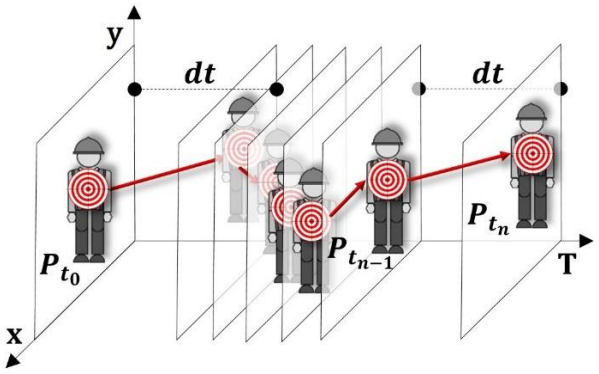

Figure 4-6: Worker's 2D motion data across time.

This motion-based matching method compares workers between Cam_{pr} and Cam_{new} by correlating their 2D trajectories over time ΔT. A candidates' correlation table (CCT) (see Figure 4-7) is introduced in order to achieve this.

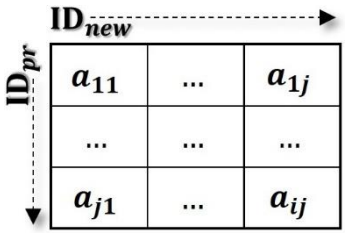

Figure 4-7: Candidates correlation table (CCT).

The vertical column contains the workers with known IDs ID_{pr}, whilst the horizontal column all the unlabelled candidates ID_{new}. Each cell of the table contains a similarity score $a_{i,j}$ which results from

the comparison between each worker from Cam_{pr} across line i and one candidate from Cam_{new} in column j. This is expressed as follows:

$$a_{i,j} = \frac{r_x + r_y}{2} \tag{4-2}$$

where r_x, r_y are the Pearson's correlation coefficients:

$$r_x = \frac{\sum_{i=1}^{i=n}(x_{pr_i} - \bar{x}_{pr})(x_{new_i} - \bar{x}_{new})}{\sqrt{\sum_{i=1}^{i=n}(x_{pr_i} - \bar{x}_{pr})^2}\sqrt{\sum_{i=1}^{i=n}(x_{new_i} - \bar{x}_{new})^2}} \tag{4-3}$$

$$r_y = \frac{\sum_{i=1}^{i=n}(y_{pr_i} - \bar{y}_{pr})(y_{new_i} - \bar{y}_{new})}{\sqrt{\sum_{i=1}^{i=n}(y_{pr_i} - \bar{y}_{pr})^2}\sqrt{\sum_{i=1}^{i=n}(y_{new_i} - \bar{y}_{new})^2}} \tag{4-4}$$

where x_{pr}, y_{pr} are the coordinates in the image of the labelled tracked worker in Cam_{pr}, and x_{new}, y_{new} are the coordinates in the image of the unlabelled tracked worker in Cam_{new}.

A worker will have the same motion pattern across cameras in terms of direction and speed. This is valid only when trajectories are positively correlated i.e. $r_x, r_y > 0$. Hence, candidates with either r_x or r_y negative are discarded. The proposed matching method searches for the "strongest" candidate to confirm speed similarity. It relies on an algorithm that performs repetitive search rounds until the maximum correlation score is confirmed between one worker with known previous ID ID_{pr} and one candidate ID_{new}. This searching occurs in two directions. Firstly, the proposed algorithm calculates for each i^{th} line the maximum horizontal $a_{i,j}$ (p_{H_i}):

$$p_{H_i} = \arg\max_{i}\{a_{i,j}\} \tag{4-5}$$

Then, at the j^{th} position of this horizontal maxima (p_{H_i}) the algorithm calculates also the vertical maximum (p_{V_j}):

$$p_{V_j} = \arg\max_{j|i}\{a_{i,j}\} \tag{4-6}$$

This two-dimensional searching for local maxima guarantees the uniqueness of a returned positive match which is expressed as follows:

$$\forall\, ID_{new} \rightarrow \begin{cases} match\,, p_{H_i} = p_{V_j} \\ non\ match, p_{H_i} \neq p_{V_j} \end{cases} \tag{4-7}$$

Hence, if p_{H_i} is equal to p_{V_j} then this method matches the compared workers and removes the candidate from the column with the unlabelled IDs ID_{new}. Matching terminates only when all candidates within ID_{new} are successfully matched with all workers from ID_{pr}. The advantage of the proposed searching for the "strongest candidate" algorithm is that it is invariant to similarity thresholds. Two representative examples in Figure 4-8 illustrate how the proposed searching algorithm works. In the example of case A, the algorithm performs three search rounds (a)-(c), until all three candidates are correctly matched (green shaded cells). This is the optimum/minimum number of search rounds and is equal to the total number of candidates. It is achieved only when each round returns a positive match. However such a case is not always feasible, especially when workers have quite similar motion patterns i.e. direction and speed. Case B shows such an example where the searching algorithm requires four search rounds (d)-(g) instead of three until all three candidates are matched to the three workers with known previous IDs. This is due to the first round R1st that returns no positive match as the proposed searching algorithm results in two workers W_1 & W_3, as potential matches for candidate C_3 in (d). When this occurs, the searching algorithm repeats in R2nd the R1st in order to update the CCT table that a_{13} is not a potential positive match for C_3 (cell a_{13} is shaded red).

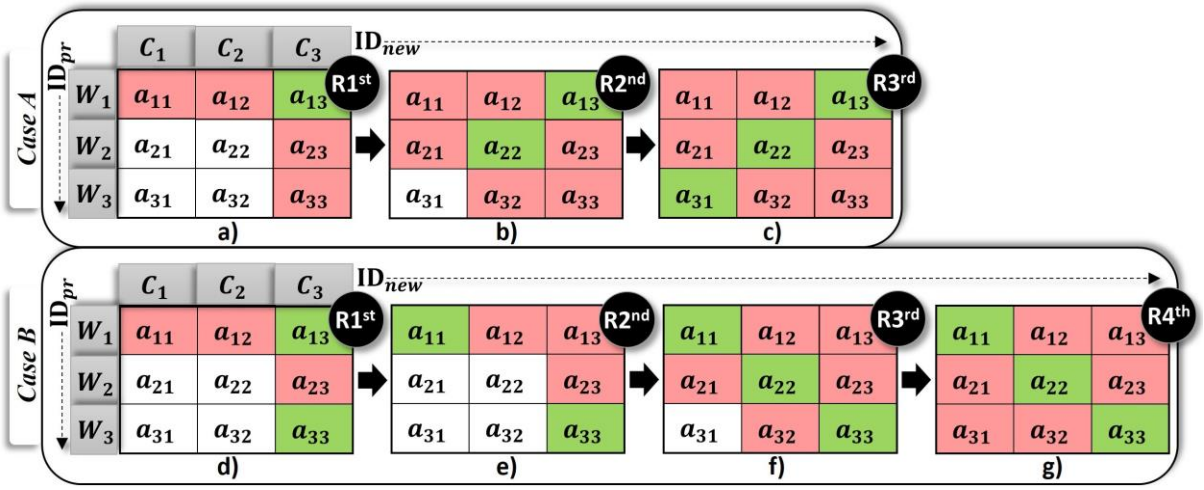

Figure 4-8: Proposed searching algorithm for the "strongest" candidate.

Lastly, the motion-based matching method is adjusted for non-conjugated cameras. Figure 4-9 illustrates analytically the reason for this. In this figure, two cameras capture the 2D trajectory of the same worker over a period of time ΔT (a). However, the angle θ between the local image coordinate systems of the two cameras as seen in (b) affects the way a worker's 2D trajectory projects on each camera's local coordinate system. Figures in (c) and (d) show how the resulting x/y and x′/y′coordinates differ in terms of magnitude.

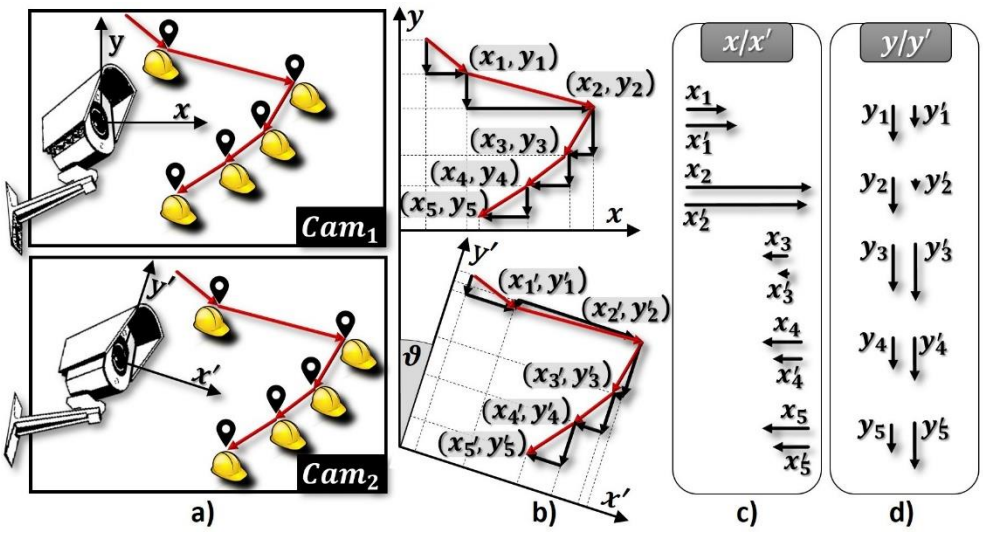

Figure 4-9: Projection of tracked 2D motion paths on the local coordinate systems of non-conjugated cameras.

All trajectories are transformed with respect to a common coordinate system to alleviate the effect of the above issue while correlating trajectories. The proposed method uses the 1st horizontal line of a chessboard as the x axis of the proposed reference coordinate system (see Figure 4-10).

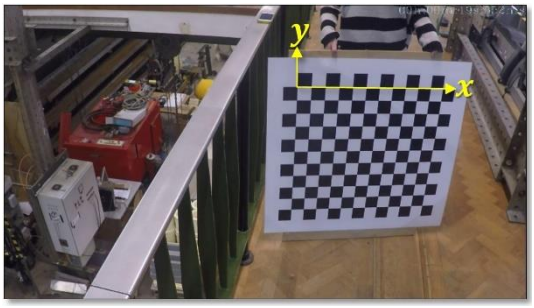

Figure 4-10: Reference coordinate system for motion matching between non-conjugated cameras.

Firstly, the corners that lie along this reference line from all camera views are detected. Then, principal component analysis (PCA) uses the coordinates of the detected corners to define the relationship between the reference system and a camera's cam_i local coordinate system. PCA extracts a 2x1 eigenvector $e_{cam_i} = [e_{x_i} \quad e_{y_i}]^T$ that fits a line through the coordinates of the detected corners. This eigenvector corresponds to the maximum eigenvalue and is used to calculate the relative rotation θ_{cam_i} between the reference system and each of the local coordinate systems i:

$$\theta_{cam_i} = \tan^{-1} \frac{e_{x_i}}{e_{y_i}} \tag{4-8}$$

Lastly, θ_{cam_i} is used in a 2D rotation matrix in order to align the local trajectories of a cam_i according to the proposed reference coordinate system:

$$\begin{bmatrix} x_{ref_i} \\ y_{ref_i} \end{bmatrix} = \begin{bmatrix} \cos\theta_{cam_i} & -\sin\theta_{cam_i} \\ \sin\theta_{cam_i} & \cos\theta_{cam_i} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad (4\text{-}9)$$

where x_{ref_i}, y_{ref_i} are the coordinates of a target W_1 tracked in cam_i with respect to a reference coordinate system, and x_i, y_i are the coordinates of the tracked target W_1 with respect to the local coordinate system of cam_i.

4.4.2. Geometry-based matching method

The matching method of the second step introduces a geometric restriction for matching workers. This method combines the epipolar constraint and the reference point as provided by the vision tracker to achieve this. This method is invariant to occlusions, scale and posture variations that occur due to the relative distance and rotation of cameras (see Figure 4-11). Traditional appearance-based matching methods would fail under such conditions since not all features are equally visible from all cameras.

Figure 4-11: Posture variations between frames of non-conjugated cameras.

Initially, all cameras are calibrated. The calibration method of Zhang & Member (2000) is implemented in order to extract the intrinsic parameters and the eight point algorithm (Hartley, 1997) is used for calculating the extrinsic. These methods return the camera matrix K, and the essential matrix E. The K is a 3X3 matrix that contains the intrinsic parameters i.e. principal point, skew coefficients, distortions, and focal length. The E is also a 3X3 matrix that depicts the extrinsic parameters i.e. relative rotation (R) and translation (T) between the fixed cameras:

$$E = T_x R \quad (4\text{-}10)$$

Then the epipolar constraint is applied in order to calculate for each worker's P_{pr} reference point in Cam_{pr} its projection as an epipolar line l_{ep} in Cam_{new}:

$$\mathrm{l}_{ep} = {K_{Cam_{new}}}^{-T} E {K_{Cam_{pr}}}^{-1} P_{pr} \tag{4-11}$$

where $K_{Cam_{new}}, K_{Cam_{pr}}$ is the camera matrix K of Cam_{new} and Cam_{pr} respectively. A positive match occurs if along one l_{ep} lies only one candidate P_{new}. However, the epipolar constraint may fail to return a match due to instabilities of the proposed reference point. This is illustrated in Figure 4-12. In this example the centroid of worker with ID "0" from Cam_{pr} does not lie along the l_{ep} (black line) of candidate with ID "1" in Cam_{new} as the tracking method fails to capture both workers equally accurately.

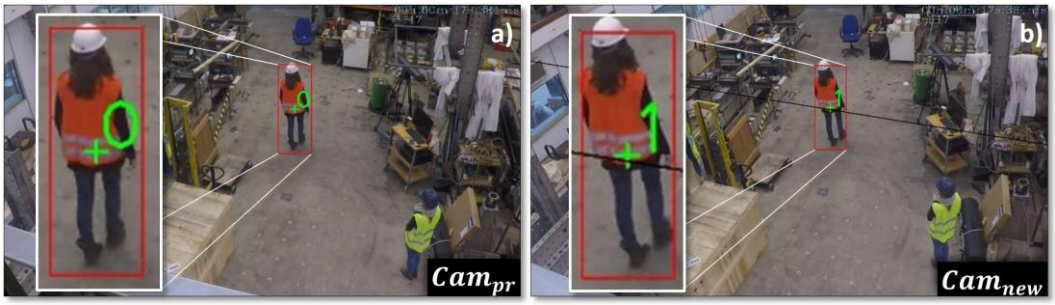

Figure 4-12: False negative (FN) matching of a single worker between two cameras (a-b).

This tracking instability becomes obvious when the bounding box of a tracker drifts away (fluctuates) from the target. Figure 4-13 presents such an example. In this figure, a kernel-based tracking method (Ross et al., 2008) drifts while a worker walks (a-d) or changes posture (e-f). The red rectangle depicts the tracking output whilst the yellow dotted rectangle the ground truth. The difference between these two rectangles highlights the tracking instability. For simplicity, the remainder of this thesis uses the term fluctuation error $error_f$ when referring to this.

Figure 4-13: Drifting issue of a computer vision tracking method under walking (a-d) and posture variations (e-f).

This geometry based method turns l_{ep} into a search band in order to alleviate the effect of $error_f$. Figure 4-14 illustrates with a continuous line the l_{ep} and with dotted lines the upper l_{ep}^{up} and the lower l_{ep}^{low} boundaries of the proposed matching search band. These boundaries are expressed as follows:

$$l_{ep}^{up} = ax_{new} + by_{new} + c + h \quad (4\text{-}12)$$

$$l_{ep}^{low} = ax_{new} + by_{new} + c - h \quad (4\text{-}13)$$

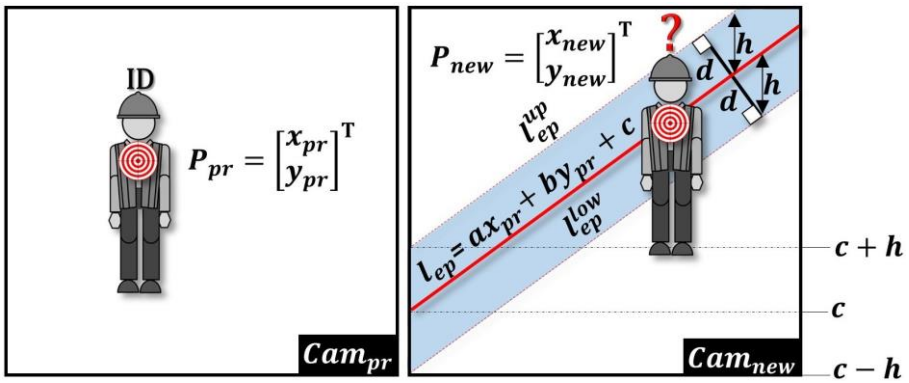

Figure 4-14: Geometry-based matching search band.

This geometry-based method defines $error_f$ as equal to the average vertical difference h between two parallel lines:

$$h = l_{ep} - (ax_{new} + by_{new} + c \pm error_f) \quad (4\text{-}14)$$

It uses l_{ep} to calculate the line that depicts the ground truth. To achieve this, we manually select for a worker P_{pr} in Cam_{pr} the reference points for which no drifting appears ($error_f = 0$). This way we restrict the appearance of $error_f$ on the tracked target of Cam_{new}. Hence, this first line is defined as $l_{ep} = ax_{pr} + by_{pr} + c$. The second line passes through the same candidate P_{new} in Cam_{new} and is equal to $ax_{new} + by_{new} + c \pm error_f$. Once we have extracted h, the width (d) of the proposed matching search band is measured through the following equation:

$$d = \frac{|ax_{new} + by_{new} + (c \pm h)|}{\sqrt{a^2 + b^2}} \quad (4\text{-}15)$$

This geometry-based method returns a positive match if only one candidate's reference point is measured within a search band. It returns no match if a search band encloses multiple candidates. This occurs due to congestion i.e. several workers within a restricted area. Figure 4-15 illustrates such an

example. In this figure, a worker will have to be compared to five candidates. This is because all the reference points of these candidates fall within the same matching search band (red shaded area).

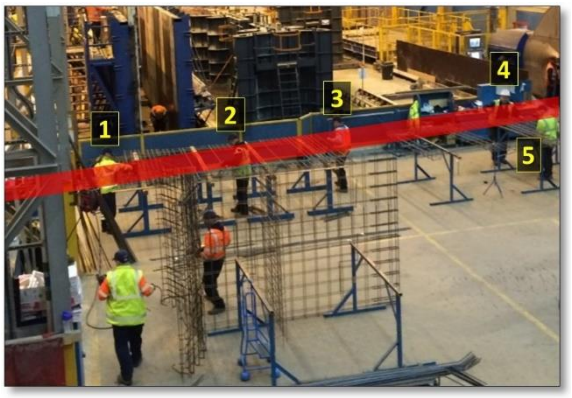

Figure 4-15: Multiple candidates within the proposed matching search band.

4.4.3. Template-based matching method

The third step is activated when the second step yields a false outcome. This step overcomes the ambiguity when the proposed matching search band encloses multiple candidates. This matching method uses candidates' correlation table (CCT) of section 4.4.1 to compare colour values instead. It is hypothesized that small differences in a worker's appearance (i.e. sleeves, colour hat) may be enough for achieving high similarity scores. This matching method, employs the normalized cross correlation method (NCC) to compare one template patch T(x, y) to a source patch I(x, y):

$$r_c = \frac{\sum_{x\prime,y\prime}(T(x\prime,y\prime)*I(x+x\prime,y+y\prime))}{\sqrt{\sum_{x\prime,y\prime} T(x\prime,y\prime)^2 * \sum_{x\prime,y\prime} I\,(x+x\prime,y+y\prime)^2}} \quad (4\text{-}16)$$

Each of the T(x, y), I(x, y) is taken equal to the colour patch as enclosed within the tracking output (bounding box) to reduce computational complexity. The template patches T(x, y) represent workers in Cam_{pr} and the source patches I(x, y) depict candidates in all Cam_{new}.

4.5. Calculation of 3D trajectories

The final step of the overall proposed method employs the mid-point triangulation method to turn the 2D corresponding reference points into 3D coordinates when a positive match is returned from any of the previous steps. This method is selected as noise (e.g. fluctuation, calibration error) prevents the rays of corresponding points projected in two images from intersecting at the original 3D point as illustrated in Figure 4-16. The proposed triangulation method alleviates this issue by calculating instead the mid-

point of the minimum vertical distance V_m between the two rays. This method is taken from the literature and is explained in more detail below.

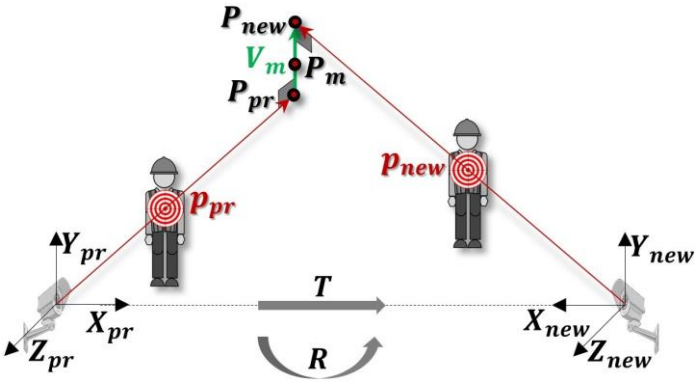

Figure 4-16: Mid-point triangulation method.

A world coordinate P_m is expressed as $P_{pr} = [X_{pr} \quad Y_{pr} \quad Z_{pr}]$ with respect to the coordinate system of camera Cam_{pr} and as $P_{new} = [X_{new} \quad Y_{new} \quad Z_{new}]$ with respect to camera Cam_{new}. If the coordinate system of Cam_{pr} is used as reference, then P_{pr} and P_{new} are linked through the following equation:

$$P_{pr} = RP_{new} + T \tag{4-17}$$

where R, T is the rotation and translation matrix respectively between cameras Cam_{pr} and Cam_{new}. P_{pr} and P_{new} are mapped in image planes at points $p_{pr} = [x_{pr} \; y_{pr} \; 1]$ and $p_{new} = [x_{new} \; y_{new} \; 1]$ respectively as follows:

$$P_{pr} = ap_{pr} \tag{4-18}$$

$$P_{new} = R^{-1}(bp_{new} - T) \tag{4-19}$$

where a, b are real numbers. A perpendicular vector V_m is equal to the vector product between P_{new} and P_{pr}:

$$V_m = c\left(p_{pr} \otimes (R^{-1}p_{new} - T) - T\right) \tag{4-20}$$

where c is a real number. The three parameters a, b, c are calculated by summing P_{pr}, P_{new}, and V_m vectors though the following equation:

$$\begin{aligned} & P_{pr} + V_m = P_{new} \\ \Leftrightarrow \; & \left(ap_{pr}\right) + c\left(p_{pr} \otimes (R^{-1}p_{new} - T) - T\right) = R^{-1}(bp_{new} - T) \\ \Leftrightarrow \; & \left(ap_{pr}\right) - (bR^{-1}p_{new}) + c\left(p_{pr} \otimes (R^{-1}p_{new} - T) - T\right) = -R^{-1}T \end{aligned} \tag{4-21}$$

Finally the 3D location P_m of a target is taken equal to:

$$P_m = P_{pr} + \frac{1}{2} V_m \tag{4-22}$$

The hypothesis tested in this chapter is that the proposed solution comprised of the methods proposed above can significantly enhance the accuracy, precision and recall in matching multiple workers within a congested jobsite under posture/scale variations, appearance similarity, and occlusions.

