ABNORMAL EVENT DETECTION IN VIDEO SURVEILLANCE

LIM MEI KUAN

INSTITUTE OF POSTGRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2014

ABNORMAL EVENT DETECTION IN VIDEO SURVEILLANCE

LIM MEI KUAN

THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENT
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF POSTGRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2014

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: LIM MEI KUAN (I.C./Passport No.: 840305-04-5106)

Registration/Matrix No.: WHA 100045

Name of Degree: Doctor of Philosophy

Title: Abnormal Event Detection in Video Surveillance

Field of Study: Computer Science (Computer Vision)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

ABSTRAK

Peningkatan permintaan untuk keselamatan dan ketenteraman awam oleh masyarakat hari ini menjurus kepada penggunaan kamera litar tertutup (CCTV) yang lebih meluas bagi tujuan pemantauan di premis awam. Kes pengeboman di Boston Marathon, dan penculikan seorang kanak-kanak warganegara British di Tasik Titiwangsa, Malaysia baru-baru ini telah membangkitkan lagi kesedaran dan permintaan untuk pelaksanaan video analitik bagi membantu penguatkuasaan undang-undang dalam mencegah jenayah. Siasatan lanjutan sekitar kes-kes tersebut mendapati bahawa tragedi tersebut boleh dielakkan sekiranya terdapat penggunaan teknologi yang dapat mengenalpasti tingkah laku suspek yang mencurigakan. Oleh itu, objektif utama tesis ini adalah untuk membangunkan dan melaksanakan algoritma visi computer untuk mengenal pasti dan mengesan tingkah laku yang mencurigakan, yang boleh menimbulkan jenayah. Pelaksanaan video analitik bermatlamat untuk mencetus peringatan kepada anggota keselamatan untuk pengawasan video yang lebih berkesan dan proaktif.

Sumbangan pertama tesis ini adalah pengenalan algoritma pengesanan visual yang mampu mengesan pergerakan objek dalam video. Salah satu cabaran besar dalam domain ini adalah keupayaan untuk menangani pergerakan objek yang drastik. Contohnya, pergerakan drastik disebabkan oleh pergerakan objek yang sering beralih daripada satu kamera kepada yang lain. Algoritma pengesanan visual yang diusulkan dalam tesis ini mengandaikan masalah pergerakan objek secara drastik sebagai masalah pengoptimuman untuk pengesanan yang lebih berkesan. Keputusan eksperimen menggunakan data awam telah menunjukkan keupayaan algoritma yang diusulkan dalam menangani masalah pergerakan drastik objek di bawah pelbagai senario, dengan ketepatan purata, 91.39%.

Penyelesaian di mana setiap object yang bergerak dikesan terlebih dahulu bagi membuat kesimpulan tentang tingkah laku mereka adalah tidak sesuai dan mustahil apabila

ratusan atau ribuan objek atau orang hadir dalam video, seperti dalam acara-acara besar seperti marathon, di mana orang ramai berkerumun. Oleh itu, sumbangan kedua tesis ini mencadangkan penyelesaian alternatif untuk menangani pengawasan orang ramai dalam kerumun. Algoritma yang dicadangkan menghapuskan syarat untuk mengesan setiap individu. Sebaliknya, algorithma yang diusulkan mengeksploitasi dinamik pergerakan orang ramai secara kerumun. Keputusan eksperimen menggunakan data awam telah menunjukkan keberkesanan rangka kerja algoritma yang dicadangkan dalam mengenal pasti tingkah laku asing dalam kerumun, dengan purata ketepatan lebih kurang 78%. Data awam tersebut merangkumi pelbagai senario seperti kesesakan dan pergerakan yang tidak teratur.

Sumbangan ketiga tesis in bertujuan untuk menyediakan platform yang lengkap untuk mengesan pelbagai aktiviti di kawasan berlainan, yang bersesuaian dengan permintaan di kawasan masing-masing. Ini adalah penting dalam pemantauan seharian, di mana peristiwa yang berbeza boleh berlaku di kawasan yang sama pada masa yang sama. Sebagai contoh, kes melepak dan objek terbiar di kawasan larangan, yang boleh membawa kepada kemungkinan kes keganasan seperti pengeboman berlaku. Penyelesaian yang dicadangkan menyediakan fleksibiliti untuk menangani persekitaran yang berbeza untuk pemahaman yang lebih mendalam tentang tempat kejadian, dengan menggunakan teori yang dikenali sebagai komposisi. Idea utama di sebalik konsep komposisi adalah untuk menguraikan maklumat yang diperolehi dari video yang diberikan kepada beberapa darjah abstraksi sebelum membuat sebarang kesimpulan. Keputusan eksperimen dalam mengenal pasti pelbagai aktiviti seperti melepak, pencerobohan, bagasi yang ditinggal dan dibiarkan, tergelincir dan terjatuh, kekecohan dalam kerumun telah menunjukkan keberkesanan platform yang dicadangkan dengan ketepatan purata, 83%.