4.6. Experiments and results

This chapter uses two data sets for validation. The first (data set A) is collected at the structures group laboratory at the Engineering Department of the University of Cambridge (see (a) in Figure 4-17). The second (data set B) is from an offsite manufacturing facility (see (b) in Figure 4-17). The latter depicts a more challenging environment since it contains more workers and heavier occlusions. Hence, data set A is used for the determination of the proposed method's parameters whilst data set B is used for the overall method's validation. Both data sets were collected through two cameras that were mounted similarly to the set-up of a typical surveillance camera system (see (c-d) in Figure 4-17). The experimental set up is described in Chapter 2 (see section 2.4.1)

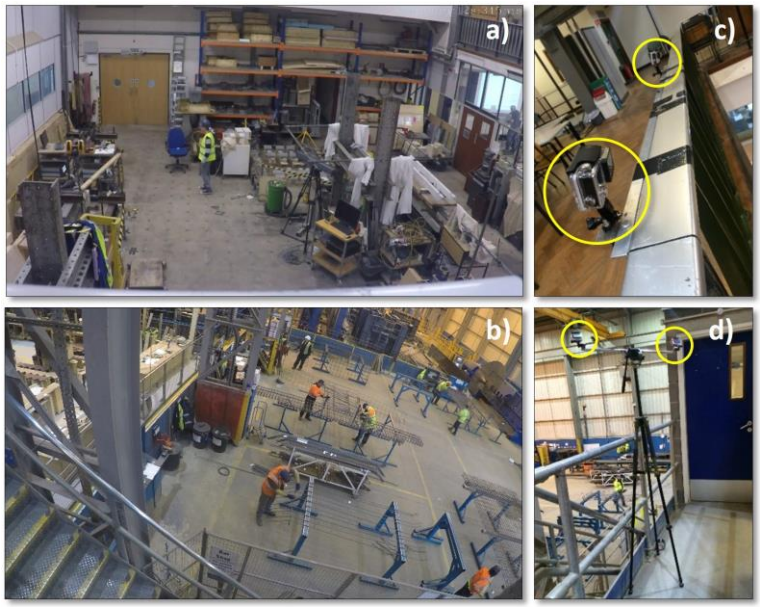

Figure 4-17: Experimental video data sets. (a) Data set A. (b) Data set B.

The quality of the proposed method's evaluation process depends on the sample size to be tested. Hence, this evaluation section defines the minimum number of required samples n through the following equation (Eng, 2003):

$$n = \frac{4Z_{crit}{}^2 p(1-p)}{D^2} \quad (4\text{-}23)$$

A confidence level of 95% is set based on normal distribution (Z_{crit} = 1.960) and a limit of error (D) equal to ±5%. This explicitly means that the accuracy of 95 out of 100 samples will fall within a range of ±5%. The p variable is the proportion of each sample with a specific characteristic and is experimentally measured. The tracked region as described by the tracking output (bounding box) is the type of samples used during evaluation.

A 13x12 chessboard with 60mm square size was used for the calibration of the cameras. Initially, the corners of this chessboard, as captured in each of the synchronized images of the experimental stereo camera set up system, are detected. Then, the known coordinates of the corresponding corners are used for the calculation of the extrinsic and intrinsic camera parameters. Figure 4-18 illustrates these correspondences with same coloured circles.

Figure 4-18: Detection of corresponding points for stereo camera calibration.

The accuracy of this calibration is evaluated by comparing the estimated (see (a) in Figure 4-19) and the real (see (b) in Figure 4-19) baseline between the two cameras (baseline). Figure 4-19 shows that the resulting error was approximately equal to 22mm.

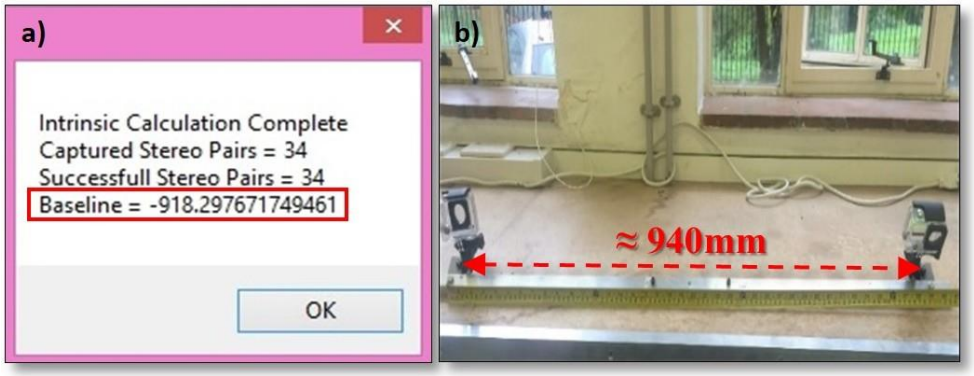

Figure 4-19: Stereo calibration accuracy.

Three metrics measure the proposed method's performance: precision, recall, and accuracy. Precision is the fraction of the total number of correctly matched tracked workers (TP, True Positive) over the total number of incorrectly and correctly matched ones (TP + FP, True Positive + False Positive):

$$\frac{TP}{TP+FP} \tag{4-24}$$

Recall represents the matching completion level and is equal to the number of workers correctly matched (TP) divided by the total number of correctly matched and missed (TP + FN, True Positive + False Negative):

$$\frac{TP}{TP+FN} \tag{4-25}$$

Lastly, accuracy is defined by the number of workers that were correctly matched as same (TP) and those that were correctly not matched (TN, True Negative), over the total sum of the matched workers:

$$\frac{TP+TN}{TP+FP+TN+FN} \tag{4-26}$$

4.6.1. Evaluation of the motion-based matching method

Initially, the performance of this method is evaluated over a variety of lengths of past motion data as seen in Figure 4-20. In this figure, a black line depicts the trajectory of one worker's motion over the past 100ms, 500ms, 1000ms and 1500ms respectively. Performance is measured only when motion exists. This section tests three workers from data set A while walking both in similar and opposite directions.

Figure 4-20: Display of previous motion data (black line) over time.

In total, 1063 pairs of stereo frames with 3189 tracked targets were processed for the determination of the optimum length of past motion data. The unique characteristic of this tested sample is motion. Hence, the p variable of Equation 4-23 is defined by the number of the non-moving targets over the total tracked targets. For the involved tracked targets of data set A, p is set equal to 8%. This method is not designed to return any results for the non-moving targets. Therefore, we consider these missed

matches as TN. Figure 4-21 displays for each tested length of past motion data the total TP and FP matches.

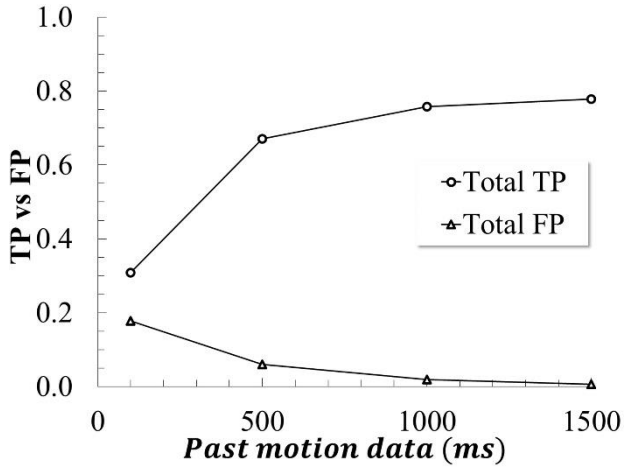

Figure 4-21: Total TP and FP matches over lengths of past motion data (per camera).

The TP and FP performance metrics in Figure 4-21, are normalized to one by dividing with the total number of tracked targets for display purposes. Figure 4-21 shows that for larger lengths of past motion data the total of TP matches increases whilst the total of FP matches decreases. The resulting optimum length is 1500ms since it scores the highest performance for data set A with 99% precision, 94% recall, and 94% accuracy (see Table 4-2). This length parameter was also validated with data set B. In this testing, p variable was measured equal to 39% for a data set of 1205 targets tracked across 1205 pairs of stereo frames. The method returned a high performance of 97% precision, 99% recall, and 98% accuracy, equal to data set A. This confirms the effectiveness of 1500ms as the optimum length of past motion data for comparing moving workers.

Table 4-2: Evaluation of proposed motion matching method over various lengths of past motion data.

Performance metrics	Past motion data (ms)				
	Data Set A				Data Set B
	100	500	1000	1500	1500
Precision	63%	92%	97%	99%	97%
Recall	45%	84%	92%	94%	99%
Accuracy	43%	80%	90%	94%	98%

Figure 4-22 illustrates the performance of the motion-based matching method by comparing workers' 2D trajectories over the last 1500ms. The examples in this figure depict candidates moving either in similar, opposite or random direction. The image pair in (a-b) of Figure 4-22 contains three positive matches. In Cam_{pr}, candidates with IDs "3" and "4" move similarly in terms of direction, whilst candidate with ID "2" walks in the opposite direction. The longer the length of the trajectory the faster

the worker moved over the past 1500ms. This is depicted in (c-d) of Figure 4-22. Candidates with IDs "1" and "2" move faster compared to candidates "3" and "4" in Cam_{pr}. Figure 4-22 in (e-f) shows two TP matches while workers are partially occluded and two TN matches. These missed matches are justified by the lack of motion of workers with IDs "2" and "3" in Cam_{pr} as they stand still while tightening steel re-bars.

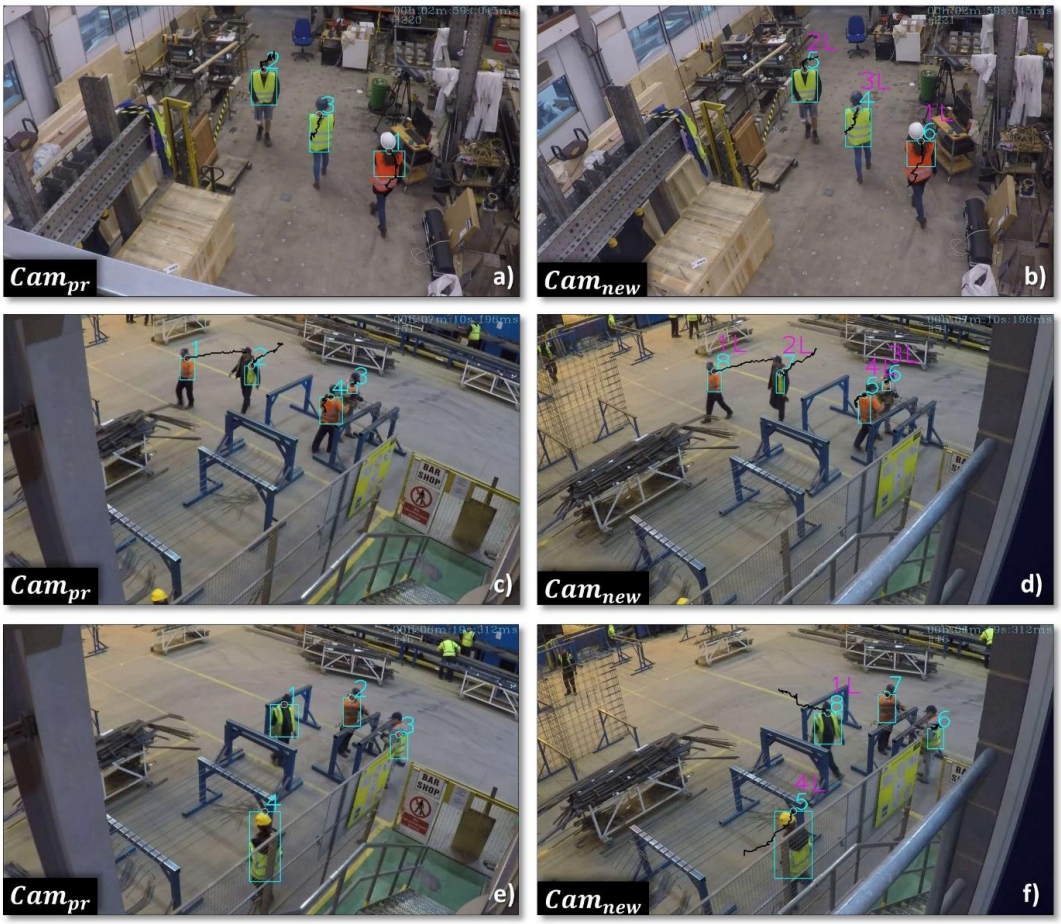

Figure 4-22: Performance examples of the motion-based matching method. (a-b) Three TP matches among two workers in similar and one worker in opposite direction (Data set A). (c-d) Four TP matches among two workers in similar and two workers in random direction (Data set B). (e-f) Two TN matches due to lack of motion and two TP matches under occlusions (Data set B).

4.6.2. Evaluation of the geometry-based matching method

The matching method of the 2nd step searches for unique candidates within the boundaries of a geometrically defined area. Equation (4-15 expresses the width of this search band in accordance to the fluctuation error ($error_f$) for the specific implemented tracking method. In this section, $error_f$ is experimentally defined by following the process as described in section 4.4.2. The sample size for this measurement depends on the proportion of targets whose reference points in Cam_{pr} fail to lie along

their corresponding epipolar lines in Cam_{new}. The p variable for this data set was measured equal to 92%. This percentage depicts that only 8% of the candidates can be matched using epipolar geometry only. In total 532 tracked targets from subsequent stereo frames of data set A are used for this evaluation. This sample captures workers while walking and under posture and scale variations. The $error_f$ for each tracked target is expressed as a percentage of his/her height. This turns the geometry-based matching method invariant to specific threshold values. Figure 4-23 illustrates the experimental (actual) and the fitted (normal) cumulative probability distribution of the resulting values of $error_f$.

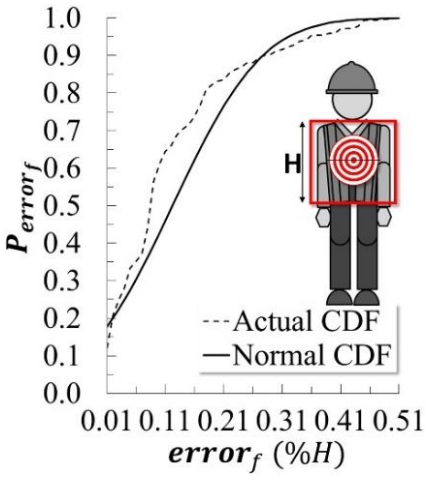

Figure 4-23: Cumulative distribution of the fluctuation error ($error_f$).

The cumulative probability distribution is used to interpolate values of $error_{f_i}$ for specific probability values ($P_{error_{f_i}}$). This interpolation is expressed as follows:

$$error_{f_i} = error_{f_{i-1}} + \left(P_{error_{f_i}} - P_{error_{f_{i-1}}}\right) \frac{error_{f_{i+1}} - error_{f_{i-1}}}{P_{error_{f_{i+1}}} - P_{error_{f_{i-1}}}} \quad (4\text{-}27)$$

Table 4-3 contains the measured width of the proposed matching search band only for $error_{f_i}$ with a probability higher than 50% ($P_{error_{f_i}} > 0.5$).

Table 4-3: Experimentally defined $error_{f_i}$.

P_{error_f}	$error_f$ (%H)
0.5	0.12
0.6	0.15
0.7	0.19
0.8	0.22
0.9	0.28
1.0	0.56

It is possible for a single matching search band in a congested environment to contain more than one candidate. The geometry-based method does not return any matches due to this ambiguity. Hence, such missed matches are considered as true negative (TN). The FN results depict the number of the not matched workers when the width of the search is not big enough to capture the reference point of the same worker due to the effect of $error_f$. Table 4-4 shows the matching performance for different widths of the proposed search band. Initially, 1199 pairs of stereo frames from data set "A" are processed for every tested width, containing a total of 3596 tracked targets in both cameras. The p variable for this testing depicts the total number of targets that are not matched due to congestion and is equal to 6%. This table highlights how effectively the proposed search band leverages the instability caused by $error_f$ and increases matching performance.

Table 4-4: Evaluation of proposed geometry-based matching method over various values of $error_f$.

Performance Metrics	$P_{error_{f_i}}$							
	Data Set A							**Data Set B**
	0.0	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**0.9**
Precision	0.98	1.00	0.99	1.00	1.00	1.00	1.00	0.93
Recall	0.08	0.73	0.79	0.82	0.87	0.90	0.99	0.94
Accuracy	0.08	0.74	0.80	0.82	0.88	0.91	0.99	0.95

The graph in Figure 4-24 defines the optimum width of the proposed matching search band. This figure displays the total normalized TN and FN matches for $error_{f_i}$ values that correspond to the chosen $P_{error_{f_i}}$ probabilities of Table 4-3. Figure 4-24 shows a trade-off between the width of the search band and the resulting matching performance.

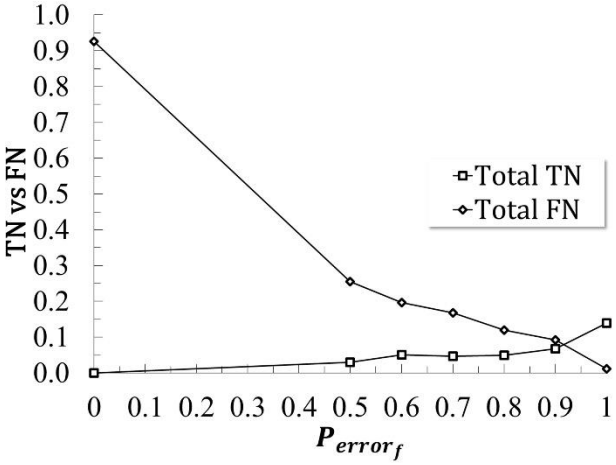

Figure 4-24: TN & FN missed matches over the probability ($P_{error_{f_i}}$) of $error_{f_i}$ values.

The more the size increases, the more total candidates fall within the same search band in a congested environment. As a result of this, the total FN matches decreases whilst the total TN matches increases.

Hence, $error_f$ is set equal to the x value at the intersection point between the TN and FN curves since beyond this point matching ambiguity escalates due to the increasing TN matches. This x value corresponds to $error_f$ with $P_{error_f} = 0.90$. Hence, the probability of getting $error_f$ larger than $\frac{28}{100}$H is only 10% based on the interpolated values of Table 4-3. This experimentally defined width is tested in the congested (p=53%) working environment of data set B to validate its effectiveness. Overall, 415 pairs of stereo frames, and 1806 tracked targets in both cameras are processed. Previous Table 4-4 shows that the geometry-based method features a high performance of 93% precision, 94% accuracy, and 95% recall for Data set B.

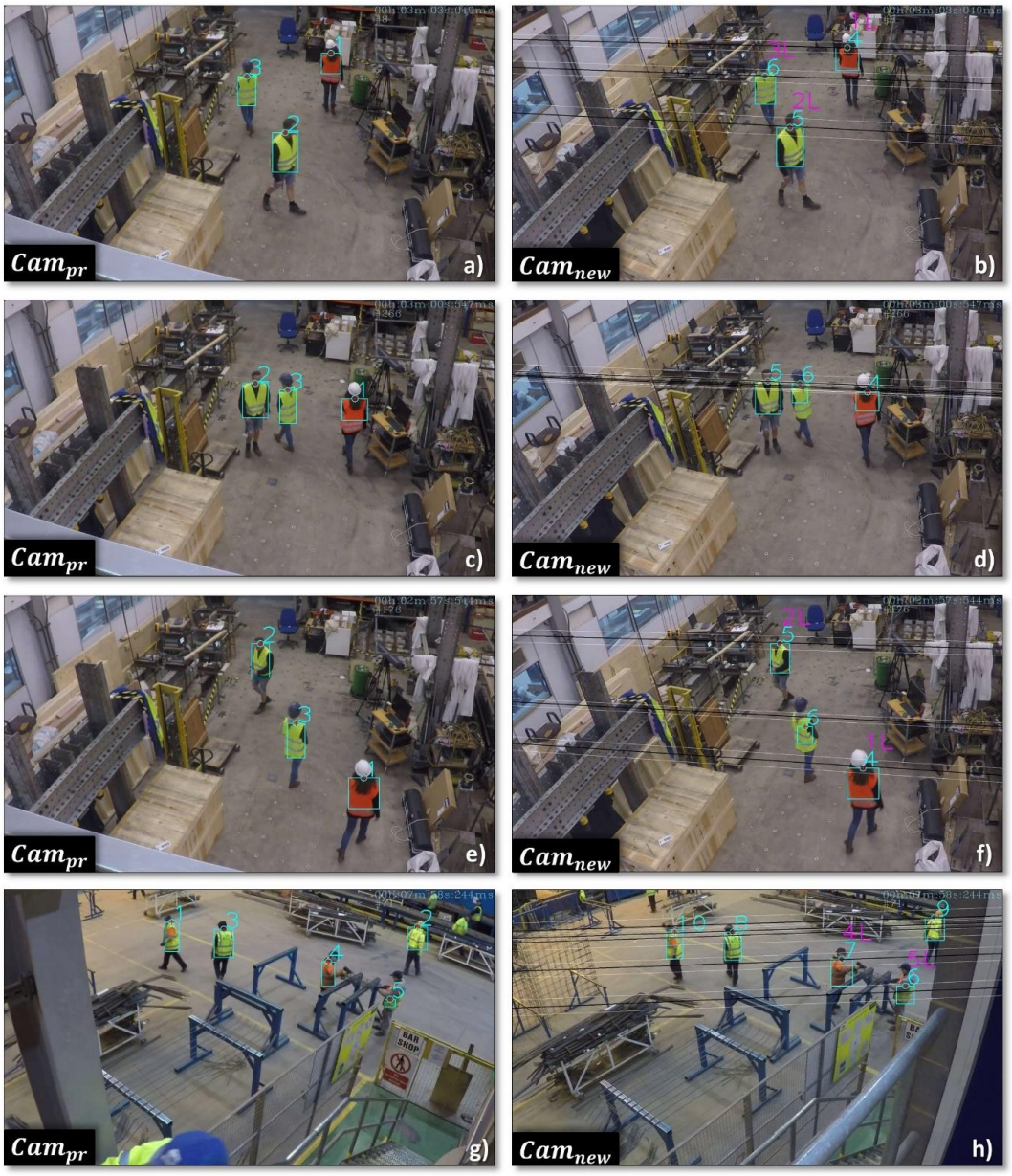

Figure 4-25: Performance examples of the geometry-based matching method. (a-b) Three TP matches, (c-d). Three TN matches due to congestion. (e-f) Two TP matches and one FN caused by tracking inaccuracy (Data set A). (g-h) Three TN matches due to congestion and two TP (Data set B).

Figure 4-25 shows representative performance examples. In these examples, white lines represent the matching search band, whilst black are the epipolar lines. In Figure 4-25 (a-b) candidates with IDs "1-3" are correctly matched (TP) to workers with IDs "4-6". However, in Figure 4-25 (c-d) the matching approach fails due to congestion and returns three TN. The same holds for candidates of data set B in $\mathrm{Cam}_{\mathrm{pr}}$ with IDs "1-3" (see Figure 4-25 (g-h)). All these TN matches result when search bands encompass more than one candidates. Figure 4-25 (e-f) illustrates two TP matches and one FN. The latter is due to the large $error_f$ of the worker with ID "6" in $\mathrm{Cam}_{\mathrm{new}}$.

4.6.3. Evaluation of the template-based matching method

This last matching method compares workers using their colour templates. The HSV colour space is selected for this purpose due to observations of previous Chapter 3 (see section 3.4.2). The template-based matching method uses as input the colour templates of workers. Minor dissimilarities like posture and appearance variations (e.g. length of sleeves, open jacket) are exploited in this step. Figure 4-26 depicts such examples. In this figure, two workers that wear Hi-Vis vests of the same colour but trousers and hard hats of different colour in (a), and two workers with Hi-Vis vests and hard hats of different colour in (b), are distinguishably described though the HSV colour space. In these examples "p" stands for Cam_{pr} and "n" for Cam_{new}. This method can be useful even if it has significant restrictions when matching is required among congested candidates with no motion. Such scenario occurs when tracking re-initializes after a sudden termination due to problems such as occlusion and posture changes.