ABSTRACT

The recent Boston Marathon bombing and the kidnap of a British boy at Lake Titiwangsa, have ignited a pressing interest for automated video content analysis to assist the law enforcement in preventing such events from recurring. Post-mortem investigations surrounding such cases often found that there were missed opportunities for using technology to detect the abnormality of the suspects, which lead to those tragedies. Therefore, this thesis aims to develop computer vision solutions to identify regions or behaviours, which could lead to unfavourable events, as a cue to direct the attention of security personnel for a more effective and proactive video surveillance.

The first contribution of this thesis introduces a robust visual tracking algorithm that is able to locate moving objects in surveillance videos. A great challenge in this domain is the capability of dealing with complex scenarios of tracking abrupt motion, such as switching between cameras, which is very common when the number of CCTV to be monitored is enormous. Conventional sampling-based predictors often assume that motion is governed by a Gaussian distribution. This assumption holds true for smooth motion but fails in the case of abrupt motion. Therefore, by considering tracking as an optimisation problem, the proposed SwATrack algorithm searches for the optimal distribution of motion model without making prior assumptions, or prior learning of the motion model. Experimental results have shown that the proposed SwATrack improves the accuracy of tracking abrupt motion, with an average accuracy of 91.39%, while significantly reduces the computational overheads, with an average processing time of 63 milliseconds per frame.

Visual tracking of objects at mass gatherings such as rallies can be daunting due to the large variations of crowd. Hence, the second contribution proposes an alternative solution that deals with dense crowd scenes. A new research direction that identifies and

localises interesting regions by exploiting the motion dynamics of crowd is proposed. Here, interesting regions refer to abnormalities, where they exhibit high motion dynamics or irregularities. This assumption alludes to the social behaviours and conventions of humans in crowded scenes. Therefore, the possibility of abnormal events taking place is considered likely, when there is high motion dynamics and irregularities. Experiment results have shown an average accuracy of 78% on the defined dataset.

The third contribution aims to provide an integrated solution to detect multiple events in different regions-of-interest of a given scene. This is very critical in the real-world scenarios where multiple events may take place in a scene at the same time. Existing solutions such as CROMATICA and PRISMATICA are commonly limited to detect single events, at a particular time. On the contrary, the proposed solution provides flexibility to deal with different environments, for a broader degree of scene understanding. The key idea is to conceptually decompose information obtained from a given scene into several intermediate degrees of abstractions. These low-level descriptions are then integrated using a basic set of rule-packages, to discriminate the different events. Experimental results on fives scenarios of abnormal events have shown an average accuracy of 83%.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartiest appreciation and gratitude to my supervisor, Dr. Chan Chee Seng, for being such a tremendous mentor. I am deeply grateful for your words of wisdom, encouragement and guidance throughout my journey in obtaining this degree. Without your unflagging support and active participation in every step of the process, this thesis may never have been realised.

I would also like to thank Professor Paolo Remagnino for being my external supervisor. You were not only supportive during my research visit in the United Kingdom, but have remained involved when I returned to Malaysia. Research has never been a dull moment with your never ending ideas and technical expertise. My warmest appreciation also goes to the various members of the Robot Vision Team in Kingston University.

I am sincerely grateful to Yayasan Khazanah for the opportunity to be one of your sponsored scholars. Without your financial support and continuous motivation, this journey would definitely be more challenging. I would also like to convey special thanks to Dr. Lai Weng Kin, who have continuously encouraged me to take up the challenge of obtaining a PhD. Throughout this journey as a graduate student in UM research lab, there have been unforgettable moments of joy and misery, which have been made more enjoyable by my research colleagues. Thank you for your friendship and for providing such a stimulating research environment.

Last but not least, my profound gratitude to my family. Words cannot express how grateful I am to my dearest mother, father, brother and father-in-law. Thank you for your unconditional love, support and understanding. Your prayers and love have been my constant motivation and driving force. And to my loving husband, Hock Poh Lim, who has been by my side through all weathers, living every single seconds of it, and without whom, I would never have the courage to embark on this.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDIX

CHAPTER 1

INTRODUCTION

Nowadays, Closed Circuit Television Camera (CCTV) systems are rapidly being deployed in public spaces to help strengthen public safety and deter crime (Anderson & McAtamney, 2011). CCTV collects images which are then transferred to a monitor-recording device, to be monitored and stored. A CCTV control room acts as a central hub, where security activities are regulated and coordinated by CCTV operators. The operators are usually responsible for monitoring and reacting to events acquiring attention, which they observe on real-time CCTV videos displayed on the screen displays.

By far, the human vision and perception are highly effective at skimming through large quantity of video sequences and providing high-level semantic interpretation of the scene. However, difficulties arise due to the sheer amount of information and growing number of CCTV to be monitored. Fig. 1.1 illustrates a typical setting in a CCTV control room in Malaysia, where the camera to operator ratio is generally low (approximately 25-30 cameras per operator). In such situation, it is almost impossible to scan and/or follow moving people or objects from camera to camera, thus leading to the risk of missing vital information (Keval & Sasse, 2006). Moreover, the attention span of human has been shown to deteriorate after 20 minutes, while manual monitoring task requires demanding, prolonged cognitive attention (N.-H. Liu, Chiang, & Chu, 2013).