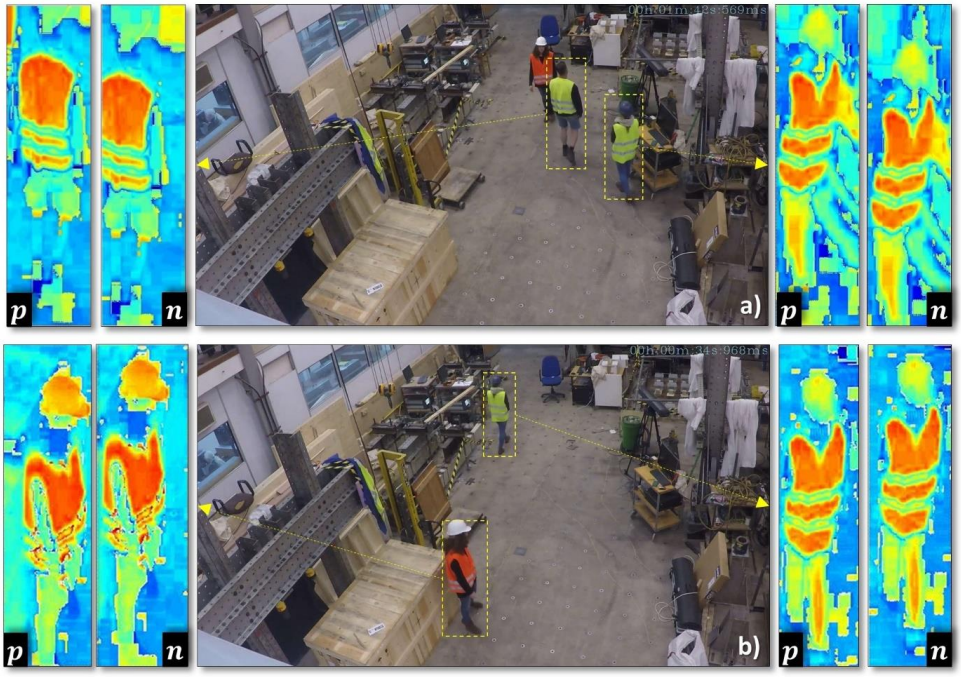

Figure 4-26: Illustration of compared colour templates with the HSV colour space. (a) Workers with same colour of Hi-Vis jacket. (b) Workers with different colour of Hi-Vis jacket.

Table 4-5 shows that the proposed template-based matching method performs better for the workers of data set A (see (a-b) in Figure 4-27) than for the workers of data set B (see (c-d) in Figure 4-27). This is because data set B has a bigger proportion of similar in appearance targets (p=80%). Overall, 385 targets for data set B and 351 for data set A (p=67%) are processed. Table 4-5 shows that precision drops by 8%, whilst recall and accuracy by 6% and 12% respectively. This is due to the almost identical appearance of workers of data set B that increased the total number of FP matches.

Table 4-5: Evaluation of proposed template based matching method.

Performance Metrics	**HSV**	
	Data Set A	**Data Set B**
Precision	0.96	0.88
Recall	0.95	0.89
Accuracy	0.91	0.79

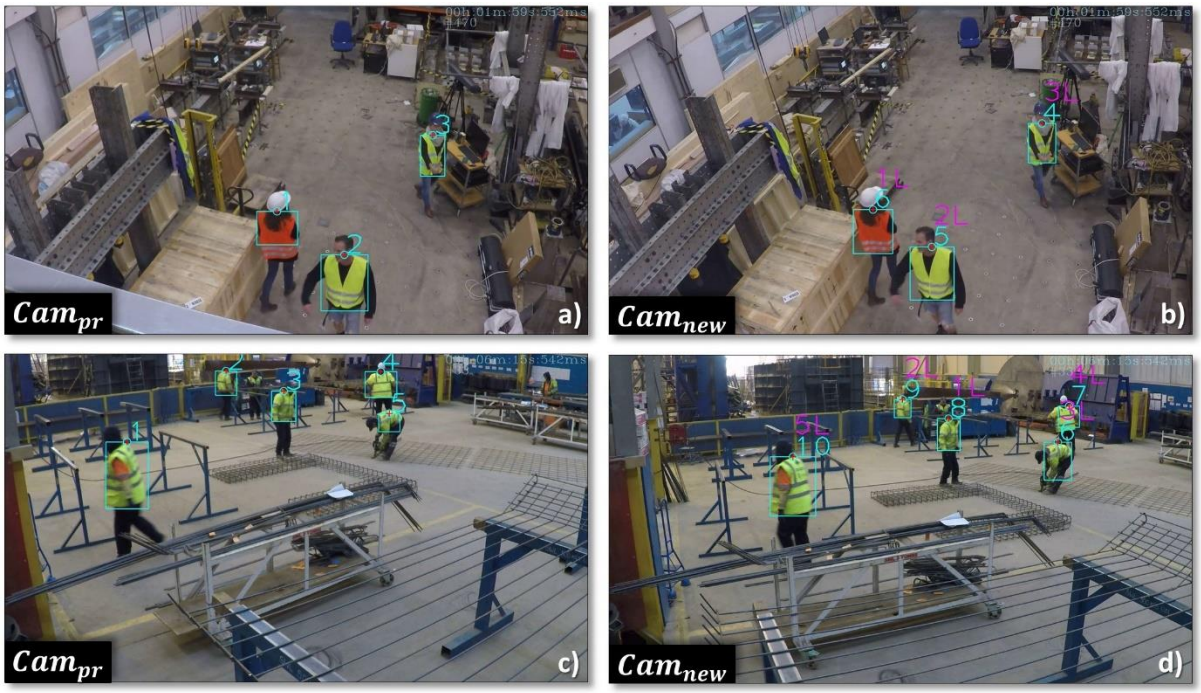

Figure 4-27: Performance examples of the template based matching method. (a-b) Three TP matches (Data set A). (c-d) Three FP and Two TP matches (Data set B).

4.6.4. Evaluation of overall proposed matching method

The previous 4.6.1-4.6.3 sections confirm the hypothesis of this chapter that the motion-based matching method is the most efficient for matching construction workers. Second follows the geometry and third the template-based method. Each of them is robust under specific conditions. Hence, their performance depends on the appearance frequency of these conditions. The two best methods, motion and geometry, are separately tested under congestion, posture/scale variations, appearance similarity and occlusions.

In total 252 pairs of stereo frames from data set B, capturing 1122 tracked targets in both cameras are processed. The sample size is defined by the proportion (p) of the tracked targets of data set B that face congestion and occlusions, have no motion, and share similar appearance. This proportion is measured equal to 53%. Table 4-6 shows that, the performance of motion and geometry-based methods drops significantly under multiple challenges in terms of recall and accuracy compared to the overall method that features 97% precision, 98% recall, and 95% accuracy. This validates the previous hypothesis that none of these methods is robust under all matching challenges at the same time.

Table 4-6: Evaluation of motion and geometry matching method under congestion, posture and scale variations, appearance similarity, and occlusions (data set B).

Performance Metrics	**Only Motion**	**Only Geometry**	**Overall Method**
Precision	0.97	0.93	0.97
Recall	0.60	0.43	0.98
Accuracy	0.59	0.42	0.95

Table 4-7 describes in detail the exact number of the total: a) correctly matched targets (TP), b) incorrectly matched (FP), c) incorrectly not matched (FN), and d) correctly not matched (TN) of the overall proposed method. The low total of FP compared to the high total of TP proves the good performance of the proposed method, whilst the low totals of FN and TN show the completion achieved.

Table 4-7: Confusion matrix of the overall proposed method.

		Predicted			
		Yes	**No**		
Actual	**Yes**	TP	FN	1065	20
	No	FP	TN	35	2

Finally, Figure 4-28 illustrates some representative examples of how the three matching methods supplement each other. Each of them corresponds to a different colour. Green stands for the motion-based method, orange for the geometry and pink for the template-based. In (a-b) of Figure 4-28, the two walking candidates with IDs "1" and "4" in Cam_{pr} are matched with motion whereas the two standing still candidates with IDs "2" and "3" are matched with the geometry and template respectively. The same holds for the examples presented in (c-f) of Figure 4-28.

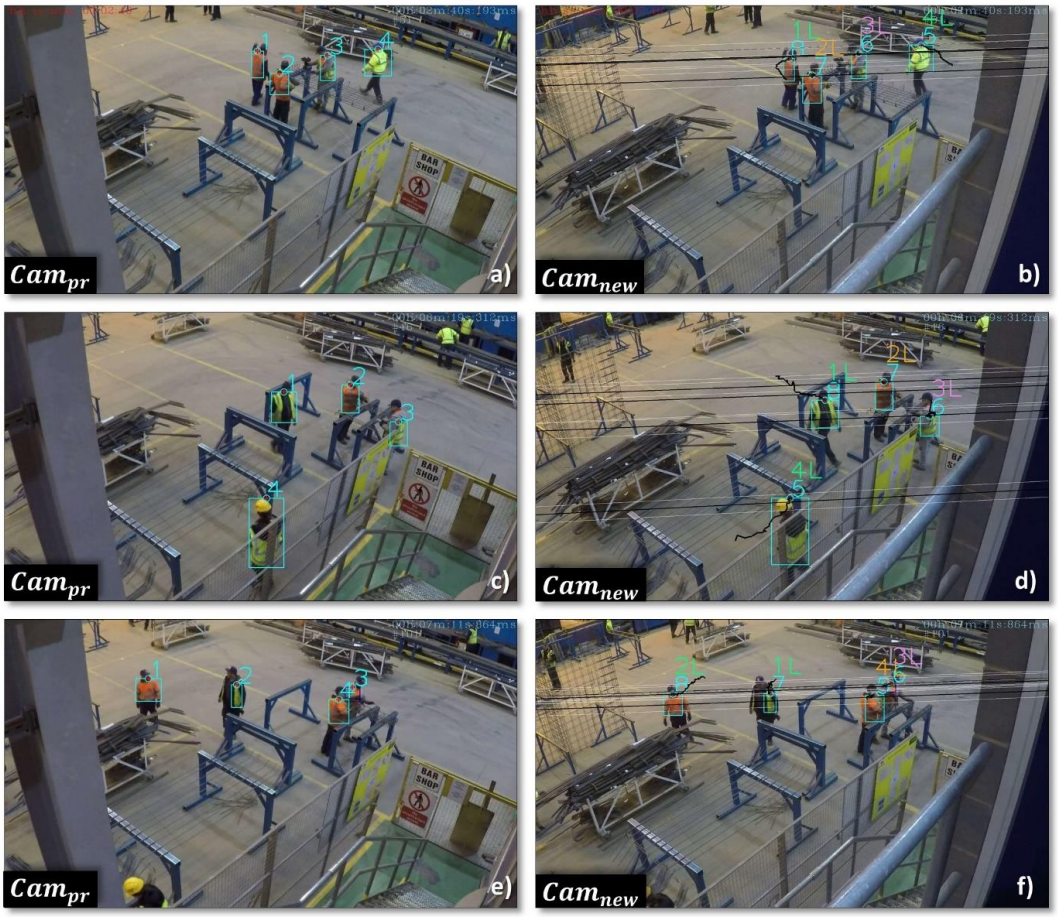

Figure 4-28: Performance of the overall vision-based matching method under congestion, appearance similarity and occlusions (green labels: motion based matching; orange labels: geometry based matching; pink labels: template based matching).

4.6.5. Accuracy of 3D trajectories

Finally, the accuracy of the extracted 3D trajectories is evaluated by using the upper middle point of tracking's bounding box for triangulation. For the evaluation, the extracted ground plane coordinates (X, Z) are exploited. The ground truth D_{real} trajectory is manually marked with yellow chalk on the floor and is equal to 6.4m (see Figure 4-29). The worker with orange jacket from data set A is tracked while walking along this line. The extracted trajectory D_{calc} is illustrated in Figure 4-30. Firstly, a line is fitted to D_{calc} in order to measure the total length with Euclidean distance which is then subtracted from the ground truth. The resulted distance error is equal to 13cm. This error falls within satisfying accuracy limits as it is smaller than the 15cm error of tagged 3D tracking methods (Chawla et al., 2010).

Figure 4-29: Ground truth trajectory.

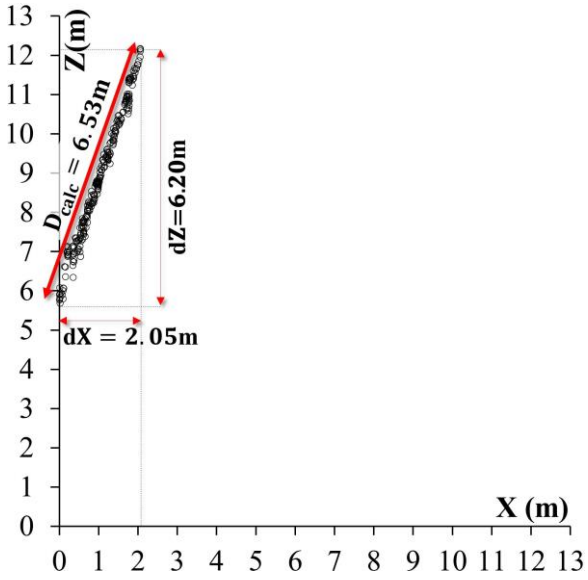

Figure 4-30: Euclidean calculation of triangulated trajectory.

4.7. Chapter overview

The current state of practice for inter-camera tracking of construction resources is time consuming and labour intensive. This is due to the manual effort involved in matching the same targets across multiple camera views. This matching is essential in order to turn the 2D image trajectories (x, y) into 3D trajectories (X, Y, Z). Up to date the current state of research has not managed to solve the matching problem for targets that: a) share greatly similar appearance such as workers, b) are under occlusions, c) appear with variant posture and scale, and d) are within a congested environment. This chapter presents a computer vision-based method for matching the same workers across all available overlapping cameras views. The novelty of the method presented in this chapter lies in the sequential

combination of three methods, each of which exploits unique features in order to achieve matching of the same workers across different cameras.

Firstly, a computer vision-based 2D tracking method that follows each worker across subsequent frames of the same camera is implemented. This tracking provides the input data of the proposed method. The first step tests the hypothesis that a small segment of previous motion data is enough for distinguishably describing a worker. This step uses a motion-based matching method that compares workers' 2D trajectories over the last 1500ms. If this fails to return a positive match for all candidates, then the second step that employs geometric restrictions in order to predict the area where a possible match lies is activated. If more than one possible candidates lie within the same search band then this geometry-based method returns no match. When this occurs the third step compares workers based on their colour templates. This method uses only the templates as enclosed by the bounding box of the vision tracking method for computational efficiency.

Each method performs well separately only under specific conditions. The two first steps employ methods that are invariant to posture and scale variations, heavy occlusions, and appearance similarity compared to the method of the third step. However, the latter performs better under congestion and lack of motion. The motion-based method features 97% precision, 99% recall, and 98% accuracy for tested walking workers. The geometry-based method achieves 93% precision, 94% recall, and 95% accuracy in a non-congested environment. Lastly, the template-based method returns 88% precision, 89% recall, and 79% accuracy for greatly similar workers which are captured under the same posture.

The performance of the two best performed methods under all conditions at the same time is also evaluated. The resulting performance of the motion-based method is significantly lower with 97% precision, 60% recall, and 59% accuracy. The same holds for the geometry-based method that scores 93% precision, 43% recall and 42% accuracy. On the other hand, for the same conditions the overall proposed method features 97% precision, 98% recall, and 95% accuracy. This proves that the hypothesis of this chapter that only the combination of all methods can efficiently tackle all matching challenges is correct.

The limitations of the research presented in this chapter are the following: a) all tests were conducted with data from indoor working environments due to the nature of the construction sites the researchers had access to, b) all data were collected from only a set of cameras (stereo camera), and c) the proposed method is designed only for cameras with overlapping views.

5

Detection of work cycles for monitoring labour productivity

In this chapter, a trajectory analysis-based method for monitoring the labour productivity of construction workers is presented, using as inputs the outputs of Chapters 3 and 4. It detects repetitive patterns in trajectories of workers. These patterns depict work cycles. The total duration of these work cycles is equal to the labour input of every worker on a task level. Labour productivity is calculated by dividing the total labour output over the total input. The method presented in this chapter focuses on the input as the monitoring of the output is quite straight forward through visual inspection at the end of work shifts (e.g. number of steel cages prepared, meters of brick wall painted). The aim of this chapter is to monitor the labour productivity of construction workers accurately and proactively.

5.1. Introduction to trajectory analysis for pattern recognition

The prevalent approach to estimating labour productivity is through an analysis of the trajectories of the construction entities. This analysis typically exploits four types of trajectory data: a) walking path trajectories, b) dense trajectories (posture), c) physiological rates such as heart rate (beats/minute) and respiratory rate (breaths/minute), and d) sound signals. The output of this analysis is the number of work cycles performed by construction workers. The total duration of these cycles is equal to the labour input of a task. Chapter 2 concludes that all methods proposed so far do not meet the requirements for proactive monitoring of labour productivity in an accurate, non-obtrusive, time and cost efficient way for multiple workers.

Several studies have focused on trajectory analysis in order to detect trajectories that depict strange behaviour profiles, for security purposes (Junejo & Foroosh, 2008; Wiliem et al., 2008). The trajectories of people are compared to those that correspond to normal behaviour to achieve this. After comparison, trajectories are grouped into clusters of similar motion patterns. Figure 5-1 illustrates an example of clustered trajectories based on their spatial similarity. The trajectories in this figure belong

to people that were monitored while walking along the corridor of a shopping mall. The trajectories in (a-h) of Figure 5-1 depict normal behaviour, whilst the trajectories in (i-l) of Figure 5-1 belong to abnormal behaviours.

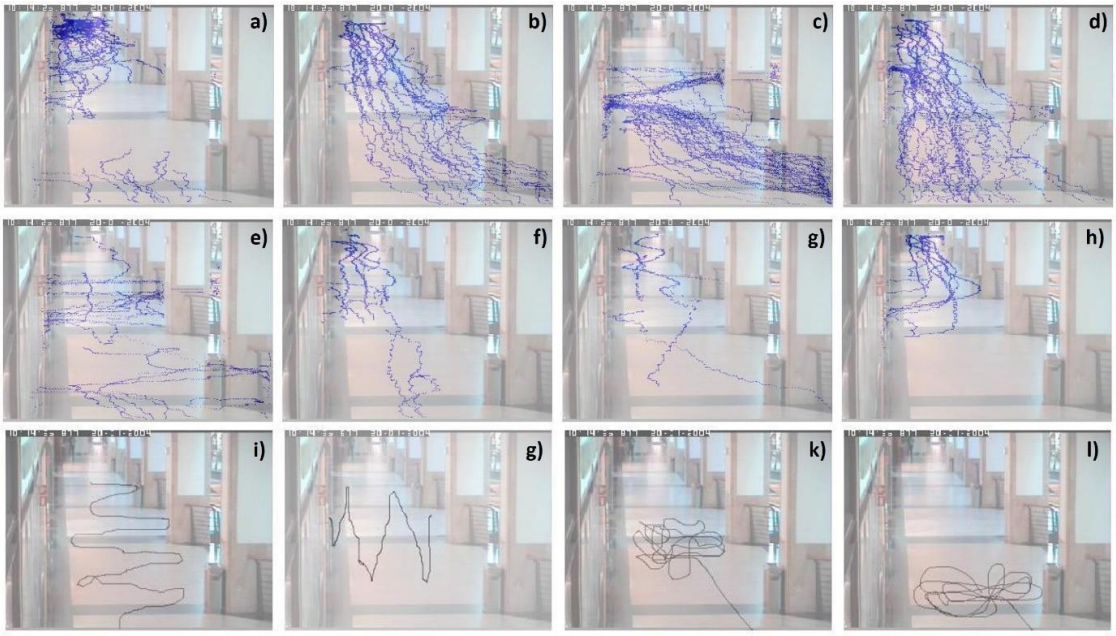

Figure 5-1: Trajectory clustering for detecting abnormal human behaviour (Wiliem et al., 2008). (a - h: trajectory clusters of normal behaviour, i-l: trajectory clusters of abnormal behaviour)

The Hausdorff distance (Rote, 1991) metric is one of the most commonly used functions for calculating the spatial similarity between trajectories with multiple sample points and different lengths (Junejo & Foroosh, 2007). Trajectories are treated as sets of points. In more detail, the Hausdorff distance for two trajectories (T_A, T_B), is equal to the maximum distance of trajectory T_A or T_B to the nearest point of trajectory T_B or T_A respectively (see Figure 5-2). Hence, such spatial proximity-based clustering implies that monitored targets should move along the same designated paths. This is feasible in shopping malls or parking lots but not at construction jobsites as in such environments, walking paths change over time as the project progresses.

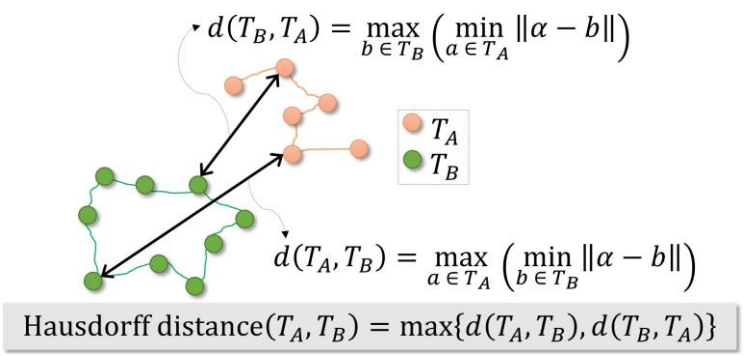

Figure 5-2: Hausdorff distance between trajectories T_A, T_B.

Researchers have included temporal features, such as speed, in clustering in order to overcome the limitations of spatial proximity-based clustering methods. In this respect, a Hidden Markov Model (HMM) (Fosler-lussier, 1998), was trained to cluster trajectories given their spatiotemporal features i.e. speed and position (Suzuki et al. 2007). This training process calculated the probability of a person moving in different state sequences. Each of these sequences consists of all possible transitions from state to state for all hidden states of a Hidden Markov Model. However, these states are not really hidden. In fact, they depict real positions as illustrated in the example of Figure 5-3.

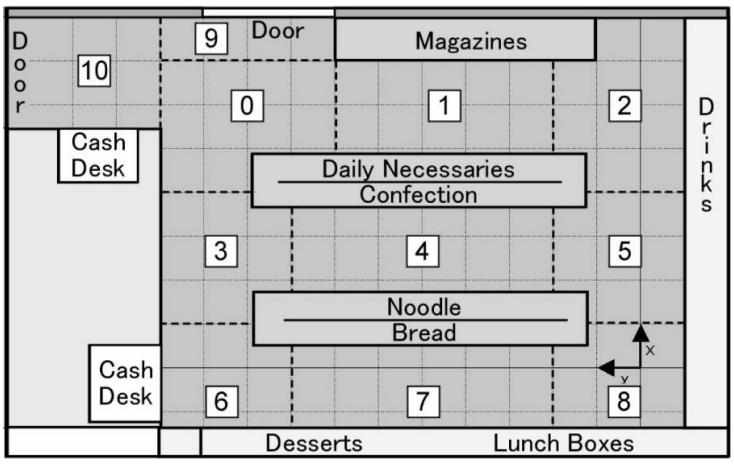

Figure 5-3: Hidden states (1-10) of an HMM within a store (Suzuki et al., 2007).

In the above figure, each state stands for a semantically important area within a store e.g. magazines area, noodles area. The trajectories of people are grouped into clusters based on their minimum distance from each cluster's centroid. The trajectories of a cluster were classified as abnormal behaviour if the maximum likelihood of this cluster's HMM was less than an empirically defined threshold. This type of clustering requires that "hidden" states are known (e.g. entrance, cashier, exit). Such states for construction are the work zones e.g. steel fixing zone, concrete pouring zone, materials storage zone, rest areas and hoisting zone. However, the locations of these work zones ("hidden" states) change over time. This is due to the dynamic nature of construction projects. For instance, a brick layering task on day "1" is described by three "hidden" states (A, B, and C). "Hidden" state A is the area where bricks are stored. "Hidden" state B is the area where brick walls are constructed. "Hidden" state C is the exit areas. When the construction of brick walls at "hidden" state B is finished on day "2", then the workers will move to another "hidden" state D in order to continue with the construction of another wall. If this new "hidden" state D is positioned on another floor level or far from the previous "hidden" state B, then most likely "hidden" states A and B will also be repositioned in order to be closer to the area where the new walls are constructed ("hidden" state D). Therefore, as obtaining prior knowledge on the "hidden" states is not generally feasible, HMMs cannot be applied to clustering workers' trajectories.

In summary, this section presents studies that compare clusters of trajectories in order to detect the ones that most likely depict abnormal motion patterns. The main limitation of such studies in terms of detecting work cycles instead of abnormal patterns, relates to the way trajectories are clustered. Hence, this chapter aims to answer the following question: "*How can trajectory data be efficiently clustered into work cycles in order to measure the labour input*?"

This chapter presents a semantic-based method for clustering workers' trajectories into work cycles regardless of the type or the numbers of tasks the workers perform. These trajectories are four-dimensional (4D) and describe the motion of workers across the jobsite over time. They are segmented into 4D sub-trajectories which are then classified as either "move" or "stop" semantic events. The "move" events describe workers while transitioning between sub-tasks, whilst the "stop" events depict workers while performing sub-tasks. A task consists of sub-tasks. For instance, the steel fixing task consists of the following sub-tasks: placing, fixing, and picking of reinforcement bars (re-bars). Each worker's classified 4D sub-trajectories are clustered together under the main assumption presented in Chapter 2 in order to detect work cycles. The total duration of these work cycles is equal to the labour input of workers. Workers while at "stop" are relatively still. This spatiotemporal stability combined with the presence of workers at pre-scheduled work zones imply productive time. The following Chapter 6, classifies these work cycles in order to achieve proactive monitoring of labour productivity.

The remainder of this chapter is organized as follows. Sections 5.1 and 5.2 present the current state of research in trajectory analysis-based methods for pattern recognition. Then, the proposed solution and methodology are presented in sections 5.3 and 5.4 respectively, and evaluated in section 5.5. Lastly, section 5.6 presents an overview of the conclusions of this chapter.

5.2. Cluster analysis outline

Machine learning methods are divided into supervised and unsupervised methods. The supervised methods use a sample of labelled data to "teach" the relationship between inputs and outputs. On the other hand, unsupervised methods use unlabelled data to "learn" to recognize similar objects without any notion of the output. Such methods are the clustering methods. Their goal is to group similar objects into clusters. Unique features of objects are exploited in order to achieve this. Each object is described by an observation of features $\{x_i\}$. These features form an N-dimensional feature vector $\vec{x} = \{x_1, x_2, \dots, x_N\}^T$(see Figure 5-4). This section presents an overview of existing clustering methods in order to identify the most suitable for grouping trajectories that depict similar construction-related patterns (i.e. tasks or sub-tasks). This review specializes on workers. Therefore, the suitability of the clustering methods reviewed in this section is evaluated based on the characteristics of workers' trajectories. The specific characteristics which are taken into account are the following: a) a worker's trajectory is of high dimensionality as he/she moves across the jobsite over a large period of time (≥

work shift), b) several workers' trajectories should be processed as construction jobsites are congested environments where numerous tasks take place simultaneously, and c) the trajectories of workers depict multiple different patterns that should be grouped in an equal number of clusters. The latter is due to the large variety of construction tasks and the fact that even the trajectory of one worker consists of multiple patterns as he/she is involved in many different tasks within a single work shift. In summary, an efficient clustering method should be able to deal with data sets (i.e. objects) of large dimensionality (i.e. long feature vector $\vec{x}$) and high diversity (i.e. many clusters).

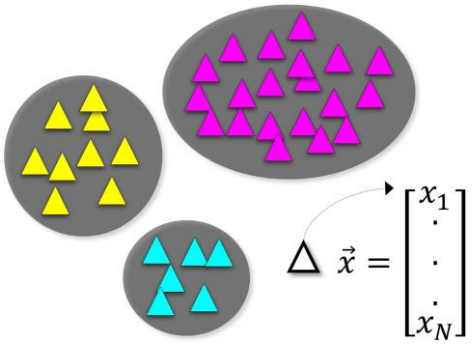

Figure 5-4: Cluster analysis of objects into three groups (clusters).

The similarity between objects is calculated through distance-based measure functions. Some of the most common are presented below:

- The Euclidean distance: $d(\vec{x},\vec{y}) = \|\vec{x}-\vec{y}\| = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$
- The Manhattan distance: $d(\vec{x},\vec{y}) = \sum_{i=1}^{N}|x_i - y_i|$
- The Minkowski distance: $d(\vec{x},\vec{y}) = \|x-y\|^{\alpha} = \left(\sum_{i=1}^{N}(x_i - y_i)^{\alpha}\right)^{1/\alpha}, for\ \alpha > 1$
- The cosine distance: $d_{\theta} = \frac{\sum_{i=1}^{N} x_i\, y_i}{\|x_i\|\|y_i\|}$

  where $\theta = \tan^{-1}\frac{\|\vec{x}\|}{\|\vec{y}\|}$
- The Mahalanobis distance: $d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \sum^{-1}(\vec{x}-\vec{y})}$

  where $\sum = covariance\ matrix$
- The Locality In-between Polylines (LIP) distance: $d(\vec{x},\vec{y}) = \sum_{i=1}^{N} Area_i * w_i$

  where N = total intersections between trajectories $(\vec{x},\vec{y})$, and $w_i = \frac{Length_{\vec{x}}(l_i,l_{i+1}) + Length_{\vec{y}}(l_i,l_{i+1})}{Length_{\vec{x}} + Length_{\vec{y}}}$

The distance-based measure functions are chosen according to the type of objects to be clustered. The Euclidean, Manhattan and Minkowski measures are mainly used to compare objects that are described by feature vectors of low dimensionality. This is because the distance between two objects increases along with their dimensionality. Hence, two objects of different dimensionality (low vs high) but of the same type, score a large (i.e. dissimilar) rather than a low similarity value (similar). As a result of this,

these objects of the same nature will be wrongly assigned to different clusters. This problem can be alleviated by the cosine distance measure as it normalizes the features of the objects to a common range. The Mahalanobis distance measure is also invariant to the dimensionality of objects since it exploits the Gaussian distribution of objects. It computes the distance of an object x to an object y given their covariance matrices Σ. Distance-based measures for specifically calculating similarities between trajectories have also been proposed. The Locality In-between Polylines (LIP) is such a measure (Pelekis et al., 2007). It treats trajectories as polylines and measures their similarity by summing up the areas formed between intersections of points of the compared trajectories (see Figure 5-5).

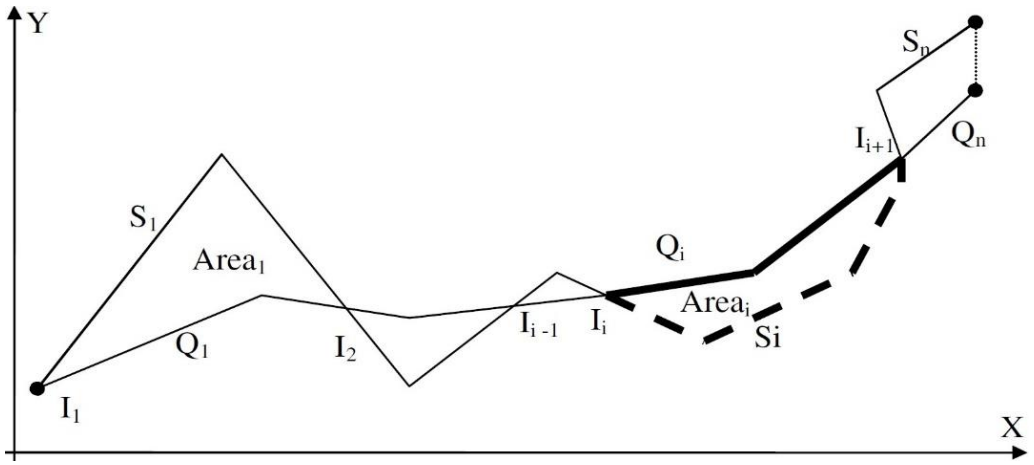

Figure 5-5: Locality In-between Polylines (LIP) distance of trajectories (Q, S) (Pelekis et al., 2007).

Figure 5-6 illustrates with two-dimensional examples the different types of clusters which are categorized as follows:

a) Well-separated clusters: All objects within a cluster have the smallest distance between them rather than with any object from another cluster.
b) Centre-based clusters: All objects within a cluster have the smallest distance from their cluster's centre rather than with any other cluster's centre.
c) Contiguous clusters: All objects within a cluster have the smallest distance to at least one object within its cluster rather than with any other object from other clusters.
d) Density-based clusters: Regions with high density as compared to their surroundings are grouped into clusters.
e) Conceptual clusters: All objects within a cluster are described by a general property.

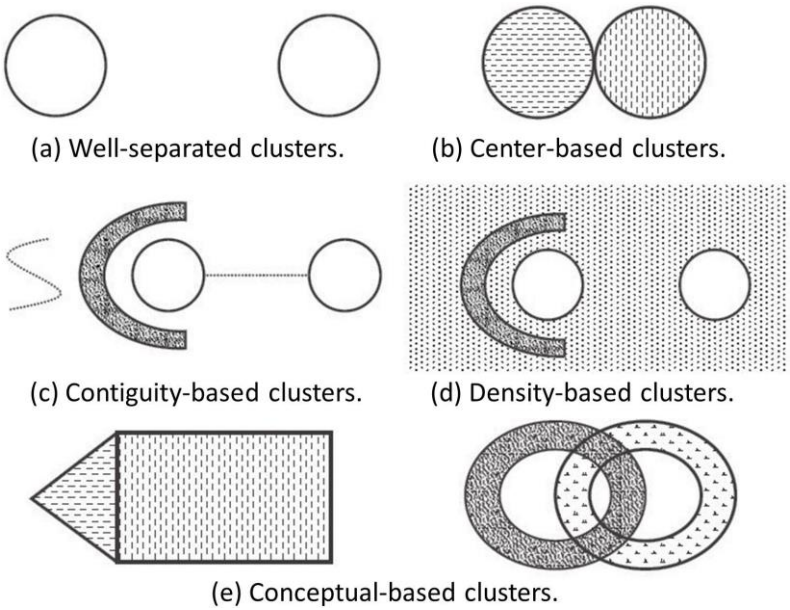

Figure 5-6: Types of clusters (Tan et al., 2005).

The advantages and disadvantages of existing clustering methods are presented below (Madhulatha, 2012; Omran et al., 2007) :

- Hierarchical clustering: Extracts clusters that can be further divided into more than one sub-cluster. The methods in this category can be either agglomerative (bottom-up) or divisive (bottom-down). The former assigns a cluster to each object and then successively merges the initial clusters into larger ones whilst the latter starts with a single cluster that contains all objects and gradually divides it into smaller clusters. Such methods are not efficient in clustering large data sets (i.e. many objects). This is because the computational and time complexity of diving clusters into subsets grows along with the size of the data. In addition, the clustering process is not reversible. Hence, there is no possibility to correct the clustering of an object in future steps.
- Partitional clustering: Assigns the objects into a predefined number of clusters at once. The k-means algorithm (Hartigan & Wong, 1979) is a representative example of this category. Such clustering methods are computationally efficient with large input data, but are limited by the requirement of manually providing the total number of clusters in advance.
- Density-based clustering: Assigns into clusters regions of arbitrary shape that have much higher density than their surroundings. Such methods are efficient in discarding noise. This is because noise has low density. However, they can only group objects with similar density into clusters. The DBSCAN algorithm (Bäcklund et al., 2011) is a representative example of this category.
- Grid-based clustering: Divides the space around the objects into a grid of cells. Then, objects are assigned to cells and form clusters if the density within each cell exceeds a threshold. The adjacent clusters are all then merged together into bigger ones. The main limitation of such methods is that the boundaries of the clusters are not flexible. They are of rectangular shape only. Because of this

restriction, the input data should be either vertically or horizontally separable. Otherwise, they will be either wrongly assigned to clusters or not assigned at all.

- Model-based clustering: Clusters in this category are governed by Gaussian distributions. An object is assigned into the cluster that scores the highest probability of being part of it. Such methods are inefficient when input data do not fit to the Gaussian model.

Another main issue of existing clustering methods is that if the common patterns of trajectories "hide" within smaller segments, then they fail to group these trajectories into the same clusters (Lee et al, 2007). This is due to the fact that clustering methods treat trajectories as objects and cluster them as a whole. In an effort to alleviate this shortcoming, previous studies introduced the partitioning of trajectories into smaller sub-trajectories before performing clustering. Partitioning methods are either based on criteria such as preciseness and conciseness (Lee et al., 2008; Lee et al., 2007) or on changes of spatiotemporal features i.e. speed and position (Liu & Schneider, 2012; Sung et al., 2012).

The former partitioning methods that use preciseness and conciseness, apply the minimum description length principle MDL (Grunwald et al. 2004) that calculates the best combination between the number of partitions (conciseness) and the discrepancy of the partitions as compared to the original trajectory (preciseness). These two criteria are contradictory to each other. The number of sub-trajectories p_i that the partitioning of a trajectory TR_i produces is equal to the value that minimizes the cost between these two criteria (see Figure 5-7).

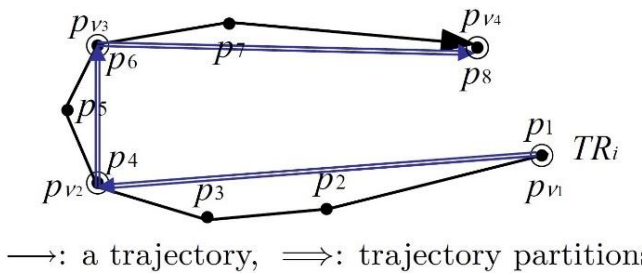

Figure 5-7: Partitioning of trajectories based on preciseness and conciseness (J. Lee et al., 2008).

The latter partitioning methods that rely on spatiotemporal features, divide trajectories into partitions every time the target changes either its direction or speed. Manually defined speed and directional thresholds (see (a-b) in Figure 5-8) (Liu & Schneider, 2012) and line simplification algorithms (see (c-d) in Figure 5-8) are exploited in order to detect these spatiotemporal changes. The main problem with partitioning of trajectories is that the extracted sub-trajectories increase even more the total number of input data. However as previously mentioned, such increase is not desirable as existing clustering methods struggle under large input data.

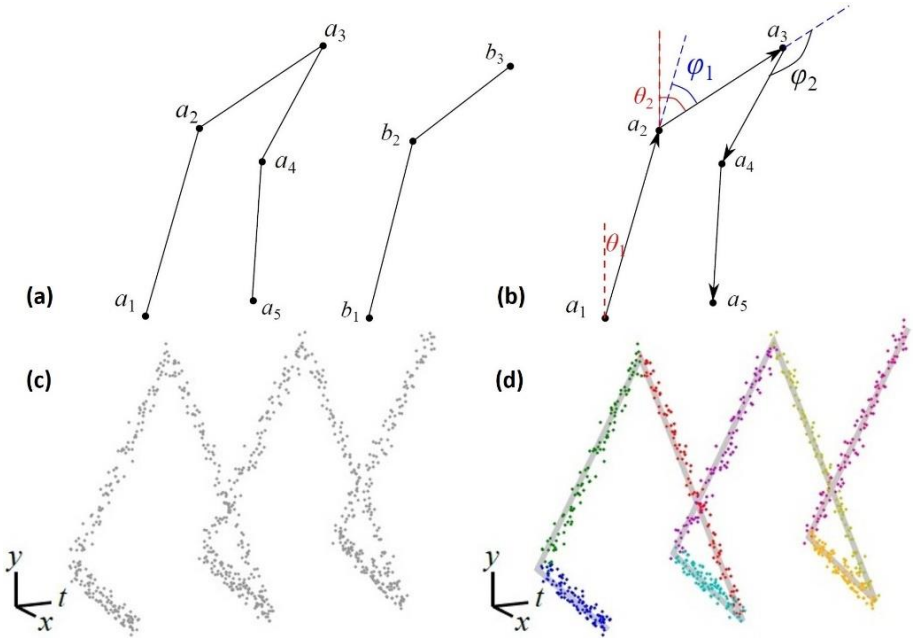

Figure 5-8: Partitioning of trajectories based on spatiotemporal changes.
(a-b: Liu & Schneider, 2012, c-d: Sung et al., 2012).

Semantic analysis was also introduced for clustering trajectories (Palma et al., 2008; Pasha & Monajjemi, 2013). These studies considered the semantic importance of "stops" and "moves" areas with the aim of localizing semantically important sub-trajectories. They used intersections of trajectories with known geographically important places provided by the user (A, B, and C regions in (a) of Figure 5-9), and lower speed values (X, Y regions in (b) of Figure 5-9). Trajectories were then clustered into similar groups using a density-based clustering method (DBSCAN), on the assumption that semantic "stops" are regions with high density of small speed values as the driver slows down to look at something interesting. The parts of trajectories that were not classified as "stops" were automatically considered as "moves" clusters depicting moving targets. This analysis-based clustering method is a promising technique for deriving semantically important clusters, and is invariant to both initialization parameters and data size. However, it is still restricted by the limitations of the applied density-based clustering method.

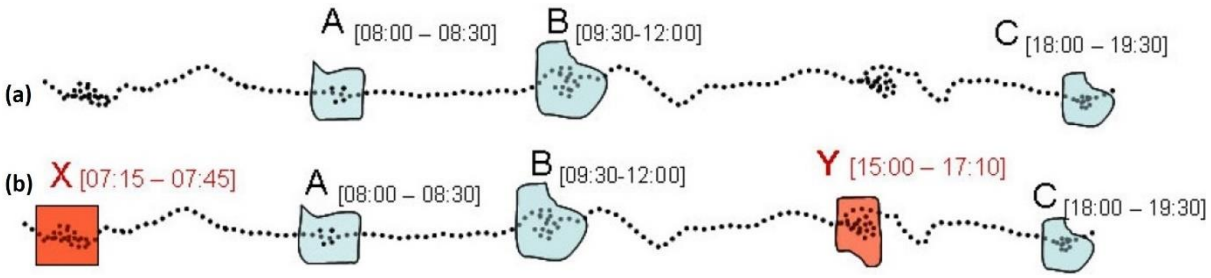

Figure 5-9: Semantic stop regions of trajectory data (Palma et al. 2008).

In summary, existing clustering methods still suffer from two main disadvantages in terms of detecting patterns (i.e. clusters) from trajectory data. Both of them relate to the large dimensionality and high diversity of worker trajectories as previously explained. Firstly, trajectories should be segmented into sub-trajectories. This is necessary as patterns "hide" in smaller segments of workers' single trajectories. When this occurs, one single trajectory (i.e. object) is converted into multiple smaller trajectories (i.e. objects). However, clustering methods as previously mentioned struggle with large data sets (i.e. objects). Secondly, clustering methods are often reliant on user-provided initialization of parameters such as the number of clusters or the threshold values of similarity metrics. Such estimations are not always possible due to the dynamically changing nature of construction tasks. For instance, a steel fixing task can either be described by two patterns (i.e. clusters) that depict fixing and picking of re-bars or by even more patterns if the worker undertakes extra tasks e.g. assists a colleague, visits the foreman for further instructions etc. Therefore the user needs to know in advance all types of tasks the workers performed to explain the high diversity (i.e. multiple clusters) resulting from their trajectories.

The main objectives of this chapter are: a) to develop a method that groups trajectories of workers into clusters of similar pattern, and b) to devise a method invariant to the large variety of construction tasks. The research question that this chapter aims to answer is: "*how can the trajectory data of a single worker be translated into labour input accurately*?"

5.3. Proposed solution

This section presents a method that turns the trajectory data into labour input. Firstly, the work cycles of construction workers are detected regardless of the type or the number of tasks they perform. The flowchart of Figure 5-10 illustrates the overall proposed method to achieve this. The skewed parallelogram shapes refer to processes and the circular to inputs/outputs. The initial inputs of the proposed method are the 4D trajectories of workers which are extracted by a 4D computer vision-based tracking method as shown in Chapters 3 and 4. A 4D trajectory T_w depicts the motion of a single worker across the jobsite over time. This motion is described by the X, Y, and Z Cartesian coordinates. Hence, it can be decomposed into three time series XT, YT and ZT. In this chapter, the X and Z axis are aligned with the floor plane of a jobsite whilst the Y axis is vertical to the floor plane. Initially, the 4D trajectories are smoothed to remove noise. Then, each single smoothed 4D trajectory T_w of a worker is partitioned into N smaller sub-trajectories $S_w = \{s_i\}_{i=1\dots N}$. This allows the proposed method to search for work cycles that correspond to smaller segments of the 4D trajectories of workers. The partitioning is time-based. A representative 3D speed value is then calculated for each 4D sub-trajectory s_i. It depicts the speed of workers along the floor $\|\vec{V}_{xz}\| = \{\|\vec{v}_i\|_{xz}\}_{i=1\dots N}$ and the vertical $\|\vec{V}_y\| = \{\|\vec{v}_i\|_y\}_{i=1\dots N}$ plane. The 4D sub-trajectories are classified as either part of "stop" or "move" semantic events based on these 3D speed values. This classification relies on the assumption presented in Chapter 2 that *if a*

worker's "move" is followed by one "stop" and a second "move" sequentially, then these three semantic events define a work cycle. The semantic "move" event depicts the transition between sub-tasks or tasks whereas the semantic "stop" event depicts the actual execution of a sub-task. Finally, the classified 4D sub-trajectories are clustered into work cycles. The time a worker spends on a task (labour input) is equal to the total duration of these cycles. This type of semantic-based clustering overcomes the disadvantages of existing clustering methods.