Therefore, over the last few decades, the computer vision community has endeavoured to bring about similar perceptual capabilities to artificial visual sensors (Gong, Loy, & Xiang, 2011). Substantial research efforts have been made towards developing video analytics solutions which are capable to automatically process and analyse the video streams. Generally, video analytics applications can perform a variety of tasks, ranging

from real-time analysis of video for immediate detection of events of interest, to analysis of pre-recorded video for the purpose of extracting events for post-mortem analysis.

(a) (b)

Figure 1.1: A typical CCTV control room in Malaysia, where the camera to operator ratio is low (Star, 2014).

1.1 Motivation

As aforementioned, the increasing demand for security and public safety by society leads to an enormous growth in the deployment of CCTV in public spaces. The heartless murder of an eight year old girl, Nurin Jazlin, at Wangsa Maju, and the most recent kidnap of a British boy, Freddie, at Lake Titiwangsa, have ignited a pressing interest for video analytics solutions to assist the law enforcement in preventing such events from happening again (M. L. Lee, 2014). Most of the post-mortem investigations surrounding such cases found that there were missed opportunities for using technology to detect the abnormality of the suspects, which lead to those tragedies. Nurin was last seen being dragged into a white van from a CCTV recording in the neighbourhood. A week after the abduction, her tortured body was found stuffed in a gym bag and abandoned in front of a shop lot. Again, the crime of abandoning Nurin's body was captured by a nearby CCTV but was left unnoticed, until someone alerted the authority when her body was discovered in the bag. The footage shows a motorcyclist carrying the sports bag where Nurin's body was

squeezed into, before abandoning it at the shop lot. Similarly, a CCTV footage recorded the entire scene when little Freddie was snatched from his mother's arms in front of their house. Snapshots of the footages are as shown in Fig. 1.2. Post-mortem investigations raise the question of whether these horrifying events can be dealt with more efficiently, or even prevented if the abnormalities of the suspect were picked up by the monitoring personnel. Therefore, this research is greatly motivated by the need to grow the role of CCTV for crime control and public safety. In addition to assisting the authority in their investigations in the aftermath of events, CCTV should also act as an extensive round-the-clock solution towards faster respond to potentially catastrophic situations, and ideally, to prevent such tragedies from recurring. Thus, this thesis aims to develop computer vision solutions to identify abnormalities, which could lead to unfavourable events, as a cue to direct the attention of security personnel for a more effective and proactive video surveillance.

(a) The first segment of the footage, lasting about two minutes features the suspect, arriving on motorcycle carrying a blue and black gym bag placed in front of him, before leaving it in front of a shop lot (Ramendran, 2007).

(b) A combo screen grab of the three minutes footage of Freddie's abduction recorded by neighbour (Star, 2013).

Figure 1.2: Sample footages of crime captured on CCTV in Malaysia.

1.2 Activity Understanding and Abnormal Event Detection in CCTV

There has been an accelerated growth in the deployment of CCTV in public places such as communal areas (e.g. parks, pedestrianised streets and parking), public transport infrastructures (e.g. airport, subway and bus stations), sport arenas, recreational centres and shopping malls (Ratcliffe, 2006). Amongst the direct benefits of implementing CCTV in public areas include triggering a perceptual mechanism in a potential offender, reducing fear of crime, assisting the authority in the detection and arrest of offenders, provision of medical assistance, and information gathering for investigations or analysis. However, one of the most difficult and expensive aspects of video surveillance has always been the need to have people monitor these cameras (Fullerton & Kannov, 2008). There would be no opportunity for immediate intervention and action without someone watching and interpreting what the cameras are recording. Typically, the monitoring personnel have all the requisite knowledge and skills for the task, but difficulties arise as the number of cameras grows. Often, many events were unnoticed due to the inherent limitations from depending solely on human monitoring. This is commonly due to i) sheer number of information and screens to be monitored, ii) boredom and human fatigue, iii) distractions and interferences, and finally iv) the complexity and uncertainty of human behaviour. In most scenarios, the consequences of not detecting abnormal activities and events which could ultimately lead to unfavourable events are irreversible and catastrophic. Hence, the last decade has seen significant advances in the field of using computers and technologies for a more proactive video surveillance. This field of research is also widely known as video analytics.