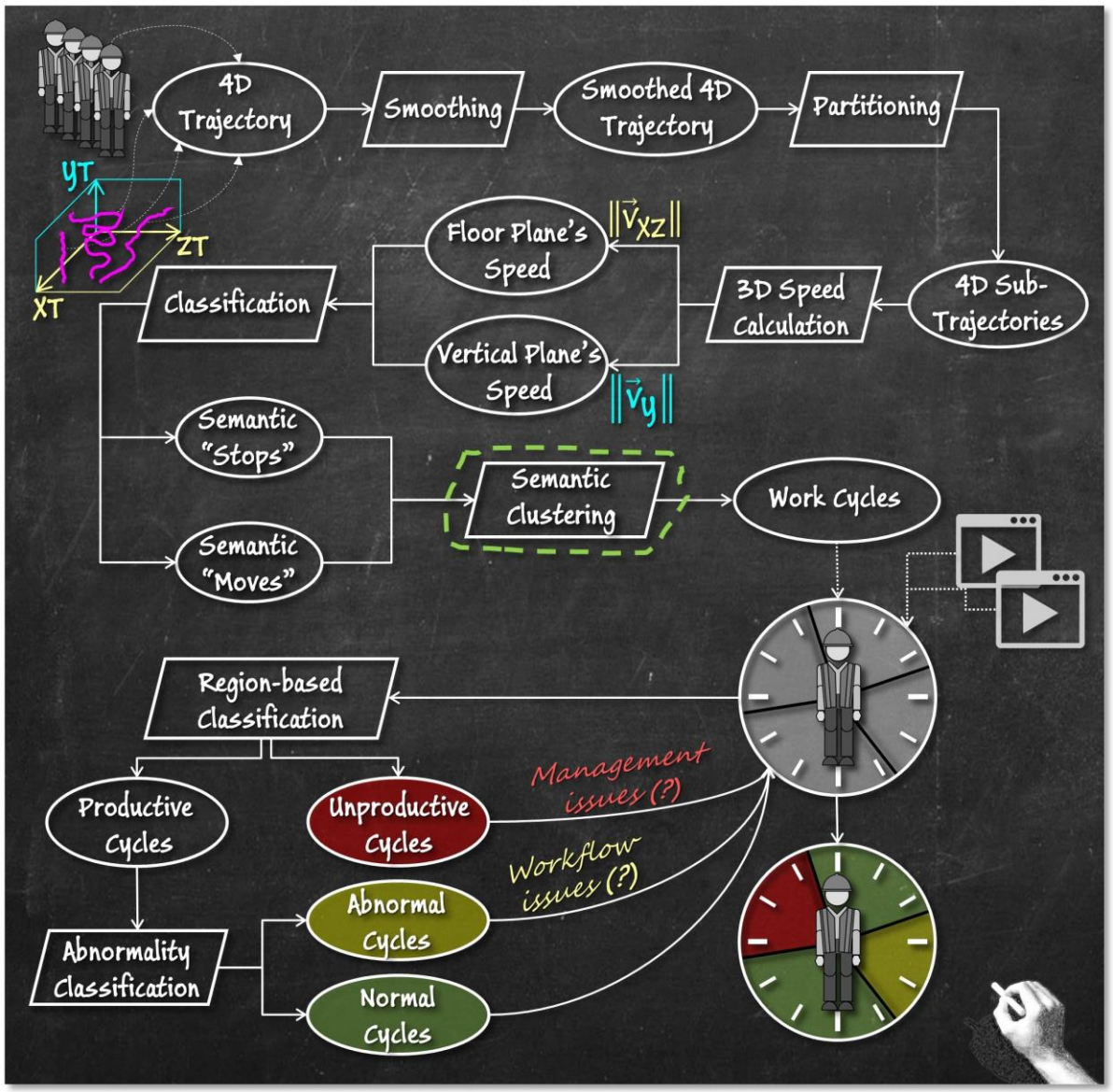

Figure 5-10: Flowchart of proposed method for monitoring the labour productivity of workers.

The detected work cycles are further classified as: a) unproductive, b) normal productive, and c) abnormal productive. This classification highlights potential management and work flow issues. All work cycles are initially classified as productive and unproductive through region–based classification. This classification categorizes the areas of a jobsite as "active" and "inactive". The former are areas

where construction related-tasks take place e.g. steel fixing, concrete pouring and brick laying. The latter are areas where no construction related-tasks take place. If the work cycles are detected within the "active" areas, then they are classified as productive. If not, then they are classified as unproductive. Given that the overall proposed framework tracks the 3D location of workers over time, this region-based classification is achieved by providing the coordinates of the "inactive" areas. Similar classification has also been proposed by previous researchers to turn the trajectory data into labour input. However, this chapter uses the region-based classification only to provide an overview of the time the workers spend at areas that are linked to management issues and low productivity rates as shown in Chapter 1. Such areas are the rest, the office and the materials' storage areas. Then, the productive work cycles are further classified as normal and abnormal based on their duration. The abnormal are those with the longest duration compared to the total cycles of a worker. This allows project managers to check first the tasks that consumed most of workers' time, if the labour output is not the desirable output. This is achieved by deriving the video footages at the exact time of the day the abnormal cycles occurred.

The method presented in this chapter makes the following assumptions: a) workers have a range of movements (i.e. bending, sideway steps) while performing a task. Hence, their 3D speed values are not zero (≠0) while at a "stop", b) the triangulation and calibration errors, which propagate as noise in the calculation of the computer vision-based 4D trajectories, have the same effect on all data sets regardless of the experimental settings i.e. distance of cameras from workers, size of calibration board, indoors/outdoors monitoring, c) the range of workers' speed values while at "stop" is more representative for classifying the 4D sub-trajectories compared to the speed values while at "move", as the latter varies due to the fact that workers also carry equipment while walking, d) the location of the "inactive" areas does not change often, and e) the longest duration work cycles imply problems such as work flow inefficiencies, congestion, disruptions and safety**.**

5.4. Proposed methodology

This section analyses the method of the proposed solution for monitoring the labour input of construction workers through their trajectory data.

5.4.1. Smoothing of 4D trajectories

This section proposes a method to remove noise from 4D trajectory data (smoothing). The overall proposed framework of this thesis uses a computer vision (tag-less) method (see Chapters 3 and 4) to calculate the 4D trajectories of workers. This way the labour productivity of workers is monitored in a non-obtrusive way. However, such visual methods entail noise. This is due to errors caused by the implementation of 2D tracking (see Chapter 3), triangulation and calibration methods (see Chapter 4). An example of a worker who remains almost totally still while performing an electrical task on the top

of a ladder (see Figure 5-11) is used to display how such types of noise, affect the quality of the trajectory data.

Figure 5-11: Screenshots of a tracked worker who remains still while performing an electrical task.

Given that the worker in the example of Figure 5-11 does not move, the XT, YT, and ZT time series should have been depicted by straight lines of zero degrees inclination. However, the ZT time series displays motion as shown in (a) of Figure 5-12. As a result, the worker in (b) of Figure 5-12 appears as if he has moved 0.45m along the X axis and 1m along the Z axis. Therefore, noise must be removed before classifying the 4D sub-trajectories as either part of "stop" or "move" events.

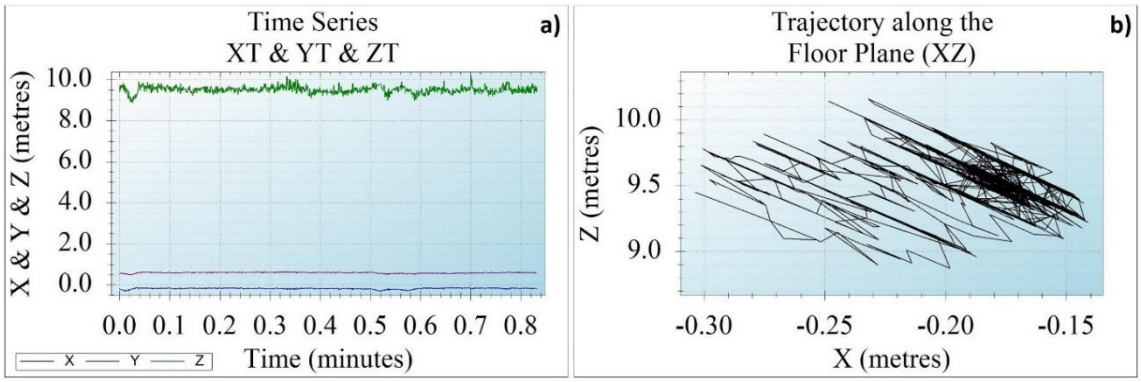

Figure 5-12: Unsmoothed trajectory data of a non-moving worker (see Figure 5-11). (a) Time series along X (blue), Y (purple), and Z (green) axes. (b) Floor plane trajectory.

Existing smoothing methods such as moving averages are robust for removing random noise from time series. They replace every data point of a time series by averaging k previous points. These methods either treat all data points with the same significance or consider that the most recent are more important. The moving average methods are: a) the simple moving average (SMA):

$$\hat{x}_t = \frac{1}{k}\sum_{i=0}^{k-1} x_{t-i},\ for\ \ t = k+1, \ldots, n \qquad (5\text{-}1)$$

b) the weighted moving average (WMA):

$$\hat{x}_t = \frac{1}{\sum_{i=0}^{k-1} w_i} \sum_{i=0}^{k-1} w_i x_{t-i}, \quad for \ t = k+1, \dots, n \tag{5-2}$$

and c) the exponential moving average (EMA):

$$\hat{x}_t = \lambda x_{t-1} + (1-\lambda)\hat{x}_{t-1} = \sum_{i=0}^{k-1} \lambda(1-\lambda)^i x_{t-k} + (1-\lambda)^t \hat{x}_0, \quad for \ 0 < \lambda \leq 1 \tag{5-3}$$

where weights w_i and constant λ of WMA and EMA respectively, are chosen either based on the experience of the user, the nature of the data or through trial and error processes (Gor & Man, 2009). All moving average methods are experimentally evaluated in the following section in order to choose the best. The problem of such smoothing methods is that they also discard non-noise peak values if a large smoothing step k is chosen (Guiñuinón et al., 2007). This section presents a method that searches for the optimum smoothing step k to alleviate this issue. It exploits the fact that all three time series (XT, YT and ZT) of a worker who does not move are straight lines of almost zero degrees inclination. Firstly, all three time series are smoothed by a randomly selected initialization step k. Secondly, simple linear regression fits a line $f_t = ax_t + b$ to every smoothed time series that minimizes the sum of residuals $e_t = x_t - f_t$ (see Figure 5-13).

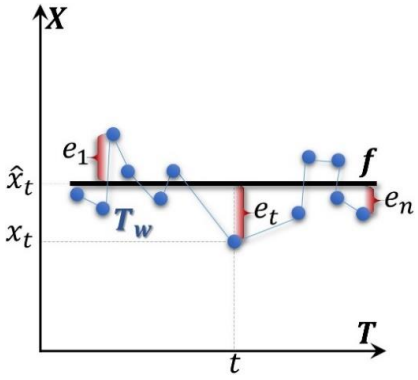

Figure 5-13: Fitting a line to a time series.

Smoothing and fitting are repeatedly performed for all three time series. During this searching, the smoothing step k is increased successively. Two metrics are commonly used for measuring this goodness of fit. These metrics are: a) the adjusted coefficient of determination or adjusted R-squared r^2_{adj} that is expressed as:

$$r^2_{adj} = 1 - \frac{(1-r^2)(n-1)}{(n-j-1)} \tag{5-4}$$

where n is the sample size, and j is the number of independent variables in the regression equation, and b) the coefficient of determination r^2:

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{5-5}$$

where SS_{res} is the total sum of squares:

$$SS_{tot} = \sum_t (x_t - \bar{x})^2 \quad (5\text{-}6)$$

and SS_{tot} is the total sum of residuals:

$$SS_{res} = \sum_t (x_t - f_t)^2 \quad (5\text{-}7)$$

The adjusted R-squared r^2_{adj} is selected as it penalizes the addition of variables that do not contribute to the goodness of fit. Therefore, the proposed method searches repeatedly for the optimum smoothing step k, if r^2_{adj} of all three time series (XT, YT, ZT) increases, and if the fitted lines have inclinations larger than zero degrees. It terminates searching, if either r^2_{adj} of any of the three time series (XT, YT, ZT) decreases, or if straight lines of almost zero degrees inclination are fitted to all time series. Finally, the value of step k, when searching terminates is used to smooth the 4D trajectories. Algorithm 5-1 illustrates the searching loop for the calculation of the optimum smoothing step k.

Algorithm 5-1: Loop of proposed smoothing method.

```
for each k
  for each XT, YT, ZT
  Apply simple linear regression
  Calculate r²_adjXT, r²_adjYT, r²_adjZT, a_XT, a_YT, a_YT
  If r²_adjXT_k ≥ r²_adjXT_k−1 & r²_adjYT_k ≥ r²_adjYT_k−1 & r²_adjZT_k ≥ r²_adjZT_k−1 & a_XT, a_YT, a_YT > 0⁰
    k = k+1;
    continue;
  else
    Smoothing step = k
    break;
  end if
end for
```

5.4.2. Partitioning of 4D trajectories

The partitioning of trajectories allows the detection of work cycles that are hidden in small segments of single trajectories. The proposed method partitions the 4D trajectories in a way that describes the best proposed semantic "stop" and "move" events. Therefore it is essential at this point to clearly define these two semantic events. Workers while at "stop" perform tasks. This entails either torso movement (i.e. forward bending, backward bending, sideways bending, and rotation) or sideway steps (see Figure 5-14). On the other hand, workers while at "move" either walk along the floor plane or along the vertical plane (e.g. climbing a ladder).

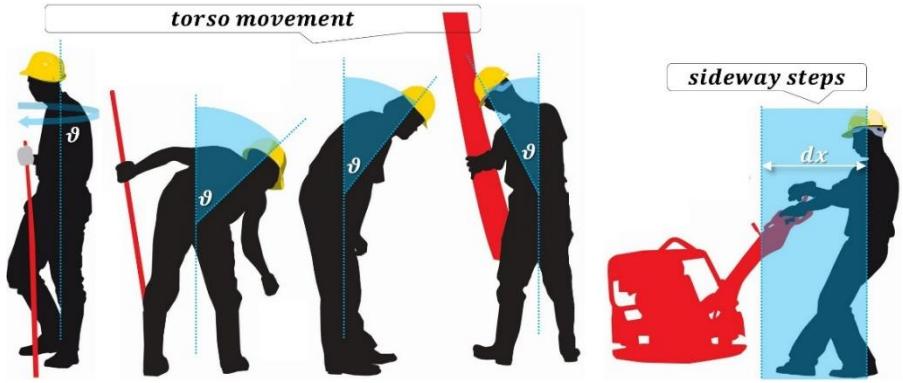

Figure 5-14: Range of workers' movements while at "stop" event.

The partitioning is time-based. It is known that the stride length of a person is 0.815m on average (Barreira et al., 2010). This is almost half of the 1.4m that a walking person covers on average in 1sec (Boonstra et al., 1993). This entails that in 1sec a worker who makes sideway steps while at "stop" has smaller speed along the floor plane than a worker who walks while at "move". Therefore, every single 4D trajectory T_w is partitioned into N 4D sub-trajectories $S_w = \{s_i\}_{i=1\dots N}$ of equal duration d_s. A 4D sub-trajectory s_i is then described by a representative 3D speed value that depicts the motion of a worker along the floor and the vertical plane and is expressed as follows:

$$\overline{\|\vec{v}_i\|}_y = \frac{\left|y_{t_{a+d_s}} - y_{t_a}\right|}{d_s} \tag{5-8}$$

$$\overline{\|\vec{v}_i\|}_{xz} = \frac{\sqrt{\left(x_{t_{a+d_s}} - x_{t_a}\right)^2 + \left(z_{t_{a+d_s}} - z_{t_a}\right)^2}}{d_s} \tag{5-9}$$

5.4.3. Classification of 4D sub-trajectories

This section classifies the 4D sub-trajectories $\{s_i\}_{i=1\dots N}$ of workers as either part of "stop" or "move" events. If the speed value of a 4D sub-trajectory s_i falls within the range of "stop" speed values, then it is classified as part of a "stop" event, if not, then it is classified as part of a "move" event. This range is experimentally defined in the following section. Both, the speed along the floor plane $\|\vec{v}_i\|_{xz}$ and the speed along the vertical plane $\|\vec{v}_i\|_y$ must return a positive "stop" to ensure a correct classification. If either $\|\vec{v}_i\|_{xz}$ or $\|\vec{v}_i\|_y$ returns a positive "move", then the 4D sub-trajectory is classified as part of a "move" event. The classification along the floor plane is expressed as follows:

$$\|\vec{v}_i\|_{xz} = \begin{cases} \text{"stop"}, if\text{: } \|\vec{v}_i\|_{xz} \leq \left[0, \|\vec{v}_i\|_{xz_{stop}}\right] \\ \text{"move"}, if\text{: } \|\vec{v}_i\|_{xz} > \left[0, \|\vec{v}_i\|_{xz_{stop}}\right] \end{cases} \tag{5-10}$$

and along the vertical plane as:

$$\|\vec{v}_i\|_y = \begin{cases} \text{“stop”}, if: \|\vec{v}_i\|_y \leq \left[0, \|\vec{v}_i\|_{y_{stop}}\right] \\ \text{“move”}, if: \|\vec{v}_i\|_y > \left[0, \|\vec{v}_i\|_{y_{stop}}\right] \end{cases} \quad (5\text{-}11)$$

where $\|\vec{v}_i\|_{xz_{stop}}, \|\vec{v}_i\|_{y_{stop}}$ is the upper threshold of a worker´s speed values along the floor plane and the vertical plane respectively while at “stop”.

5.4.4. Clustering of 4D sub-trajectories into work cycles

This section presents how all classified 4D sub-trajectories are clustered into work cycles. Initially, all 4D sub-trajectories that are sequentially classified as either part of “stop” or “move” events are grouped into semantic events of larger duration. Then, the proposed clustering method relies on the main assumption of this thesis as shown in Chapter 2 to detect work cycles $\{c_i\}$. This assumption dictates that if one “move” event is followed by one “stop” and one “move” event sequentially, then all 4D sub-trajectories that belong to these three sequential semantic events depict a work cycle c_i. This is expressed as follows:

$$\text{Work Cycle } c_i = \text{"move"}_{t_{end}}^{t_{start}} + \text{"stop"}_{t_{end}}^{t_{start}} + \text{"move"}_{t_{end}}^{t_{start}} \quad (5\text{-}12)$$

The starting t_{start} and ending t_{end} time of a semantic event is equal to the starting and ending time of its first and last 4D sub-trajectory respectively. Figure 5-15 illustrates how the semantic events are clustered into work cycles. The minimum duration of a detectable semantic event is equal to 1sec given that the proposed method partitions the 4D trajectories every 1sec.

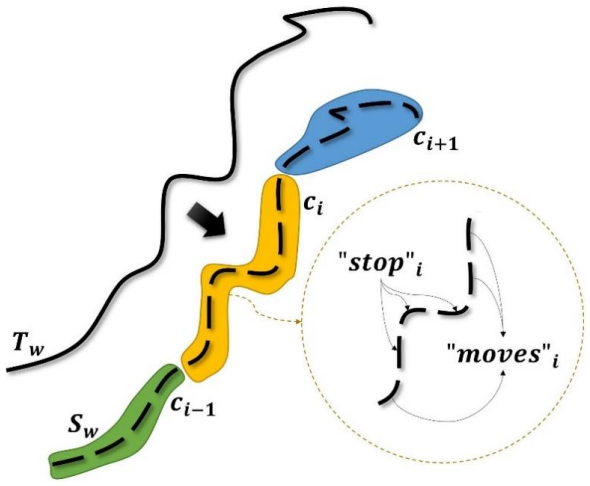

Figure 5-15: Clustering of 4D sub-trajectories into three work cycles c_i (blue, yellow, green).

5.5. Experiments and results

This section presents the performance of the method proposed in this chapter in terms of: a) detecting work cycles, and b) determining the labour input of construction workers.

The performance of the proposed method is evaluated with two data sets. The first, (data set steel fixing) captures one worker while performing a steel fixing task. The second, (data set electrical) consists of two workers who perform an electrical task. Data set steel fixing was recorded at a pre-manufacturing facility (Bison), whilst data set electrical was collected at a jobsite in Cambridge (James Dyson building). The total durations of data sets steel fixing and electrical are 35minutes and 51.5minutes respectively. Figure 5-16 illustrates the camera setup for each data set.

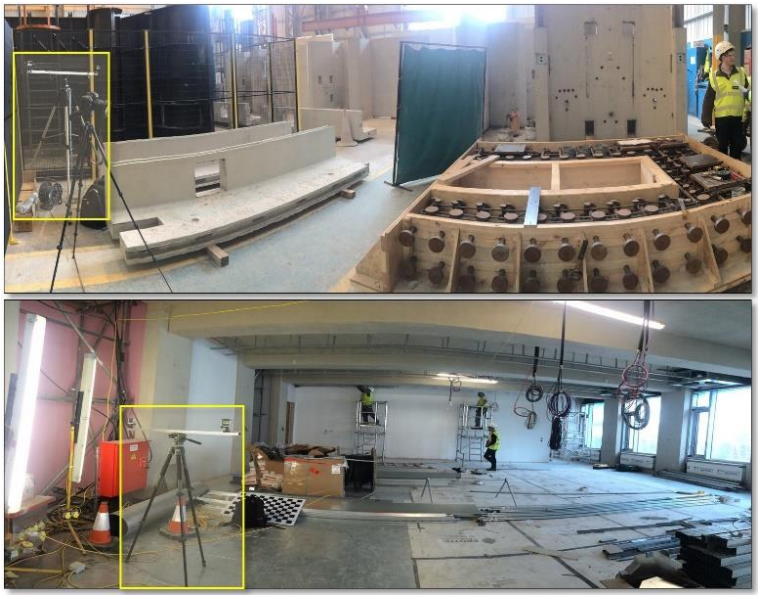

Figure 5-16: Tested data sets (from top to bottom: data set steel fixing, data set electrical).

Figure 5-17 illustrates with yellow dotted cubes the tracked areas. Because of the limited number of cameras used, the proposed method could not track the workers at "inactive" areas (i.e. rest, office, and materials' storage areas). Due to this, the time the workers spend outside the tracked areas is automatically returned as unproductive. This section classifies as abnormal those work cycles that last at least 50% longer. This threshold is not experimentally defined given that the scope of the overall proposed framework is to approach the monitoring of labour productivity in a generalized way for all types of tasks.

Figure 5-17: Tracked areas (from left to right: worker "1" of data set steel fixing, workers "2" and "3" of data set electrical).

The ground truth consists of 85 work cycles in total. This sample corresponds to a confidence level of 95% (Z_{crit} = 1.960) and a limit of error D equal to ±6% based on the following equation (Eng, 2003):

$$n = \frac{4{Z_{crit}}^2 p(1-p)}{D^2} \quad (5\text{-}13)$$

In the above equation, work cycles are the sample type. Therefore, the p variable which is the proportion of sample with a specific characteristic, is equal to the proportion of the unproductive work cycles. Productive work cycles are those when workers actually perform sub-tasks while at "stop", whilst unproductive cycles depict workers who are not involved in any construction-related task. In total, steel and electrical data sets contain 78 productive work cycles and 7 unproductive. Hence, the p variable is equal to 8%.

Precision, recall, and accuracy metrics are used for the evaluation of this chapter's proposed method. Precision is the fraction of the total number of correctly detected work cycles (TP, True Positive) over the total number of incorrectly and correctly detected work cycles (TP + FP, True Positive + False Positive). Recall depicts the detection completion level and is equal to the total number of correctly detected work cycles (TP) divided by the total number of correctly detected and incorrectly not detected work cycles (TP + FN, True Positive + False Negative). Lastly, accuracy is defined by the number of correctly detected work cycles (TP) and the number of work cycles which were correctly not detected (TN, True Negative), over the total sum of work cycles. The work cycles are manually measured for the evaluation of the proposed method.

5.5.1. Definition of parameters

This section defines experimentally the parameters of this chapter's proposed methodology.

5.5.1.1 Definition of smoothing parameters

This section determines experimentally the smoothing step k in order to remove noise from the 4D trajectories of workers. This section uses the data sample of the still (non-moving) worker of previous Figure 5-11 to achieve this. The method proposed in 5.4.1 returns a smoothing step k equal to 80 points with EMA, 459 with for SMA and 570 points with WMA (see Table 5-1).

Table 5-1: Smoothing step k of the method proposed in section 5.41 with SMA, EMA, and WMA.

Method	Smoothing step (k)								
SMA	**459**								
λ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
EMA	71	77	77	**80**	72	73	70	57	14
WMA	**570**								

Both the speed values along the floor $\{\|\vec{v}_i\|_{xz}\}_{i=1\dots N}$ and the vertical $\{\|\vec{v}_i\|_y\}_{i=1\dots N}$ plane must be close to zero given that smoothing was performed on trajectories of a non-moving worker. Figure 5-18 illustrates the 3D speed values of this worker for each of the smoothing steps of Table 5-1.

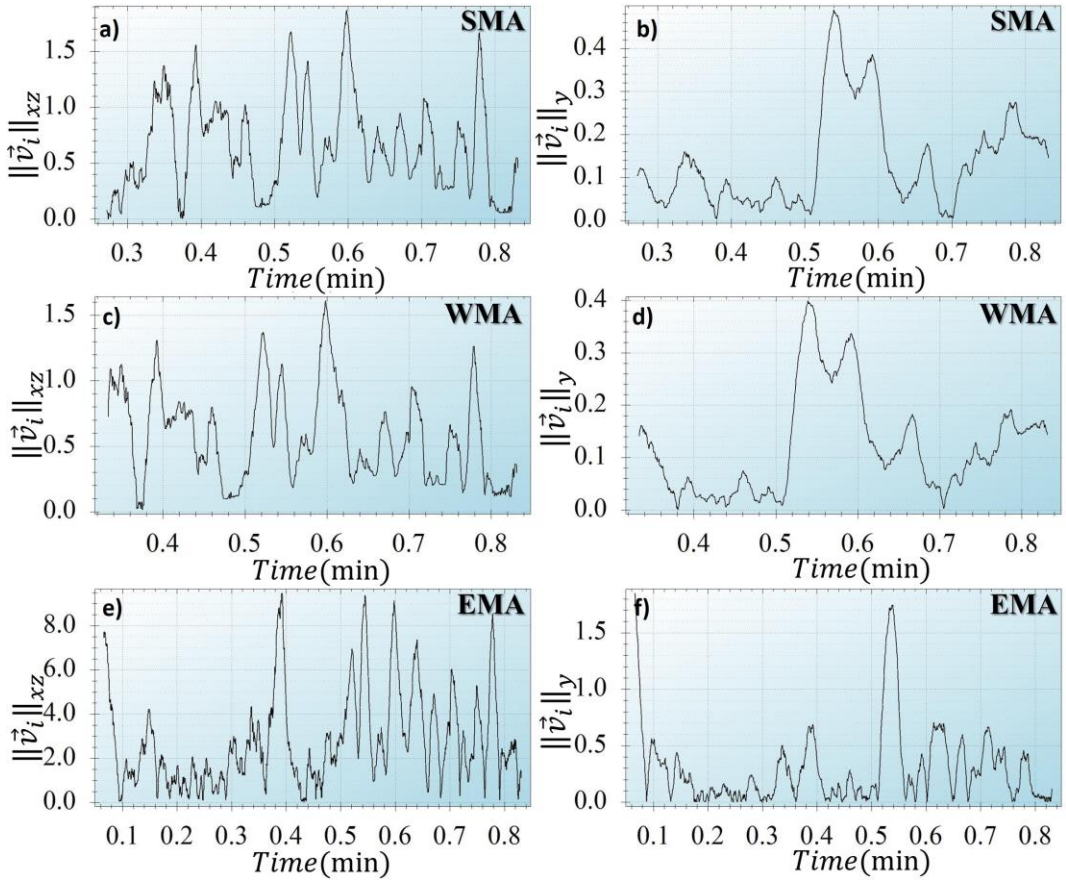

Figure 5-18: 3D speed values of the smoothed trajectory data of a non-moving worker (see Figure 5-11). Speed of worker along the floor plane $\{\|\vec{v}_i\|_{xz}\}_{i=1\dots N}$ with SMA (a), WMA (c), and EMA (e). Speed of worker along the vertical plane $\{\|\vec{v}_i\|_y\}_{i=1\dots N}$ with SMA (b), WMA (d), and EMA (f).

In Figure 5-18 it appears that EMA performs the worst compared to SMA and WMA, whilst the latter performs the best as it returns the lowest 3D speed values. The speed error along the vertical plane with WMA is almost equal to zero. However, the speed error along the floor plane reaches the 1.5meters/min. This is because the smoothing method of section 5.4.1 did not manage to fit straight lines of zero inclination to all time series due to the large noise of the ZT time series as seen in previous Figure 5-12. Figure 5-19 shows that all three smoothed time series (XT: blue, YT: purple and ZT: green) are almost perfectly linear. The maximum distance error is equal to 4cm along the X axis and equal to 10cm along the z axis. The smoothing step k with WMA ($k = 570$) is finally selected as it returns the smallest speed error.

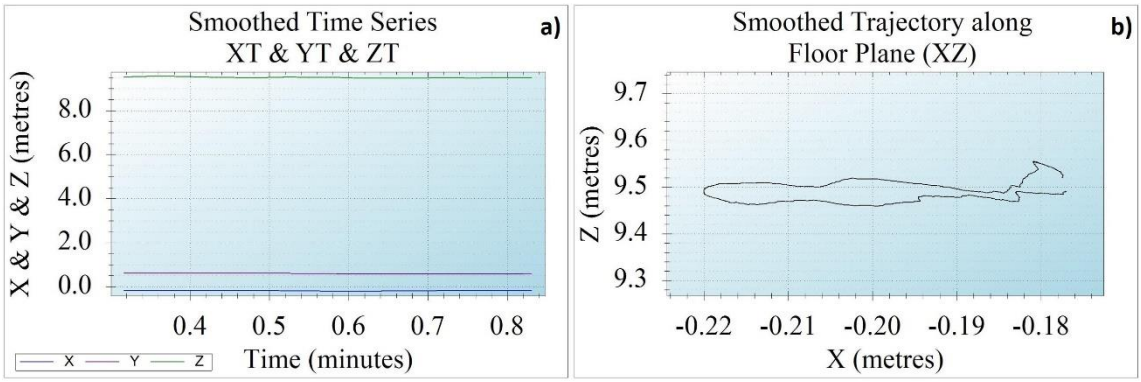

Figure 5-19: Smoothed trajectory data of the almost still worker of previous Figure 5-11. (a) Smoothed time series of X (blue), Y (purple), and Z (green) axis. (b) Smoothed floor plane trajectory.

5.5.1.2 Definition of classification parameters

Worker "1" from data set electrical was randomly selected in order to experimentally determine the range of 3D speed values of any worker while at "stop". This worker performs a steel fixing task for 17minutes. During this time he is not visible from the cameras for 4.15minutes. The performance of the proposed classification method is manually measured for different ranges of 3D speed values through precision, recall, and accuracy. The 3D speed values are normalised by dividing: a) the speed values along the floor plane with the known average human walking speed (Boonstra et al., 1993), and b) the speed values along the floor (XZ) plane with the known average climbing speed (Chang et al., 2004).

Figure 5-20 shows that the best performance was achieved for a range of $\left[0, \|\vec{v}_i\|_{xz_{stop}}\right] = [0, 0.4]$ along the floor plane, and for a range of $\left[0, \|\vec{v}_i\|_{y_{stop}}\right] = [0, 0.6]$ along the vertical plane. For these ranges of 3D speed values, the proposed method returns an accuracy, precision, and recall of 100% in terms of detecting work cycles. Therefore, they are selected in order to evaluate the performance of the proposed method.

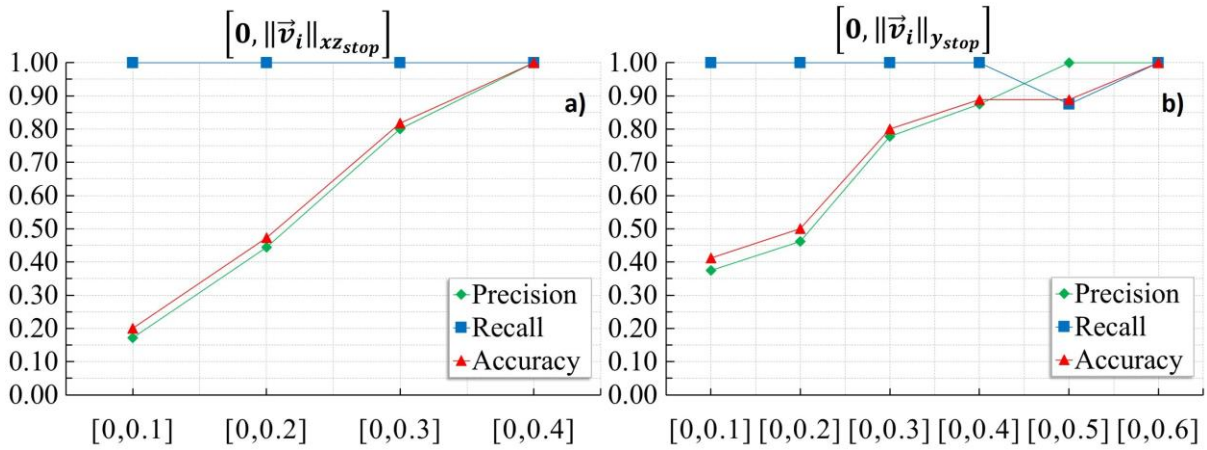

Figure 5-20: Precision, recall, and accuracy graphs for the determination of the range of speed values of any worker while at "stop" along the (a) floor $\left[0, \|\vec{v}_i\|_{xz_{stop}}\right]$, and (b) vertical $\left[0, \|\vec{v}_i\|_{xz_{stop}}\right]$ plane.

Figure 5-21 illustrates the normalized speed values $\{\|\vec{v}_i\|_{xz}\}_{i=1\ldots N}$ along the floor plane of worker "1". It displays with green colour the parts of the diagram that correspond to "moves" and with red colour the time the worker was not visible from the cameras.

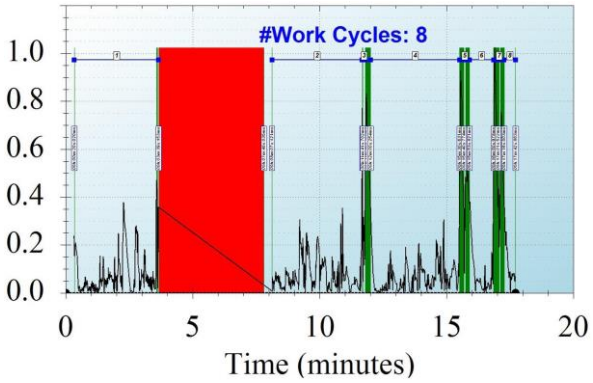

Figure 5-21: Normalized speed values of worker "1" along the floor plane $\{\|\vec{v}_i\|_{xz}\}_{i=1\ldots N}$.

5.5.2. Evaluation of work cycles' detection

This section evaluates the performance of the method presented in this chapter in terms of translating the trajectory data into labour input. It colours red the unproductive, yellow the abnormal productive and green the normal productive work cycles.

Data set steel fixing

This data set consists of two recordings (part A, B) with a duration of approximately 17minutes each. Both recordings were collected the same day. Table 5-2 shows the ground truth of the manually collected work cycles. In total, worker "1" performed 29 work cycles that depict the following sub-

tasks: a) fixing steel re-bars, b) picking re-bars or equipment, and c) reading drawings. Only one is unproductive, whilst none of the productive work cycles corresponds to idle time.

Table 5-2: Manually collected ground truth of semantic "stops" of worker "1" (data set steel).

	Part A (GT)		Part B (GT)		
#	Start - End	#	Start - End	#	Start - End
1.	00:00:033-0:05:105(TN)	11.	00:00:033-00:13:680(TN)	21.	14:15:388-15:02:268(FN)
2.	00:12:645-03:37:884(TP)	12.	00:19:919-00:21:287(TN)	22.	15:03:670-16:15:175(TP)
3.	03:39:153-07:48:035(TP)	13.	00:25:258-00:30:697(TP)	23.	16:18:645-16:28:988(TP)
4.	07:50:036-11:41:067(TP)	14.	00:33:867-00:34:768(FN)	24.	16:32:359-16:42:102(TP)
5.	11:42:268-11:44:070(TP)	15.	00:37:771-01:24:150(TP)	25.	16:47:173-17:02:422(TP)
6.	11:52:545-15:31:798(TP)	16.	01:28:855-02:36:322(TP)	26.	17:05:658-17:08:628(FN)
7.	15:35:101-15:40:273(TP)	17.	02:40:293-02:47:667(FN)	27.	17:09:896-17:18:171(TP)
8.	15:45:645-16:49:609(TP)	18.	02:49:202-02:51:071(FN)	28.	17:23:943-17:29:416(FN)
9.	15:57:317-17:02:989(TP)	19.	02:52:205-05:53:687(FN)	29.	17:30:517-17:41:227(TP)
10	17:11:131-17:42:962(TP)	20.	05:55:021-14:14:087(TP)		

Start/End → min:sec:msec

The proposed method detects 9 TP, 1 TN, 0 FP and 0 FN work cycles in part A, and 10 TP, 2 TN, 7 FN and 1FP work cycles in part B. The pie charts in these figures show the durations of the detected work cycles. The 3 TN results (#1, #11, and #12) result from the way trajectories are smoothed. The smoothing step k (see section 5.4.1) is equal to 19sec if divided by camera frame rate i.e. $\frac{k}{fps} = \frac{570}{30}$. Hence, all smoothed time series are 19sec shorter in length at the beginning compared to the unsmoothed. Hence, work cycles that fall within the initial 19sec cannot be detected. Table 5-2 confirms that all TN results occur at the beginning of each recording. The missed #17 and the 1FP work cycles of part B are due to instabilities of the implemented computer vision-based 2D tracking method. All the rest FN work cycles are of short duration (< 4sec). This shows that that the proposed method does not perform well in terms of detecting work cycles of such short duration.

In Part A, the work cycle #3 shows that worker "1" was unproductive i.e. definitely not performing the steel fixing task, for 4.15minutes. Work cycles #5, and #7 to #10 are all classified as normal productive whilst the work cycles with the largest duration #2, #4, and #6, are all classified as abnormal productive (see (a) in Figure 5-22). In part B, nine out of ten work cycles of this recording, are returned as normal (#13, #15, #16, #22 to #25, #27 and #29) and only one is classified as abnormal (#20) (see (b) in Figure 5-22). The interesting observation about this recording, is that the abnormal cycle has a duration of 11.56minutes which is by far the largest compared to the rest cycles of both parts A and B. This raises an ambiguity about the performance of worker "1" during this time. It can be easily observed, if we check the video footages at the exact time the abnormal cycle #20 occurred, that worker "1" could not fit a reinforcing steel bar in the formwork due to complexity of drawings. This is a common issue that affects labour productivity negatively and is identified as one of the on-site factors in Chapter 1. If we sum all the TP normal and abnormal work cycles, then the labour input of the worker

"1" is equal to 28.29minutes for both parts (A, B). The manually calculated labour input is equal to 30.62minutes. Therefore, the proposed method measured the total labour input of the steel worker with an accuracy of 92%.

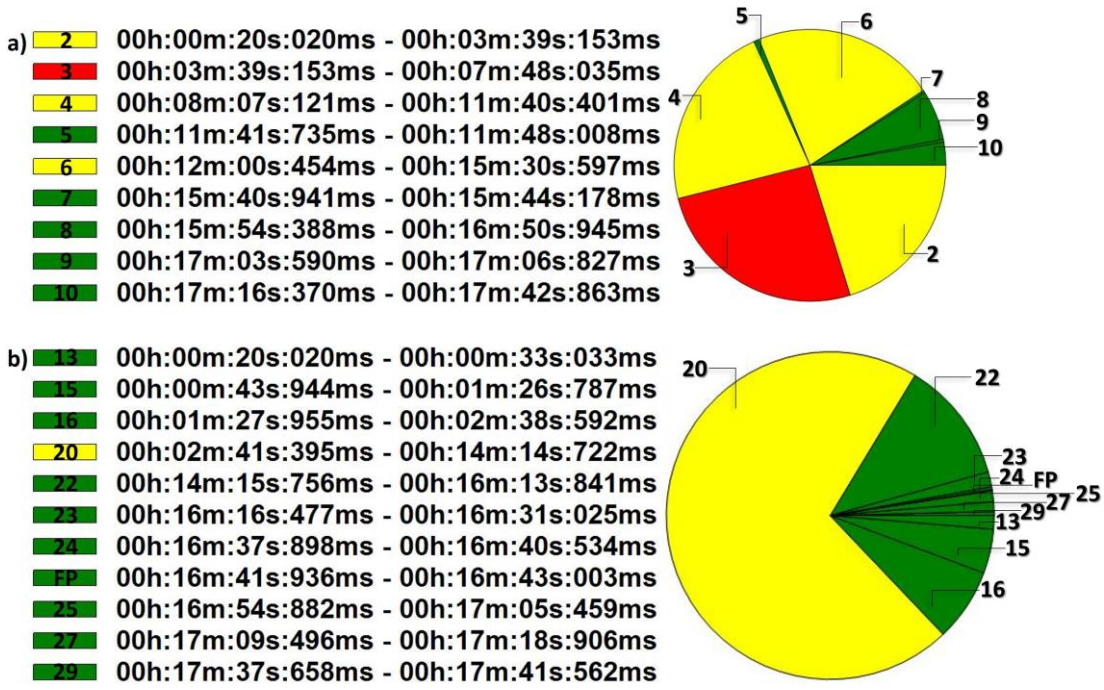

Figure 5-22: Detected work cycles of worker "1" from data set steel fixing part A (a), and B (b).

Data set electrical

This data consists of three recordings (part A, B, C) that were all collected the same day. The duration of part A is 12minutes, of part B is 7minutes and of part C is 11minutes. Table 5-3 and Table 5-4 illustrate the manually collected work cycles (ground truth/GT) of worker "2" and worker "3" respectively. Overall, worker "2" performed 24 work cycles and worker "3" performed 32. From the total work cycles of both workers, only 6 are unproductive. The rest 50 are productive work cycles that depict sub-tasks such as picking of material or equipment and placing of electrical cables. None of the productive work cycles of both workers corresponds to idle time.

The proposed method detects 17 TP, 3 TN, 1 FP and 4 FN for worker "2", and 22 TP, 3 TN, 1 FP and 7 FN for worker "3" for all three A, B, and C parts. Similarly to data set steel, all TN results of data set electrical of both workers are due to the shortening of the length of trajectories by 19sec. These TN are detected either at the beginning of each recording (see #18 and #19 for worker "2") or when workers re-enter the camera field of view (see #12 for worker "2", and #3, #9, #10 for worker "3"). The proposed method fails again to detect work cycles of relatively short duration (< 4sec). Only, the #21 work cycle of worker "3" that lasts 16sec is missed due to performance issues of the implemented computer vision-based 2D tracking method. Such an example is presented in Figure 5-23. The same holds for all FP work cycles of this data set.

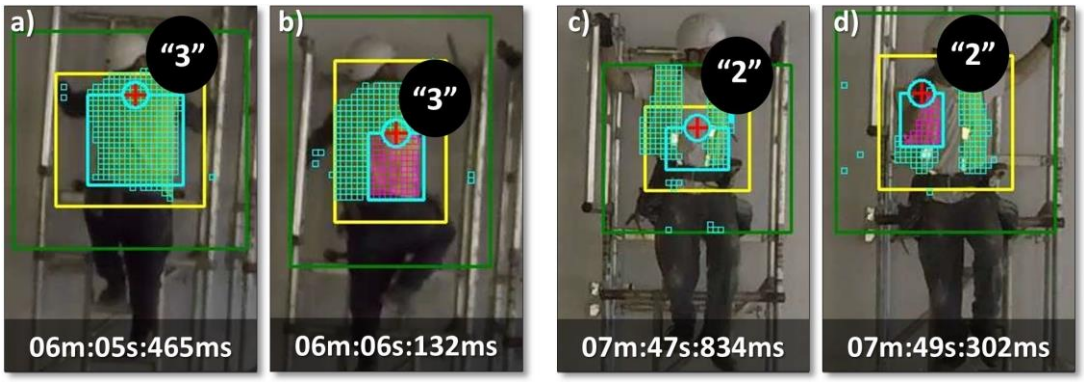

Figure 5-23: Significant fluctuation of implemented computer vision-based 2D tracking method.

Table 5-3: Manually collected ground truth of semantic "stops" of worker "2" (data set electrical).

#	Part A (GT) Start - End	#	Part B (GT) Start - End	#	Part C (GT) Start - End
1.	00:00:033-03:48:261(TP)	8.	00:00:033-02:26:513(TP)	18.	00:00:033-00:01:868(TN)
2.	03:50:263-03:56:569(TP)	9.	02:29:549-02:32:819(FN)	19.	00:03:003-00:08:341(TN)
3.	04:02:375-07:46:265(TP)	10.	02:46:566-02:48:134(TP)	20.	00:14:948-04:20:593(TP)
4.	07:55:641-08:11:190(TP)	11.	02:50:370 03:18:064(TP)	21.	04:23:963-04:34:407(TP)
5.	08:15:561-08:38:551(TP)	12.	03:20:833-03:24:804(TN)	22.	04:37:610-04:39:412(FN)
6.	08:40:219-08:41:954(TP)	13.	03:33:346-03:44:857(FN)	23.	04:41:381-04:43:983(FN)
7.	08:50:463-11:49:608(TP)	14.	03:47:894-05:49:916(TP)	24.	04:47:253-11:07:166(TP)
		15.	05:50:016-05:51:584(TP)		
		16.	05:51:584-06:43:636(TP)		
		17.	06:58:518-07:00:653(TP)		

Start/End → min:sec:msec

Table 5-4: Manually collected ground truth of semantic "stops" of worker "3" (data set electrical).

#	Part A (GT) Start - End	#	Part B (GT) Start - End	#	Part C (GT) Start - End
1.	00:00:033-05:16:049(TP)	18.	00:00:033-07:00:420(TP)	19.	00:00:033-05:52:185(TP)
2.	05:23:823 05:31:331(TP)			20.	05:57:690-06:03:329(FN)
3.	05:33:566-05:39:839(TN)			21.	06:08:201-06:24:250(FN)
4.	05:41:974-05:43:876(FN)			22.	06:27:353-06:33:025(TP)
5.	05:53:920-06:55:748(FN)			23.	06:35:128-06:36:896(TP)
6.	07:04:824-07:07:593(TP)			24.	06:50:176-06:53:246(FN)
7.	07:11:063-07:31:017(TP)			25.	07:02:422-07:09:128(TP)
8.	07:33:920 08:17:263(TP)			26.	07:11:164-07:25:845(TP)
9.	08:18:698-08:21:667(TN)			27.	07:28:414-07:30:283(TP)
10.	08:25:872-08:33:145(TN)			28.	07:36:856-07:52:805(TP)
11.	08:36:749-08:39:652(FN)			29.	07:57:877-08:00:580(TP)
12.	08:41:988-08:42:655(FN)			30.	08:11:757-09:24:063(TP)
13.	08:45:992-09:04:911(TP)			31.	09:26:566 09:28:634(TP)
14.	09:08:180-09:13:286(TP)			32.	09:39:178-11:07:166(TP)
15.	09:36:075-09:56:562(TP)				
16.	10:01:000-10:13:746(TP)				
17.	10:19:352-11:49:608(TP)				

Start/End → min:sec:msec

Figure 5-24 shows that worker "2" performed 7 abnormal work cycles (#1, #3, #7, #8, #14, #20 and #24), 7 normal (#2, 5, #6, #10, #16, #17, and #21) and 3 unproductive (#4, #11 and #15). All abnormal work cycles have durations close to 3minutes. On the other hand, all 4 normal work cycles have very short durations of less than 1minute. Therefore, not any particular "suspicious" performance, that could imply work flow issues or negatively influencing factors, is obvious. The proposed method returns a total labour input equal to 26.21minute for worker "2". The manually measured labour input is 27.55minutes. Therefore, the accuracy of the proposed method in terms of estimating the total labour input is equal to 95%.