1.2.1 What is Video Analytics?

Generally, video analytics are computer vision algorithms monitoring live or recorded video to understand behaviour and identify 'interesting' events. The automatic analysis of video images takes motion detection to a new level, where analytics utilise local or processing power to detect, recognise or learn 'interesting' events which contextually may be defined as 'suspicious', (Lavee, Khan, & Thuraisingham, 2007), 'irregular', (Y. Zhang & Liu, 2007; Wiliem, Madasu, Boles, & Yarlagadda, 2008), 'unusual ', (Zhong, Shi, & Visontai, 2004; Jäger, Knoll, & Hamprecht, 2008; Jiang, Wu, & Katsaggelos, 2009) or 'abnormal' (Maxion & Tan, 2000; C.-K. Lee, Ho, Wen, & Huang, 2006; Xiang & Gong, 2008; D. H. Hu, Zhang, Zheng, & Yang, 2009; Mehran, Oyama, & Shah, 2009; Varadarajan & Odobez, 2009) in terms of behaviours, events or activities. Cameras are integrated with video analytics solution to recognise various events; from simple recognition task such as whether a moving object is an animal or person, to more complex scenarios such as identifying a particular kind of shoplifting known as sweethearting. Amongst the well-known suppliers of video analytics solutions include CROMATICA, IBM, Bosch, Honeywell, Siemens, Aimetic, Vidient, Panasonic, PRISMATICA, ObjectVideo, VCA, Cisco, Agent Vi, IndigoVision and VCA Technology (Gouaillier & Fleurant, 2009).

Video analytics can assist security personnel by identifying 'interesting' activity or events for closer examination. This indirectly leads to the change of human role from observer to overseer, for a more effective and proactive surveillance. In addition, the recordings of people and activity in a space allow collection of metadata for forensic investigations as well as search for unanticipated events. Nevertheless, despite the many advantages of video analytics, it is important to understand that video analytics are neither fully autonomous nor perfect. Human intervention in the surveillance loop remains necessary as there will always be questionable situations of false alarms and missed events.

1.2.2 What is Activity Understanding?

The goal of intelligent surveillance and analytics solutions is to extend the capability of conventional surveillance to not only detect, classify and track objects in the scene, but to describe or infer the activity or event taking place. Thus far, two different terms including *action* and *event* are used interchangeably in the literature to refer to *activity*.

In this thesis, the following taxonomy as proposed by (Xiang & Gong, 2006; W. Lin, Sun, Poovendran, & Zhang, 2008) is applied. In particular, *activity* comprises sequential actions which are described by a combination of *features* or *attributes*. For example, a set of human activities such as *walking* and *running* can be differentiated using a combination of features, including *body profile* and *speed*. Each feature can be decomposed further to; *body profile = {vertical,horizontal}* and *speed = {slow,fast}*. The *walking* activity is described by vertical body profile and slow speed while *running* is described by vertical body profile and fast speed. Accordingly, *activity understanding* is defined as the establishment of high-level interpretation or semantic description of low-level features (Xiang & Gong, 2006). Since the terms *activity* and *event* are often used in the literature interchangeably, this thesis does not attempt to explicitly differentiate the two. At this point, the term *activity understanding* refers to general activity and do not discriminate between normal and abnormal events in the scene.

1.2.3 What is Abnormal Event Detection?

In the literature, there has been a variety of terms used to refer to abnormal events including *interesting, irregular, suspicious, anomaly, uncommon, unusual, rare, atypical, salient and outlier*. The definition of abnormal events has been causing much debate and confusion in the literature due to the subjective nature and complexity of human behaviors. In particular, they can be categorised into 2 broad understanding, where an event is considered abnormal if:

1. There is deviation from the ordinary observed or learned events (i.e. the event having low occurrence or statistical representation in the learned model)

2. The event is not known or it is outstanding.

Similarly, there is no clear distinction between abnormal activities, events and behaviours as their descriptions often overlap one another. Nonetheless, this thesis deems events as abnormal if they obey the predefined notions that are derived from these common understanding. One section of this thesis which deals with crowded scenes, describes abnormality as regions with high motion dynamics and irregularities in the crowd motion. Meanwhile, in the later section of this thesis, abnormal events are inferred by imposing predefined notions to the set of associated attributes to provide semantic descriptions of activities. This is inspired by the principle of compositionality, that is, the relationship between an object and its associated attribute gives no meaning; unless a rule is applied to the relationship (Pelletier, 1994). Thereafter, the term abnormal and salient are used interchangeably to refer to regions acquiring attention, or precarious.

1.3 Objectives

Often, computer vision solutions for video analytics are organised according to the general pipeline as shown in Fig. 1.3 (Dee & Velastin, 2008). Input videos are firstly fed into motion estimation module to estimate the motion trajectory of moving individuals in the scene. The trajectories are then fed into an analysis module to interpret the activities taking place in the scene. This is followed by higher level analysis and/or event detection module to discriminate between normal and abnormal events, before triggering alert to indicate events acquiring attention.

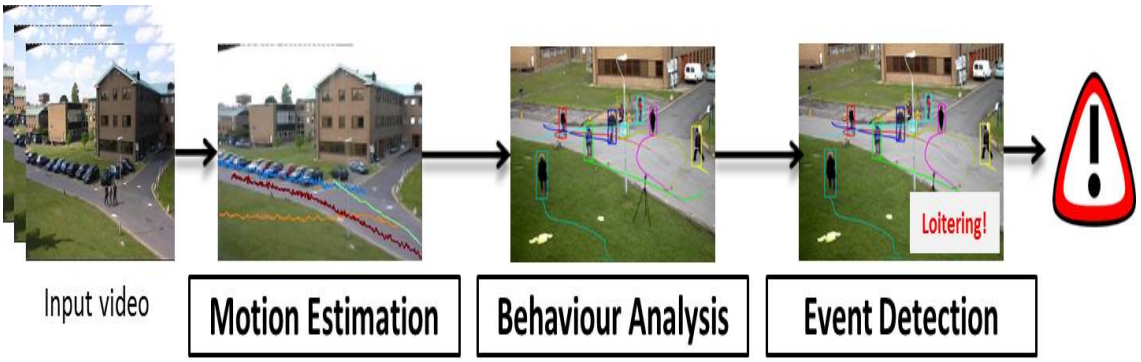


Figure 1.3: The general framework or pipeline of computer vision solutions for video analytics comprising 3 main modules; i) motion estimation ii) behaviour analysis and iii) event detection.