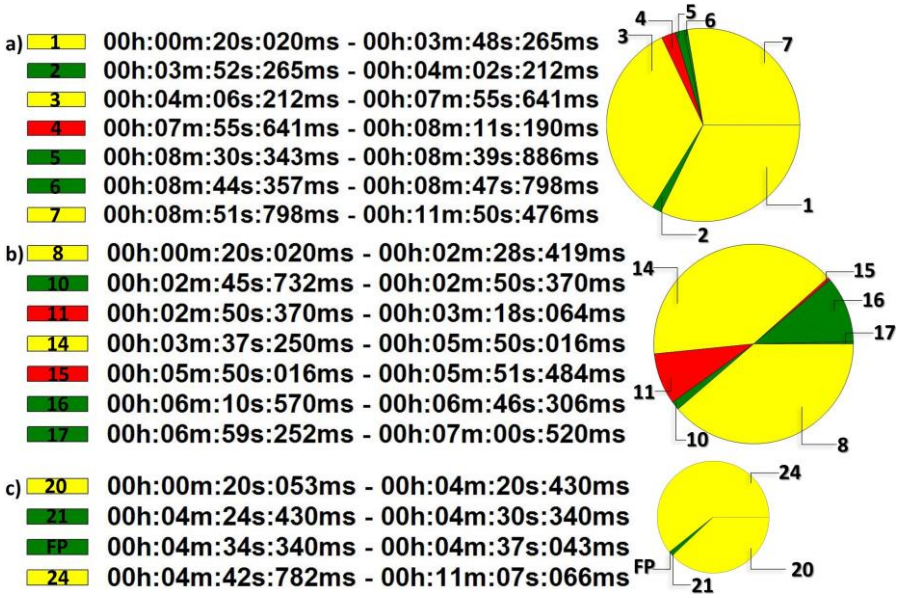

Figure 5-24: Detected work cycles of worker "2" from data set electrical part A (a), B (b), and C (c).

Figure 5-25 shows that worker "3" performed 3 unproductive work cycles (#2, #8, and #31) of short duration (< 1minute), 3 abnormal (#1, #18 and #19) that last close to 6minutes and 16 normal (#6, #7 #13 to #17, #22, #23, #25 to #30, and #32) with a duration less than 1minute. If we sum up all normal and abnormal productive cycles, then the total labour input of worker "3" is equal to 25.55minutes. This compared to the 26.35minutes of the ground truth returns an accuracy of 97% in terms of estimating the labour input. Overall, the total labour input of both workers "2" and "3" is similar. The only difference, is that worker "3" has longer in duration abnormal cycles (5.88mintes on average) compared to worker "2" (3.4minutes on average). If we go back to the video data, we observe that this is because worker "3" spends more time to pick up equipment and materials compared to worker "2". This is a work flow issue that could be resolved if materials were brought either closer to both workers or if one worker was assigned to assist workers "2" and "3" until they finish their electrical task on top of the ladders. However, it is on the project manager's judgment to apply changes on the workflow of the electrical task if he/she is not satisfied with the labour output of workers. The advantage of the method

proposed in this last chapter is that it provides all the information that a project manager needs in order to monitor the labour productivity of all three workers proactively.

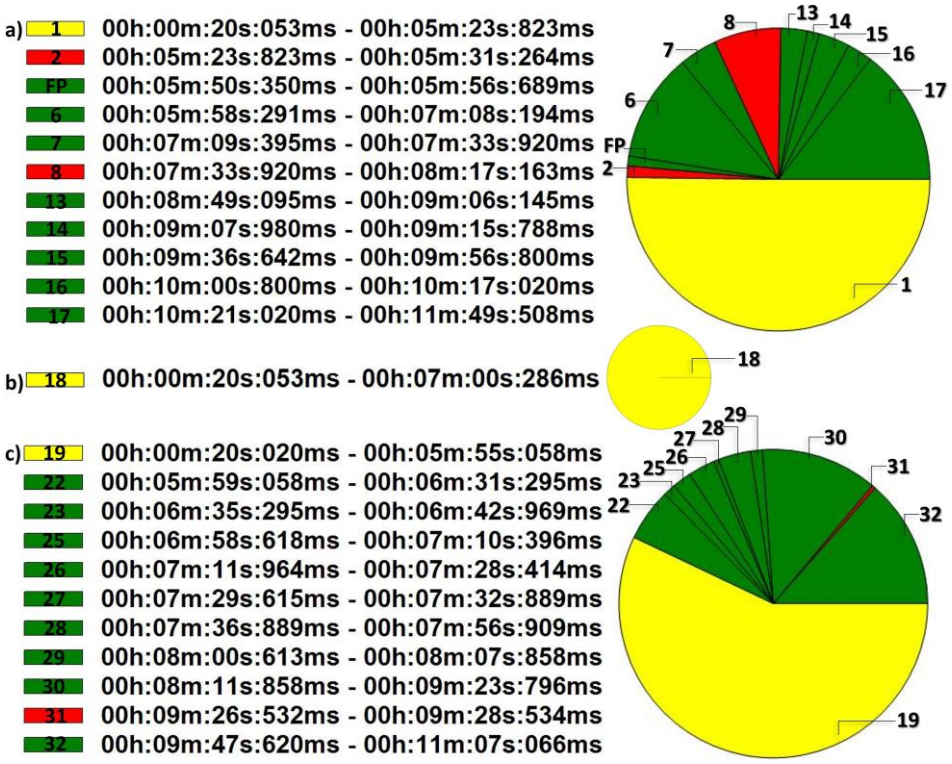

Figure 5-25: Detected work cycles of worker "3" from data set electrical part A (a), B (b), and C (c).

Table 5-5 summarizes the performance of the method proposed in this chapter in terms of detecting work cycles for the purpose of converting the trajectory data into labour input. This table presents in detail the exact number of the total: a) correctly detected (TP), b) incorrectly detected (FP), c) incorrectly not detected (FN), and d) correctly not detected (TN) work cycles for both data set steel and data set electrical. The method features 95% precision, 76% recall, and 76% accuracy. The small number of FP results shows that the proposed method is not significantly affected by noise, whilst the high rate of FN indicates that the proposed method is not efficient in detecting work cycles of short duration (<4sec).

Table 5-5: Confusion matrix of proposed method for detecting work cycles.

		Predicted	
		Yes	**No**
Actual	**Yes**	TP	FN
	No	FP	TN

→

58	18
3	9

Table 5-6 shows that the proposed method returns an accuracy of 95% on average in terms of calculating the labour input i.e. the time workers spend on performing construction-related tasks. This accuracy

outperforms the 59% of Jun Yang et al. (2016) and the 86% of Gong & Caldas (2011) and addresses the second aim of this thesis.

Table 5-6: Quantitative summary of the labour input of the monitored steel fixing and electrical task.

	GT labour Input	**Detected labour input**	**Accuracy**
Worker "1" (data set steel)	30.62	28.29	92%
Worker "2" (data set electrical)	27.55	26.21	95%
Worker "3" (data set electrical)	26.35	25.55	97%
		Average	95%

An advantage of the proposed method is that it does not need any prior knowledge about the type or the number of tasks. Therefore it is applicable to multiple workers at the same time. Additionally, workflow inefficiencies and potential management issues can be identified through the abnormal productive cycles or through the trajectories of workers. Up to date, project managers have to manually draw the spaghetti diagrams that depict workers' motion on the floor plane (Nyström & Per, 2009) to achieve this. The main limitation of the proposed method is that it also detects as productive work cycles that depict idle time i.e. workers simply standing without performing any task. However, as shown in chapter 2, idle time is not one of the main reasons of low labour productivity in the construction sector. Another limitation relates to the lack of many cameras. Due to this restriction, the benefit of monitoring workers across the entire range of jobsites is not efficiently evaluated.

In summary, with the method presented in this chapter, the proposed framework monitored the labour productivity of three workers: a) accurately, through the detection of work cycles, b) regardless of any knowledge of the type or of the number of tasks these three workers performed, c) time and cost efficiently, as none of the intermediate methods of the proposed framework required any human intervention or tags besides the cameras that in the real case scenario would be the cameras of the surveillance system, and d) unobtrusively, due to the computer vision-based methods employed in the proposed framework. For all these reasons, it is proved that the framework presented in this thesis can monitor the labour productivity of multiple construction workers at the same time i.e. proactively. This addresses the third aim of this thesis.

5.6. Chapter overview

The current state of research as presented in Chapter 2 has not yet proposed a method that performs a non-obtrusive, accurate, cost efficient and generalized monitoring of labour productivity for construction workers. Existing studies focusing on trajectory analysis for security purposes, fail to detect repetitive patterns in trajectories of workers due to limitations of clustering methods. Such clustering methods struggle: a) to cluster large data sets accurately, due to the computational and time

complexity that grows in parallel to the size of the data, and b) to achieve automation as clustering depends on users' estimations. This chapter presents a clustering method that addresses these issues in order to detect repetitive patterns in the trajectories of construction workers that depict work cycles. The total duration of these work cycles is equal to the labour input of workers.

The novelty of the proposed method lies in clustering. Firstly, the 4D trajectories of workers are smoothed in order to remove noise. Then, they are segmented into 4D sub-trajectories and classified as either "move" or "stop" semantic events. The former event depicts the motion of workers along the floor plane, whilst the latter depicts the motion of workers along the vertical plane. The classified 4D sub-trajectories are finally grouped into clusters based on the main assumption of this thesis that: *every work cycle is described by two semantic "move" events and one semantic "stop" event*.

The main limitations of the method presented in this chapter are the following: a) work cycles that depict workers who while at "stop" do not perform any task (idle time) are mistakenly detected as productive, and b) the productivity of workers who perform tasks mainly characterized by motion such as transferring materials, supervising work progress etc. cannot be monitored. This is because the "move" events depict the actual labour input instead of the "stop" events in such cases. This second limitation indicates that the automated monitoring of workers presented in this chapter cannot be applied to the entire range of construction related tasks. Only if the proposed method was updated with the type of tasks of workers it would be possible to turn also the detected "move" events into labour input.

The method presented in this chapter is not tested on complicated scenarios involving multiple workers performing different tasks. This is mainly due to the time restrictions of this research project. However, the proposed productivity monitoring method is successfully tested on a statistically sufficient sample size which was collected at real construction jobsites. In particular, it features a precision of 95%, a recall of 76% and an accuracy of 76% in terms of detecting work cycles and an average accuracy of 95% in terms of determining the total productive time (labour input) that three workers spent on a steel fixing and an electrical task. Such high performance shows that the proposed productivity monitoring method is accurate, and hence addresses the second aim of this thesis. Finally, the combined results of chapters 3, 4 and 5 address the third and final aim of this thesis as: a) the method presented in chapter 5 monitors the labour productivity of three workers accurately without any prior knowledge about the types of tasks the workers performed, and b) the methods of chapters 3 and 4 show that multiple workers can be tracked along the jobsite simultaneously. These two achievements entail that the framework presented in this thesis can monitor the labour productivity proactively i.e. monitor the labour productivity rates of multiple construction workers at the same time on a daily basis.

6

Conclusions and future work

This final chapter, summarizes the research presented in this thesis, highlights the contributions achieved and suggests a number of directions for future work.

6.1. Conclusions

Labour productivity is the fraction of the labour output over the labour input. In construction, the labour input is equal to the time workers spend on construction tasks, whilst the output quantifies what workers achieved during this time such as the number of concrete buckets poured, the number of steel cages prepared, the meters of brick walls constructed etc.

Chapter 1 describes the problem of labour productivity in construction. It is reported that, since 1960 up to date, the construction sector has not managed to improve labour productivity. On the contrary, the non-farm industries (i.e. part of the domestic economy that does not include activities related to private households, government, farm and no profit organizations) have managed to double their labour productivity. The numerous factors that affect labour productivity in a negative way are behind this lack of growth in construction. Most of these factors, with a percentage of 64.77%, are easily detectable during the monitoring of labour productivity. They are problems that appear on-site and are related to overtime, safety, resources, scheduling, rest areas, transportation, congestion, disruptions, site layout, supervision, rework, skills, fatigue, absenteeism, late arrivals and unscheduled breaks. Considering that these factors are not periodic phenomena, significant amounts of time might be lost until they are detected. Therefore, monitoring of labour productivity should be performed proactively. However, this is not feasible with current practices. Such practices are extremely labour intensive and time consuming as they rely on manual observation and work sampling techniques.

Chapter 2 discusses the current state of research in monitoring of labour productivity. Region-based and activity-based studies focused on the calculation of the labour input in order to achieve this. The labour output is not researched by current studies as its calculation is quite straight forward through visual inspection at the end of work shifts. The region-based studies monitor the labour input by linking

the location of workers to predefined zones of specific management interest (e.g. excavation zone, brick laying zone). If workers are located at the correct zones given their assigned task, then the time they spend at these zones is simultaneously considered productive. On the other hand, the activity-based studies monitor the labour input by detecting and linking the activities of workers (e.g. bending, stretching, sound, strain) to specific tasks (e.g. nailing, brick laying). However, none of the existing studies has proposed a method for monitoring labour productivity of multiple construction workers at the same time, accurately, unobtrusively, time and cost efficiently. Therefore, the main objective of the research conducted in this thesis is to develop a fully automated computer vision-based framework for monitoring labour productivity of construction workers regardless of the type or number of tasks they perform through their work shift. The specific aims are to: a) track workers unobtrusively, b) extract the labour input of construction workers accurately, and c) monitor the labour productivity of construction workers proactively. This thesis hypothesizes that task productivity of construction workers can be monitored through their trajectory data.

Chapter 2 closes with a detailed summary of the overall proposed framework for monitoring labour productivity of construction workers. The research presented in this thesis assumes that all work cycles fit to the same pattern. The total duration of these work cycles per worker is equal to his/her labour input. The proposed framework is formed based on this assumption. It consists of two main methods which are sequentially applied. The first, performs computer vision-based 4D tracking of construction workers. This method uses as input the video data streamed from multiple cameras with overlapping field of view. The output of this tracking method are 4D trajectories of workers that depict their 3D location over time. The second, uses as input these 4D trajectories in order to detect work cycles.

Chapter 3 describes a computer vision-based 2D tracking method of construction workers. This type of tracking matches the same worker across subsequent frames of a single camera (intra camera tracking) and returns his/her 2D trajectories. This tracking method is designed for complex working environments. None of the existing computer vision-based tracking methods has succeeded to track multiple targets like workers that share similar appearance under illumination/scale/posture variations, and abrupt movements in the long term. This is mainly because construction jobsites are complex environments due to congestion, background clutter and occlusions. This chapter proposes a novel computer vision-based method that tackles all these challenges. The proposed 2D tracking method outperforms the latest state of the art method that also focuses on tracking of construction workers. It returns an F-measure metric equal to 72.17% and an average distance error d_{error} of 9.81 pixels compared to the 42.38% and 22.97 pixels respectively of the existing state of the art method that is used for comparison.

Chapter 4 describes a computer vision-based method for matching the same workers across multiple cameras (inter camera) with overlapping view for automated 4D vision tracking of construction workers. This type of matching is challenging due to the greatly similar high visibility apparel of

workers, occlusions, and congestion. This chapter presents a novel method that addresses all these issues. The proposed matching method uses as input the output of the computer vision-based 2D tracking method of Chapter 3**.** It searches for potential matches in three sequential steps. This searching stops only when a positive match is returned for all workers. The first step searches for the strongest match by correlating 1500ms of workers' past 2D trajectories. If this step fails to return a positive match, then the second step applies geometric restrictions in order to define the area within an image that most likely contains a positive match for a worker. If more than one potential match is detected within this geometrically defined area, then the proposed matching method activates the third step that correlates workers' colour intensity values. The proposed matching method features a very promising performance of 97% precision, 98% recall, and 95% accuracy. After all workers are matched across multiple cameras, their 3D locations over time are calculated through the mid-point triangulation method. The successful performance of the method presented in this chapter addresses the first aim of this thesis as workers are tracked unobtrusively.

Chapter 5 describes how the 4D trajectories of Chapter 4 are converted into labour input. This chapter concludes the framework presented in Chapter 2. Existing studies that focus on trajectory analysis for detecting abnormal behavior profiles fail to detect patterns in trajectories of workers due to restrictions of current clustering methods. It has been noted by previous studies, that trajectories must be divided into smaller sub-trajectories in order to detect patterns that "hide" in smaller segments of trajectories. Such segmentation automatically increases the size of the data. In addition, the trajectories of workers contain several patterns due to the large variety of tasks they perform. These patterns depict an equal number of clusters. Considering the large number of workers that must be monitored on a daily basis, initialization clustering parameters should be provided in an automated way for time and cost efficiency. However, existing clustering methods struggle with large data sets and are reliant on user-provided initialization of parameters (e.g. number of clusters). Therefore, this chapter proposes a clustering method that addresses these shortcomings in order to detect repetitive patterns in trajectories of workers. This method is based on the assumption presented in Chapter 2 that *every work cycle is described by two semantic "moves" and one semantic "stop"*. This assumption supports that workers remain at "stop" while performing tasks and "move" only to start new ones. Therefore, the labour input of any worker is equal to the total duration of his/her work cycles. The 4D trajectories are initially partitioned into smaller 4D sub-trajectories in order to detect work cycles of short duration. Then, these 4D sub-trajectories are classified as either semantic "stop" or "move" events. The method proposed in this chapter detects work cycles by clustering the classified 4D sub-trajectories in accordance to the main assumption of this thesis.The detected work cycles are finally returned as unproductive, normal productive, and abnormal productive. The unproductive work cycles are detected through region-based classification. They depict the time workers spend at areas of the jobsite where no construction-related tasks take place directly. Such areas are the rest, office and materials' storage areas given that they are linked to low productivity rates as shown in Chapter 1. The abnormal and normal work cycles are both

considered productive. Their main difference is that the former have by 50% larger duration compared to the latter. This threshold is arbitrarily defined by the researcher as such classification aims only to highlight the work cycles that consumed most of the time of workers through their work shift. This simplifies the assessment of workers' performance if the labour output is not the desirable. The productivity monitoring method of this chapter features an accuracy of 95%, recall of 76% and precision of 76% in terms of detecting work cycles, and an average accuracy of 95% in terms of determining the total time workers spend on construction-related tasks. This time is the actual labour input of construction workers. This good performance addresses the second aim of this thesis. In addition, it is proved that labour productivity of multiple workers can be monitored at the same time as no prior knowledge of either the type or the number of tasks that workers perform is needed. This addresses the third aim of this thesis and proves true our hypothesis that task productivity of construction workers can be monitored through their trajectory data.

6.2. Contributions

This research contributes to the civil engineering community by providing an automated framework for monitoring the labour productivity of construction workers proactively. The developed framework achieves this for multiple workers at the same time regardless of the type or number of tasks that workers perform. More specifically, the contributions of the research conducted in this PhD are the following:

1. Developed a novel computer vision-based method that performs 2D tracking of construction workers in complex environments.
2. Developed a novel computer vision-based matching method for automated and unobtrusive 4D (3D location over time) tracking of construction workers.
3. Developed a novel trajectory analysis-based method for converting the trajectory data of construction workers into labour input accurately.

6.3. Recommendations for future work

The research presented in this thesis tried to address most of the challenges related to monitoring of labour productivity of construction workers. However, there are still some recommendations for future work that could improve the efficiency of the proposed framework.

One such recommendation would be the recognition of workers (i.e. name, task assigned). This could be achieved by attaching on the hard hats of workers unique identification features similar to QR codes. A computer vision-based method could be developed in order to detect firstly the hard hats and

then the QR codes. This additional information could improve both the proposed 4D tracking and the productivity monitoring method. With regards to the former, the proposed 4D tracking method tracks the 3D location of workers over time under the condition that workers are captured within the overlapping view of two cameras at least. If workers leave the view of all cameras and return after a while a new ID will be assigned to them. However, if the recognition system is added, then the proposed framework will attach to these workers their previous IDs the moment they re-enter the view of cameras. This way, all trajectories of the same worker will be automatically sorted out. With regards to the latter, the proposed productivity monitoring method could also generate detailed crew balance charts if the tasks of workers are known. This could be achieved by attaching to every detected work cycle a label such as hammering, pushing wheel barrel, nailing. However, this entails further research on the detection of a large variety of construction-related objects such as hammers, nail guns, wheel barrels, concrete buckets.

A second recommendation relates to the monitoring of earthmoving equipment labour productivity. Although existing studies have been successful for such construction entities, there is still some space for improvement. Earthmoving equipment usually has an optimum productivity pace compared to workers. This is because the number of sub-tasks they perform are very limited. For instance a crane performs only two sub-tasks i.e. loading and unloading, with a relatively steady pace given that the cranes are usually fixed in place. Taking that into account the proposed framework could be able to notify the project managers about when the productivity pace ($work\ cycles/time$) of such entities is less than the expected optimum. The abnormality classification method presented in Chapter 5 could return such information if only the optimum pace was known.

A third recommendation is to apply the framework to a larger camera network that will cover the entire jobsite (both “active” and “inactive” areas). The “inactive” areas as mentioned in Chapter 5, are the rest, office and materials’ storage areas. This way the benefits of monitoring the time workers spend at areas that are not directly related to construction tasks will be better evaluated. In general, the time spent at these areas has been proven to be responsible of low labour productivity rates in construction. Existing studies have proposed tagged based methods. However, such methods have not yet been proven applicable due to the obtrusive nature of this type of monitoring.

Bibliography

Abu-Ain, W., Abdullah, S. N. H. S., Bataineh, B., Abu-Ain, T., & Omar, K. (2013). Skeletonization Algorithm for Binary Images. *Procedia Technology*, *11*(Iceei), 704–709.

Accuware.com. (2017). Video Tracker – Accuracy – Accuware – Support. *[Online] Available at: Https://www.accuware.com/support/video-Tracker-Accuracy/ [Accessed 3 Aug. 2017]*.

Akhavian, R., & Behzadan, A. H. (2016). Smartphone-based construction workers' activity recognition and classification. *Automation in Construction*, *71*(Part 2), 198–209.

Al-qahtani, M. H., S. Ali, M., & Al-Qanhtani, M. (2007). *Productivity Study on a Selected Construction Site*. King Fahad University of Petroleum & Minerals Construction Engineering and Management.

Alinaitwe, H., Mwakali, J., & Hansson, B. (2006). *Labour productivity in the building industry*. Lund University, Sweden. Retrieved from https://lup.lub.lu.se/search/publication/940c69c2-ce84-41f7-a6e0-20d2a5376e15

Allen, J. G., Xu, R. Y. D., & Jin, J. S. (2006). Object Tracking Using CamShift Algorithm and Multiple Quantized Feature Spaces. In *VIP '05 Proceedings of the Pan-Sydney area workshop on Visual information processing* (Vol. 36, pp. 3–7). Sydney, Australia, USA.

AMAC Consultants. (2004). *Productivity Measurement and Analysis*. *Productivity Measurement and Analysis*. The University of British Columbia.

An, L., Chen, X., & Yang, S. (2016). Person re-identification via hypergraph-based matching. *Neurocomputing*, *182*, 247–254.

Axis.com. (2016). Facial recognition | Axis Communications. *[Online] Available at: Http://www.axis.com/ae/en/solutions-by-Application/facial-Recognition [Accessed 9 Dec. 2016]*.

Bäcklund, H., Hedblom, A., & Neijman, N. (2011). DBSCAN A Density-Based Spatial Clustering of Application with Noise. Retrieved from www.itn.liu.se

Bai, Y., Huan, J., & Peddi, A. (2008). *Development of Human Poses for the Determination of On-site Construction Productivity in Real-time*. Final Report, National Science Foundation, 90 pgs.

Barreira, T. V., Roew, D. A., & Kang, Mi. (2010). Parameters of Walking and Jogging in Healthy Young Adults. *International Journal of Exercise Science*, (September 2017).

Bay, H., Tuytelaars, T., & Gool, L. Van. (2008). SURF : Speeded Up Robust Features. *Computer Vision and Image Understanding*, *110*, 346–359.

Beinat, E., Steenbruggen, J., & Wagtendonk, A. (2007). *Location Awareness 2020: A foresight study on location and sensor services*. Report E-07/09, Vrije Universiteit Amsterdam, Spatial Information Laboratory.

Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., & Torr, P. (2016). Staple: Complementary Learners for Real-Time Tracking. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1401–1409). Las Vegas, USA.

Boonstra, A. M., Fidler, V., & Eisma, W. H. (1993). Walking speed of normal subjects and amputees : aspects of validity of gait analysis. *Prosthetics and Orthotics International*, *17*(2), 78–82.

Breitenstein, M. D., Reichlin, F., Leibe, B., Gool, E. K. L. Van, & Leuven, K. U. (2009). Robust Tracking-by-Detection using a Detector Confidence Particle Filter, *i*(Iccv).

Brereton, R. G., & Lloyd, G. R. (2010). Support Vector Machines for classification and regression. *Analyst*, *135*(2), 230–267.

Bügler, M., Ogunmakin, G., Teizer, J., Vela, P. A., & Borrmann, A. (2014). A Comprehensive Methodology for Vision-Based Progress and Activity Estimation of Excavation Processes for Productivity Assessment. In *EG-ICE* (pp. 1–10). Cardiff, United Kingdom.

Bureau of Labor Statistics. (2008). *Employed persons in nonagricultural industries by sex and class of worker*. Retrieved from http://www.bls.gov/cps/cps_aa2008.htm

Business.panasonic.co.uk. (2016). Facial Recognition Technology - Security Solutions | Panasonic Business. *[Online] Available at: Http://business.panasonic.co.uk/security-Solutions/face-Detection-Software [Accessed 9 Dec. 2016]*.

Carrasco, V., Hall, B., & Sweany, J. (2013). *Labor and Productivity Analysis*. Denver International Airport - South Terminal Redevelopment Program.

Čehovin, L., Leonardis, A., & Kristan, M. (2015). Visual object tracking performance measures revisited. *IEEE Transactions on Image Processing*, *v1*, 1261–1274.

Chang, W., Chang, C.-C., Matz, S., & Ho Son, D. (2004). Friction requirements for different climbing conditions in straight ladder ascending. *Safety Science*, *42*, 791–805. https://doi.org/10.1016/j.ssci.2004.02.002

Chawla, K., Robins, G., & Zhang, L. (2010). Object localization using RFID. In *IEEE 5th International Symposium on Wireless Pervasive Computing* (pp. 301–306). Modena, Italy.

Cheikh, F. A., Saha, S. K., Rudakova, V., & Wang, P. (2012). Multi--people tracking across multiple cameras. *International Journal on New Computer Architectures and Their Applications (IJNCAA)*, *2*(1), 23–33.

Chen, J., Qiu, J., & Ahn, C. (2017). Construction worker's awkward posture recognition through supervised motion tensor decomposition. *Automation in Construction*, *77*, 67–81.

Chen, K., Lai, C., & Hung, Y. (2008). An Adaptive Learning Method for Target Tracking across Multiple Cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*. Anchorage, Alaska, USA.

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

Cheng, T., Teizer, J., Migliaccio, G. C., & Gatti, U. C. (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.

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

Choi, J., Moon, D., & Yoo, J. (2015). Robust Multi-person Tracking for Real-Time Intelligent Video Surveillance. *ETRI Journal*, *37*(3), 551–561.

Choi, K., & Seo, Y. (2011). Automatic initialization for 3D soccer player tracking. *Pattern Recognition Letters*, *32*(9), 1274–1282.

Chyronhego.com. (2016). TRACAB Optical Tracking. *[Online] Available at: Http://chyronhego.com/sports-Data/tracab [Accessed 9 Dec. 2016]*.

Comaniciu, D., Ramesh, V., & Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(5), 564–577.

Connell, C. (2015). What's The Difference Between Measuring Location By UWB, Wi-Fi, and Bluetooth? *[Online] Available at: Http://www.electronicdesign.com/communications/what-S-Difference-between-Measuring-Location-Uwb-Wi-Fi-and-Bluetooth [Accessed 5 Oct. 2017]*.

Cundecha, M. M. (2012). *Study of Factors Affecting Labor Productivity at a Building Construction Project in the USA : Web Survey*. M.S. Thesis, North Dakota University.

da Silva, I. N., Hernane Spatti, D., Andrade Flauzino, R., Liboni, L. H. B., & dos Reis Alves, S. F. (2017). Artificial Neural Network Architectures and Training Processes. In *Artificial Neural Networks* (pp. 21–29). Switzerland: Springer International Publishing.

Dai, J., Goodrum, P. M., Maloney, W. F., & Sayers, C. (2005). Analysis of Focus Group Data Regarding Construction Craft Workers. In *Construction Research Congress* (pp. 1–10). San Diego, California, United States.

Dai, J., Goodrum, P. M., Maloney, W. F., & Srinivasan, C. (2009). Latent Structures of the Factors Affecting Construction Labor Productivity. *Journal of Construction Engineering and Management*, *135*(5), 397–406.

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886–893).

Ding, H., & Zhang, W. (2012). Multi-target tracking with occlusions via skeleton points assignment. *Neurocomputing*, *83*, 165–175.

Dozzi, S. P., & AbourRizk, S. M. (1993). *Productivity in Construction*. Ottawa: Institute for Research in Construction, National Research Council.

El-gohary, K. M., & Aziz, R. F. (2014). Factors Influencing Construction Labor Productivity in Egypt. *Journal of Management in Engineering*, (February), 1–9.

Eng, J. (2003). Sample size estimation: How many individuals should be studied? *Radiology*, *227*(2), 309–13.

European Construction Industry Federation. (2017). *Annual Report*. Retrieved from

http://www.fiec.eu/en/library-619/annual-report-english.aspx

Facefirst.com. (2016). Advanced Facial Recognition Technology | FaceFirst. *[Online] Available at: Http://www.facefirst.com/ [Accessed 9 Dec. 2016]*.

Ferguson, H. (2012). Offsite Construction. *Ingenia*, Issue 51. Retrieved from http://www.ingenia.org.uk/Ingenia/Articles/777

FINE. (2017). http://www.projectfine.eu. *Available at: http://www.projectfine.eu/assets/deliverables/D3.1.pdf (Accessed: 11 January 2017).*

Fosler-lussier, E. (1998). Markov Models and Hidden Markov Models: A Brief Tutorial, *1198*(510).

Frinault, Y. (2015). Is construction stuck in the 1960s? *[Online] Available at: Http://www.fieldwire.com/blog/collaboration/is-Construction-Stuck-in-the-1960s/ [Accessed 27 Sep. 2017]*.

Gatti, U. C., Migliaccio, G. C., Bogus, S. M., & Schneider, S. (2014). An exploratory study of the relationship between construction workforce physical strain and task level productivity. *Construction Management and Economics*, *32*(6), 548–564.

Gaxiola, L. N., Diaz-Ramirez, V. H., Tapia, J. J., & García-Martínez, P. (2016). Target tracking with dynamically adaptive correlation. *Optics Communications*, *365*, 140–149.

Gilbert, A., & Bowden, R. (2006). Tracking Objects Across Cameras by Incrementally Learning Inter-camera Colour Calibration and Patterns of Activity. In *9th European Conference on Computer Vision (ECCV)* (Vol. 3952, pp. 125–136). May, Graz, Austria.

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

Gong, J., & Caldas, C. H. (2010). Computer Vision-Based Video Interpretation Model for Automated Productivity Analysis of Construction Operations. *Journal of Computing in Civil Engineering, ASCE*, *24*(6), 252–263.

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

Gor, R. M., & Man, M. (2009). *INDUSTRIAL STATISTICS AND OPERATIONAL MANAGEMENT Chapter 6 : FORECASTING TECHNIQUES*. Retrieved from http://nsdl.niscair.res.in/jspui/browse?type=author&value=Gor%2C+Ravi+Mahendra

Gray, D., & Tao, H. (2008). Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *10th European Conference on Computer Vision (ECCV)* (pp. 262–275). Berlin, Heidelberg.

Grunwald, P. D., Myung, I. J., & Pitt, M. A. (2004). *Advances in Minimum Description Length : Theory and Applications*. Cambridge, Massachusetts.

Guiñón, J. L., Ortega, E., García-antón, J., & Pérez-herranz, V. (2007). Moving Average and Savitzki-

Golay Smoothing Filters Using Mathcad. In *International Conference on Engineering Education – ICEE* (pp. 1–4). University of Coimbra, Portugal.

Halligan, D. W., Demsetz, L. A., Brown, J. D., & Pace, C. B. (1994). Action-response model and. *J. Constr. Eng. Manage*, *120*(1), 47–64.

Halpin, D. W., & Riggs, L. S. (1992). *Planning and Analysis of Construction Operations*. John Wiley & Sons.

Han, M., & Kim, I. (2013). Hue modeling for object tracking in multiple non-overlapping cameras. In *7th International Workshop on Multi-disciplinary Trends in Artificial Intelligence* (pp. 69–78). Krabi, Thailand.

Hanna, A. S., Taylor, C. S., & Sullivan, K. T. (2005). Impact of Extended Overtime on Construction Labor Productivity. *ASCE, Journal of Construction Engineering and Management*, *131*(June), 734–739.

Hartigan, J. A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, *28*(1), 100–108.

Hartley, R. I. (1997a). In defence of the 8-point algorithm. *Proceedings of IEEE International Conference on Computer Vision*, *19*(6), 1064–1070. https://doi.org/10.1109/ICCV.1995.466816

Hartley, R. I. (1997b). In defence of the 8-point algorithm. *Proceedings of IEEE International Conference on Computer Vision*, *19*(6), 1064–1070.

Hildreth, J., Vorster, M., & Martinez, J. (2005). Reduction of Short-Interval GPS Data for Construction Operations Analysis. *Journal of Construction Engineering and Management*, *131*(August), 920–927.

Horman, M. J., & Kenley, R. (2005). Quantifying Levels of Wasted Time in Construction with. *Journal of Construction Engineering and Management*, *131*(January), 52–61.

Hu, W., Zhou, X., Li, W., Luo, W., Zhang, X., & Maybank, S. (2013). Active contour-based visual tracking by integrating colors, shapes, and motions. *IEEE Transactions on Image Processing*, *22*(5), 1778–92.

Jarkas, A. M., & Bitar, C. G. (2012). Factors Affecting Construction Labor Productivity in Kuwait. *Journal of Construction Engineering and Management*, (July), 811–820.

Ji, Q., Zhang, P., & Du, J. (2014). Robust vision tracking by online random ferns and template library. *Signal Processing: Image Communication*, *29*, 590–598.

Jiang, H., Lin, P., Qiang, M., & Fan, Q. (2015). A labor consumption measurement system based on real-time tracking technology for dam construction site. *Automation in Construction*, *52*, 1–15.

Joglekar, J., Gedam, S. S., & Csre, I. I. T. (2010). Image matching with SIFT features--a probabilistic approach. *Proceedings of IAPRS*, *38*, 7–12.

Juels, A. (2006). RFID security and privacy: a research survey. *IEEE Journal on Selected Areas in Communications*, *24*(2), 381–394.

Junejo, I. N., & Foroosh, H. (2007). Trajectory Rectification and Path Modeling for Video Surveillance.

In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007* (pp. 1–7). Rio de Janeiro.

Junejo, I. N., & Foroosh, H. (2008). Euclidean path modeling for video surveillance. *Image and Vision Computing*, *26*(4), 512–528.

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

Khosrowpour, A., Fedorov, I., Holynski, A., Niebles, C., & Golparvar-Fard, M. (2014). Automated Worker Activity Analysis in Indoor Environments for Direct-Work Rate Improvement from long sequences of RGB-D Images. In *Construction Research Congress 2014* (pp. 140–149).

Kong, H., Akakin, H. C., & Sarma, S. E. (2013). A Generalized Laplacian of Gaussian Filter for Blob Detection and Its Applications. *IEEE TRANSACTIONS ON CYBERNETICS*, *43*(6), 1719–1733.

Kos, T., Markezic, I., & Pokrajcic, J. (2010). Effects of Multipath Reception on GPS Positioning Performance. In *52nd International Symposium ELMAR* (pp. 399–402). Zadar, Croatia.

Kuykendall, C. J. O. (2007). *Key factors affeting labor productivity in the construction*. M.S. Thesis, University of Florida.

Lee, J., Han, J., Li, X., & Gonzalez, H. (2008). TraClass : Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering. *Proceedings of the VLDB Endowment*, *1*(1), 1081–1094.

Lee, J., Han, J., & Whang, K.-Y. (2007). Trajectory Clustering : A Partition-and-Group Framework. In *ACM SIGMOD International Conference on Management of Data* (pp. 593–604). Beijing, China.

Lee, L., Romano, R., & Stein, G. (2000). Monitoring Activities from Multiple Video Steams: Establishing a CommonCoordinateFrame. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 758–767.

Li, H., Liu, Y., Wang, C., Zhang, S., & Cui, X. (2016). Tracking Algorithm of Multiple Pedestrians Based on Particle Filters in Video Sequences. *Computational Intelligence and Neuroscience*, *2016*.

Lim, E., & Alum, J. (1995). Construction productivity: Issues encountered by contractors in Singapore. *International Journal of Project Management*, *13*(1), 51–58.

Lim, E. C. (1996). *The analysis of productivity in building construction*. Doctoral Thesis, Loughborough University. Retrieved from https://dspace.lboro.ac.uk/2134/7150

Lin, X. J., Wu, Q. X., Wang, X., Zhuo, Z. Q., & Zhang, G. R. (2016). People recognition in multi-cameras using the visual color processing mechanism. *Neurocomputing*, *188*, 71–81.

Liu, H., & Schneider, M. (2012). Similarity measurement of moving object trajectories. *Proceedings of the Third ACM SIGSPATIAL International Workshop on GeoStreaming - IWGS '12*, 19–22.

Liu, Q., Zhao, X., & Hou, Z. (2014). Survey of single-target visual tracking methods based on online learning. *IET Computer Vision*, *8*(5), 419–428.

Liu, Y., Wang, S., & Chen, Y. Q. (2014). Automatic 3D tracking system for large swarm of moving

objects. *Pattern Recognition*, *52*, 384–396.

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

Madhulatha, T. S. (2012). AN OVERVIEW ON CLUSTERING METHODS. *IOSR Journal of Engineering*, *2*(4), 719–725.

Makulsawatudom, A., Emsley, M., & Sinthawanarong, K. (2004). Critical Factors Influencing Construction Productivity in Thailand. *The Journal of KMITNB*, *14*(3), 1–6.

Martinez, J. C., & Ioannou, P. G. (1994). General purpose simulation with stroboscope. In *Winter Simulation Conference*. Orlando, Florida.

Mazzeo, P. L., & Spagnolo, P. (2008). Object Tracking in Multiple Cameras with Disjoint Views. In *Object Tracking*.

Merrow, E. W., Sonnhalter, K. A., Somanchi, R., & A.F., G. (2009). *Productivity in the UK Engineering Construction Industry* (Vol. 44).

Nasirzadeh, F., & Nojedehi, P. (2013). Dynamic modeling of labor productivity in construction projects. *International Journal of Project Management*, *31*(6), 903–911. https://doi.org/10.1016/j.ijproman.2012.11.003

Nasr, E., Shehab, T., & Vlad, A. (2013). Tracking Systems in Construction : Applications and Comparisons. In *49th ASC Annual International Conference Proceedings*. 9-13 April 2013, San Luis Obispo, California, USA.

Navon, R., & Goldschmidt, E. (2003). Can Labor Inputs be Measured and Controlled Automatically ? *Journal of Construction Engineering and Management*, *127*(August), 437–445.

Navon, R., & Sacks, R. (2006). Assessing research issues in Automated Project Performance Control (APPC). *Automation in Construction*, *16*(4), 474–484.

Ng, S. T., Skitmore, R. M., Lam, K. C., & Poon, A. W. C. (2004). Demotivating factors influencing the productivity of civil engineering projects. *International Journal of Project Management*, *22*(2), 139–146.

Nyström, D., & Per, S. (2009). *Productivity increase valve and pipe assembly An investigation of how to improve the manufacturing process in a large variant production environment*. Master Thesis, Chalmers University of Technology.

Omran, M. G. H., Engelbrecht, A. P., & Salman, A. (2007). An overview of clustering methods. *Intelligent Data Analysis*, *11*, 583–605.

Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. *Proceedings of the 2008 ACM Symposium on Applied Computing - SAC '08*, 863.

Park, H.-S. (2006). Conceptual Framework of Construction Productivity Estimation. *KSCE Journal of Civil Engineering*, *10*(5), 311–317.

Park, M.-W., & Brilakis, I. (2012). Construction worker detection in video frames for initializing vision

trackers. *Automation in Construction*, *28*, 15–25.

Park, M.-W., Palinginis, E., & Brilakis, I. (2012). Detection of Construction Workers in Video Frames for Automatic Initialization of Vision Trackers. In *Construction Research Congress 2012, ASCE* (pp. 940–949).

Park, M. W., & Brilakis, I. (2016). Continuous localization of construction workers via integration of detection and tracking. *Automation in Construction*, *72*, 129–142.

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

Pasha, P., & Monajjemi, Z. (2013). *A Clustering-based Approach for Enriching Trajectories with Semantic Information Using VGI Sources*. University of Twente, Enschede, Netherlans.

Pawłowski, E. (2015). Experimental study of a positioning accuracy with GPS receiver. In *The 12th Conference on Selected Problems of Electrical Engineering and Electronics WZEE*. Kielce, Poland.

Pekuri, A., Haapasalo, H., & Herrala, M. (2011). Productivity and Performance Management – Managerial Practices in the Construction Industry. *International Journal of Performance Measurement*, *1*, 39–58.

Pelekis, N., Kopanakis, I., Marketos, G., Ntoutsi, I., Andrienko, G., & Theodoridis, Y. (2007). Similarity Search in Trajectory Databases. *14th International Symposium on Temporal Representation and Reasoning (TIME'07)*, 129–140.

Peng, P., Tian, Y., Wang, Y., Li, J., & Huang, T. (2015). Robust multiple cameras pedestrian detection with multi-view Bayesian network. *Pattern Recognition*, *48*(5), 1760–1772.

Picard, H. E. (2004). Driving down construction project labor cost. *Journal of Construction Management Association of America*, 1–10.

Projectfine.eu. (2010). D3 . 1 Specification of FINE ' s camera architecture. *[Online] Available at: http://www.projectfine.eu/assets/deliverables/D3.1.pdf [Accessed 20 Apr. 2017]*. Retrieved from http://www.projectfine.eu/

Ross, D. a., Lim, J., Lin, R.-S., & Yang, M.-H. (2008). Incremental Learning for Robust Visual Tracking. *International Journal of Computer Vision*, *77*(1–3), 125–141.

Rote, G. (1991). Computing the minimum Hausdorff distance between two point sets on a line under translation. *Information Processing Letters*, *38*(May), 123–127.

RTD FasTracks. (2014). *[Online] Available at: Http://www.rtd-fastracks.com/i225_112 [Accessed 5 Oct. 2017]*.

Sedehi, A. J. (2010). *Leveraging radio frequency technology identification for productivity analysis in high-rise construction*. M.S. Thesis, Georgia Institute of Technology.

Shah, J. H., Lin, M., & Chen, Z. (2016). Multi-camera handoff for person re-identification. *Neurocomputing*, *191*, 238–248.

Shehata, M. E., & El-Gohary, K. M. (2011). Towards improving construction labor productivity and

projects' performance. *Alexandria Engineering Journal*, *50*(4), 321–330.

Shen, X., Sui, X., Pan, K., & Tao, Y. (2015). Adaptive pedestrian tracking via patch-based features and spatial–temporal similarity measurement. *Pattern Recognition*, *53*, 163–173.

Sun, W., Zhao, C., Chen, L., Li, D., Bai, Y., Jia, W., & Sun, M. (2015). Learning based particle filtering object tracking for visible-light systems. *Optik - International Journal for Light and Electron Optics*, *126*(19), 1830–1837.

Sung, C., Feldman, D., & Rus, D. (2012). Trajectory clustering for motion prediction. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1547–1552.

Suolan, L., Jia, W., & Changyin, S. (2016). Targets Association across Multiple Cameras by Learning Transfer Models. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *9*(1), 185–196.

Suzuki, N., Hirasawa, K., Tanaka, K., Kobayashi, Y., Sato, Y., & Fujino, Y. (2007). Learning Motion Patterns and Anomaly Detection by Human Trajectory AnalysisLearning Motion Patterns and Anomaly Detection by Human Trajectory Analysis. In *IEEE International Conference on Systems, Man and Cybernetics, 2007.* (pp. 498–503). Montreal, Quebec.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Cluster Analysis: Basic Concepts and Algorithms. In *Introduction to Data Mining (First Edition)*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Teicholz, P. (2004). Labor Productivity Declines in the Construction Industry : Causes and Remedies. *AECbytes Viewpoint #4*, *4*. Retrieved from http://www.aecbytes.com/viewpoint/2004/issue_4.html

Teicholz, P. (2013). Labor-Productivity Declines in the Construction Industry : Causes and Remedies (Another Look). Retrieved from http://www.aecbytes.com/viewpoint/2013/issue_67.html

Teixeira, L. F., & Corte-Real, L. (2009). Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition*, *30*(2), 157–167.

Teizer, J., & Vela, P. a. (2009). Personnel tracking on construction sites using video cameras. *Advanced Engineering Informatics*, *23*(4), 452–462.

Thomas, B. H. R., Maloney, W. F., Smith, G. R., Handa, V. K., & Sanders, S. R. (1990). Modeling construction labor productivity. *Journal of Construction Engineering and Management*, *116*(July), 705–726.

Ubisense.net. (2017). Ubisense location intelligence products. *[Online] Available at: Https://ubisense.net/en [Accessed 5 Oct. 2017]*.

Wambeke, B. W., Hsiang, S. M., & Liu, M. (2011). Causes of Variation in Construction Project Task Starting Times and Duration. *Journal of Construction Engineering and Management*, *137*(September), 663–677.

Wang, H., Wang, X., Zheng, J., Deller, J. R., Peng, H., Zhu, L., … Bao, H. (2014). Video object matching across multiple non-overlapping camera views based on multi-feature fusion and

incremental learning. *Pattern Recognition*, *47*(12), 3841–3851.

Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, *34*(1), 3–19.

Weerasinghe, I. P. T., & Ruwanpura, J. Y. (2010). Automated Multiple Objects Tracking System (AMOTS). In *Construction Research Congress* (pp. 11–20). May 8-10, Banff, Alberta, Canada.

Wiliem, A., Madasu, V., Boles, W., & Yarlagadda, P. (2008). Detecting Uncommon Trajectories. In *IEEE Conference, Digital Image Computing: Techniques and Applications* (pp. 398–404). IEE.

Yang, D. K., Chung, P. C., & Huang, C. R. (2014). Unsupervised path modeling across multiple cameras with disjoint views for foreground object tracking. *International Conference on Information Science, Electronics and Electrical Engineering (ISEEE)*, *2*, 1161–1165.

Yang, J., Arif, O., Vela, P. A., Teizer, J., & Shi, Z. (2010). Tracking multiple workers on construction sites using video cameras. *Advanced Engineering Informatics*, *24*(4), 428–434.

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

Yang, J., Vela, P., Teizer, J., & Shi, Z. (2014). Vision-Based Tower Crane Tracking for Understanding Construction Activity. *Journal of Computing in Civil Engineering*, *28*(1), 103–112.

Yang, Y., & Cao, Q. (2013). A fast feature points-based object tracking method for robot grasp. *International Journal of Advanced Robotic Systems*, *10*.

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A Survey. *ACM Computing Surveys (CSUR)*, *38*(4).

Zhang, Z., & Member, S. (2000). A Flexible New Technique for Camera Calibration æ. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, *22*(11), 1330–1334.

Zhao, H., Xiang, K., Cao, S., & Wang, X. (2016). Robust visual tracking via CAMShift and structural local sparse appearance model. *Journal of Visual Communication and Image Representation*, *34*, 176–186.

Zhou, H., Fei, M., Sadka, A., Zhang, Y., & Li, X. (2014). Adaptive fusion of particle filtering and spatio-temporal motion energy for human tracking. *Pattern Recognition*, *47*(11), 3552–3567.

Zhou, Z., Wang, Y., & Khwang, E. (2016). A framework for semantic people description in multi-camera surveillance systems ☆. *Image and Vision Computing*, *46*, 29–46.

Zhu, Z., Ren, X., & Chen, Z. (2016). Visual Tracking of Construction Jobsite Workforce and Equipment with Particle Filtering. *Journal of Computing in Civil Engineering*, *30*(6), 1–15.

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