Although the ultimate goal of this work is to devise computer vision algorithms for activity understanding and abnormal event detection in video surveillance, specifically, the thesis is driven towards solving the three main issues in conjunction with the three main modules in the general pipeline of video analytics. The first objective aims to provide a robust visual tracking algorithm that deals with abrupt motion. The second is to identify salient regions, which could ultimately lead to unfavourable events in dense crowd scenes. Finally, the third objective aims to provide an integrated framework to detect multiple events in different regions-of-interest of a given scene.

The following section discusses the underlying challenges, as well as the problem formulation that serves as the motivation of this study towards achieving the aforementioned objectives.

1.4 Challenges and Problem Formulation

Driven by the proliferation of high-powered computers, the availability of high quality and affordable video cameras, and the ever decreasing cost of digital media storage, there is a growing demand for sophisticated video analytics solutions (K. C. Smith, 2007). One of the most fundamental building blocks for complete video analytics solutions is visual tracking. Basically, there are three key steps in video analysis: the detection of

object-of-interest, tracking of such objects from one frame to another, and analysis of their trajectories or other features to recognise their behaviour. The importance of visual tracking stems from the fact that it is pertinent to the tasks of motion based recognition, detection of abnormal events, video indexing, human computer interaction and traffic monitoring (Yilmaz, Javed, & Shah, 2006; H. Yang, Shao, Zheng, Wang, & Song, 2011).

In its simplest form, visual tracking can be defined as the problem of estimating the locality of an object-of-interest as it moves around in the scene. Over the years, significant progress has been made in the area of visual tracking and numerous approaches have been proposed. For instance, the introduction of optical flow (Horn & Schunck, 1981), utilising predictors such as the Kalman Filter (KF) (Welch & Bishop, 1995; Wan & Van Der Merwe, 2000; Oussalah & Schutter, 2000), Particle Filter (PF) (Isard & Blake, 1998; Arulampalam, Maskell, Gordon, & Clapp, 2002; H. Liu & Sun, 2012; Chan, Liu, David, & Kubota, 2008; Chan & Liu, 2009), or linear regression techniques (Ellis, Dowson, Matas, & Bowden, 2011). Tracking is often simplified by imposing constraints on the motion of objects, where the motion is assumed smooth with no abrupt change. It can be further constrained by supposing that the motion is of constant velocity and acceleration based on a priori information.

These assumptions tend to fail in the case of abrupt motion, which is fairly common given the growing number of cameras deployed. Abrupt motion can be caused by inconsistent or rapid speed of the object itself (e.g. the motion of a tennis ball), switching between cameras (i.e. object appears in random part between subsequent frames) and low frame rate videos (i.e. jerky motion). Furthermore, it has always remained a challenge to handle the trade-off between tracking precision and its computational cost. Many of the existing tracking algorithms argue that tracking precision can be improved by increasing the number of particles or samples to represent the conditional state density (Isard & Blake, 1998). However, the increase in the number of particles leads to a surge in the

computational cost. This is a major challenge in most applications, including the analytics proposed in this thesis, where activity understanding and abnormal event detection must often work at real-time, or near real-time rates (Chan, 2008). Therefore, the first challenge of this thesis is on developing a visual tracking algorithm that deals with abrupt motion, while striving for an optimised trade-off between accuracy and processing cost.

Secondly, there can be a range of mass of objects in real world video sequence; from non-crowded scenes comprising less than 3 individuals to dense crowd scenes where hundreds or even thousands gather, as illustrated in Fig. 1.4. While it is ideal to be able to track each moving object to infer their activity, this is not possible at large events such as rallies and marathons, where crowds of hundreds or even thousands gather. Visual tracking of such events is daunting due to the large variations of crowd densities and severe occlusions. Besides, tracking enormous number of individuals in a crowd would require hours or maybe days of processing (Lerner, Chrysanthou, & Lischinski, 2007). Therefore, finding interesting regions in a crowded scene is generally accomplished by firstly learning an activity model of the scene, followed by using the learned model to identify anomalies (Kuettel, Breitenstein, Van Gool, & Ferrari, 2010; Hospedales, Li, Gong, & Xiang, 2011; X. Tang, Wang, & Zhou, 2012). A major drawback of these methods is the need to have a large amount of video data for training during the learning stage. They are not general enough to be adapted to other deployment scenarios. Furthermore, based on the notion that human behaviours are indeed complex and diversified, and the infrequent occurrence of anomalies in real world scenes, learning is made unfeasible. The second challenge is to introduce a framework that identifies and localises interesting regions in crowded scenes, without the requirements of tracking individuals, prior information of the scene or extensive learning.

Thirdly, market research and forecast have shown that surveillance technology is gaining support from numerous governments internationally, in light of various security

threat and protocols required for public space. The last decade has witnessed significant advances in the field of video surveillance systems in Malaysia, particularly. Malaysia video surveillance market was estimated at over US$ 65 million in 2008 with compound annual growth rate of 27% by 2013 (Frost & Sullivan, 2009). Most recently, the analysis of video surveillance market in Malaysia has been discussed further in (J. Lin, 2014), where the actual market revenues for Malaysian market from year 2011 to 2014 demonstrated a steady growth rate of 17%. Furthermore, as tabled in the Malaysian parliament under the budget 2014 themed, 'Strengthening Economic Resilience, Accelerating Transformation and Fulfilling Promises', one of the government's priority is in reducing crime rate, focusing on crime prevention measures. A total of RM3.9 billion fund was injected to increase efficiency of the security force, including the provision of modern and sophisticated equipment; a commendable measure to curb crime (Tun Haji Abdul Razak, 25 October 2013). The enormous growth in the number of CCTVs deployed in public spaces gives rise to the need for video analytics (Shafie, 2008). Often, due to the large number of channels they have to closely observe, as shown in Fig. 1.1, as well as human fatigue, it is extremely challenging for human operators to monitor and analyse the behaviour of each individual that appears in the scene. Thus, a system that can interpret the activity and event of individuals in a constrained environment and trigger the alarm to alert the human operators will be very beneficial towards efficient and proactive surveillance.

Thus far, there have been considerable efforts in the industry as well as academia, focusing on the different algorithms, techniques and models to develop the analytics solutions (Albusac, Vallejo, Castro-Schez, Glez-Morcillo, & Jiménez, 2014; Chen, Wu, Huang, & Fan, 2011; Khoudour et al., 1997; Velastin, Boghossian, Lo, Sun, & Vicencio-Silva, 2005). These systems are commonly designed for specific surveillance applications, which arise in favour of social welfare and public safety and they include traffic monitoring, loitering detection and intrusion detection. However, most of these systems

are focused at detecting singular events at a particular region. Meanwhile, systems that provide multiple events detection are subjected to handling specific scenarios such as railway, and require extensive fine-tuning or learning when deployed in other environment such as an open market. Thus, in summary, there is an open challenge in the domain of analytics to deal with multiple events and provide flexibility in handling the different environments of public surveillance; indoor and outdoor. The enhancement from singular event to multiple events provides a broader degree of scene understanding in video surveillance. Furthermore, usually in real-world scenarios, different events may take place in a particular scene at the same time. For example, it is very likely that a loitering event take place alongside an abandoned object or luggage in a given scene. The third and final challenge is to introduce a framework that deals with multiple events, on different regions-of-interest (ROI), at a particular time, while utilising the low-level features of a given scene, for a broader degree of scene understanding.

1.5 Contributions

This thesis aims to alleviate three major problems in video surveillance which has been discussed in the earlier section. The contributions of this thesis to video-based activity understanding and abnormal events are as the following:

Contribution 1: The first contribution of this thesis aims to provide a robust visual tracking algorithm that deals with abrupt motion. The proposed swarm-intelligence based approach deems tracking as an optimisation problem. Closer work to ours include, (X. Zhang, Hu, Maybank, & Zhu, 2008; Thida, Remagnino, & Eng, 2009; W. Li, Zhang, & Hu, 2009; X. Zhang, Hu, & Maybank, 2010), where the swarm optimisation algorithm, Particle Swarm Optimisation (PSO), is adopted to perform visual tracking. Generally, the interactions and exchange of information between particles are utilised to allow the search for the optimal distribution of motion model.

(a) Sample shots of non-crowd - car park , boxing game and airport scenes.

(b) Sample shots of sparse crowd - airport, basketball game and junction scenes.

(c) Sample shots of dense crowd - stadium, train station and marathon scenes.

Figure 1.4: Sample shots of the variation of crowd in real world scenarios.

However, one major drawback of the traditional PSO is the need to fine-tune parameters as well as the uncertainty in finding the multiple optimal solutions; most work applied fix parameters settings as recommended in (Eberhart & Kennedy, 1995). In contrast, this thesis introduces an adaptive mechanism that detects and responds to changes in the search environment to allow on-the-fly tuning of the parameters. Also, an optimised sampling strategy is presented to trade-off between the exploration and exploitation of the state space, in search for the optimal proposal

distribution. By combining these two sampling strategies within the PSO framework, the proposed tracker allows robust motion estimation, without making prior assumptions, or need to learn the motion model before-hand. Thus, the proposed visual tracking algorithm provides flexibility in tracking smooth and abrupt motion, while keeping the computational cost at its minimal.

Contribution 2: The second contribution of this thesis aims to identify interesting regions, which could ultimately lead to unfavourable events, in crowded scenes, where tracking approaches are not possible. Particularly in dense crowd scenes, such as illustrated in Fig. 1.4c. The proposed method identifies and localises interesting regions or salient, by exploiting the motion dynamics of crowd; where the motion field is projected into global similarity structure to characterise the dynamics of the crowd. Analysing the motion dynamics through the manifold structure has alleviated the need to perform tracking of individuals, prior information requirement or extensive learning to identify instability or abnormal crowd behaviours.

Contribution 3: Finally, the third contribution aims to provide an integrated framework to detect multiple events in different regions-of-interest of a given scene. This is very critical in the real-world scenarios where multiple, different events may take place in a particular scene at the same time. Conventional solutions such as CROMATICA in (Khoudour et al., 1997), PRISMATICA in (Velastin et al., 2005), and EAGLE in (Schwerdt, Bernas, & Paul, 2005) are limited to detect single events only. The proposed solution provides flexibility to deal with the different environments for a broader degree of scene understanding, by utilising the known theory of compositionality into the domain of video content analysis. The key idea is to conceptually decompose information obtained from a given scene into several intermediate degrees of abstractions. These low-level descriptions are then integrated

and combined using a basic set of rule-packages, which discriminate between the different abnormal events to build a complete knowledge of the given scene.

In summary, the collective impact of the three contributions will constitute to a complete video analytics framework that is able to:

- infer activities and abnormal events in video surveillance, and
- assist the law enforcement in preventing events related to security and public safety,

in the hope of building a better and safer society.

1.6 Outline of Research

This thesis is organised into 6 main chapters as described in the followings:

Chapter 1 has presented the overview of video surveillance in general, while providing the motivation and dire need for robust analytics solutions for a more effective and proactive surveillance. In addition, a brief review on the various approaches and understanding of activity understanding and abnormal event detection in the context of surveillance is presented.

Chapter 2 reviews the state-of-the-art solutions and strategies which are relevant to the three broad problem statements that this thesis is addressing. Also, the challenges and current state of the problems are discussed.

Chapter 3 provides detailed explanation on the proposed optimised visual tracking solution, SwATrack that deals with smooth and abrupt motion. It describes and formulates the optimal distribution of motion model in a Bayesian tracking framework. Experimental results challenges the common understanding in the sampling-based tracking approaches, where an increase in the number of samples used will lead to an increase in accuracy. This is motivated by the meta-level question raised by the research community recently, on whether complex and sophisticated methods are really necessary.

Chapter 4 presents detailed description on the proposed method to identify anomalies in crowded scenes using the global similarity structure. It discusses the projection of motion obtained from the optical flow information into the global similarity structure, followed by a two stages segmentation process that combines the outputs of a coarse and fine segmentation to alleviate the need for exhaustive fine-tuning. Specifically, the proposed method investigates regions with high motion dynamics as opposed to conventional methods, where regions with high motion dynamics are often disregarded as noise.

Chapter 5 explains the extension of detection of singular abnormal events to multiple events, while providing the flexibility to deal with the different environments for a broader degree of scene understanding. In particular, the key idea behind the proposed framework, which is the principle of compositionality is discussed and formulated. Experimental results are demonstrated to further validate the effectiveness of the adaptation of the knowledge-based architecture, principle of compositionality as proposed in (Pelletier, 1994) into the domain of video surveillance.

Chapter 6 provides conclusions for each of the three broad problems and suggests a number of areas for future investigations.

CHAPTER 2

BACKGROUND RESEARCH

The United Kingdom is amongst the forerunner in the use of public video surveillance as a primary tool to monitor public activities and prevent terrorism. A substantial amount of fund has been spent by the UK Government on new technology, making it the country with the most security cameras than any other countries in the Europe (La Vigne, Lowry, Markman, & Dwyer, 2011). While surveillance cameras are widely employed in the business sector to improve security, until recently their use to monitor public spaces has been much less common in the United States, in part due to concerns about privacy and civil liberties (Hernon, 2003). Meanwhile, the implementation of public video surveillance in Malaysia, was first carried out in the 1990s and the growth of Closed Circuit Television Camera (CCTV) in public places did not begin in earnest until 2003 (Shafie, 2008).

Traditionally, CCTVs were mainly used to display images on monitors and manned by operators. The implementation of CCTVs satisfies the goal of safe patrolling from a control room, while reducing manpower by performing the role of watchdog or guard. The role of CCTVs can be categorised into two broad categories: *passive* and *active* (Welsh & Farrington, 2008). *Passive* surveillance systems rely upon the retrieval of previously recorded images, which are reviewed after-the-fact as needed, or reactive policing (Welsh & Farrington, 2008). Whereas, *active* surveillance systems are monitored in real time, typically by police or private security personnel for immediate investigation and intervention (Welsh & Farrington, 2008). The effectiveness of active monitoring depends on how frequently the images from each camera are displayed, the ratio of operators to video monitoring screens, and the training that operators receive on how to detect and respond to abnormal activity, or suspicious.

Most recently, the availability of computing power and the reduction in cost within the overall surveillance system has enabled the required demand in automation. Amongst the automated solutions include, video monitoring for specific events such detection of loitering, crowding or grouping and unattended objects. The scientific challenge is to devise and implement computer vision and machine intelligence algorithms that are able to automatically obtain detailed information about the activities and behaviours of people observed by a single camera or by a network of cameras, and alert the operator when necessary. At this point, detection thresholds are usually biased in favour of false positives, since these can usually be quickly recognised and disregarded by human observers. On the other hand, missing real indicators that lead to incidents could be a serious deficiency (Davies & Velastin, 2005). Video analytics solutions can be leveraged to serve as accurate secondary detection, permitting security personnel early notification and the opportunity to investigate, or ideally, to prevent incidents from happening.

The past decades has witnessed the integration of computer vision and machine intelligence techniques for task automation in the video surveillance domain. The computer vision algorithms, especially, have progressed from simple motion-based detections to sophisticated techniques that recognise activities and detect abnormal events. Often, the organisation of computer vision systems is hierarchical and this chapter is organised as such; from visual tracking that deals with abrupt motion, to behaviour analysis in crowded scenes and followed by multiple events detection systems. In fact, majority of the computer vision systems for surveillance are organised in this way, with low-level image processing techniques feeding into tracking algorithms which in turn, feed into higher level scene analysis and/or behaviour analysis modules (Dee & Velastin, 2008). This chapter provides a short overview of each of the respective fields, with the focus on the advantages and disadvantages of these approaches.

2.1 Ongoing Trends in Video Analytics

As technology advances, smarter systems integration is trending to real, cost-effective benefits. In general, analytics solutions demonstrate trends towards the different aspects of video surveillance and are as discussed in the following:

At the edge analytics: Video analytics embedded in a network camera represents a growing segment where applications run "at the edge" with integrated software. Ultimately, analytics algorithms are embedded on dedicated multimedia DSP boards which are directly connected to the sensors. Thus far, the technology which can be embedded into DSP boards are minimal and performs only simple analytics such as motion detection.

Analytics for special events: With events such as the 2008 Beijing Olympics and 2014 Sochi Winter Olympic, which bring millions of tourists to the country, security and surveillance have became an essential priority. Situation management solutions, including video analytics, which can be used by field agents and police to access, live and playback video feeds have been widely used for this purpose. Similarly, upcoming events such as the 2014 World Cup in Brazil will see wide implementation of intelligent surveillance solutions for crime prevention and crowd control.

Complex and customised analytics: In a report published by the National Criminal Justice Reference Service in (Krahnstoever, 2011), the development of a wide range of analytics relevant to law enforcement and corrections are described. These include features of video surveillance that can help to enable early detection and possibly prevention of criminal incidents. Beyond the traditional motion detection, demands for complex activity and understanding such as predict fight and aggression behaviours are increasing. In addition, there are numerous ways to package and customised video analytics to serve, for example, the different requirements and aspects of public and private needs.

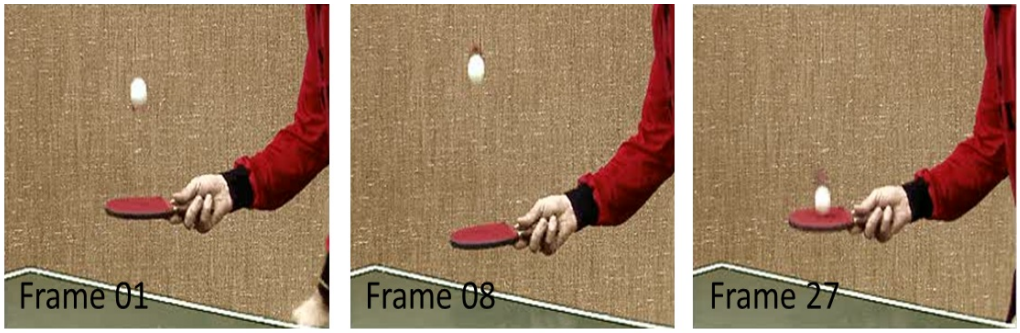
2.2 Visual Tracking

In support of the first trend in video analytics, where simple computer vision algorithms are integrated into multimedia DSP boards for "at the camera" processing, along with the hierarchical organisation of computer vision systems, this section begins with visual tracking. It provides readers with the current state and insights into the problem of visual tracking.

Generally, visual tracking is amongst the most common, yet challenging research topics in computer vision. It represents the basic processing step for most video analytics applications, and its output can be easily interpreted, to provide simple analytics such as intrusion and loitering detection (Höferlin, Höferlin, Weiskopf, & Heidemann, 2011). Consequently, the performances of these applications are significantly dependent on the accuracy and robustness of the tracking algorithms (Dore, Soto, & Regazzoni, 2010). The goal of tracking solutions is to associate foreground pixels over time as belonging to a particular moving or occasionally stationary object.

In this study, the review is narrowed down to statistical correspondence methods, where the task of motion estimation is implemented by sampling-based predictors such as the Kalman Filter (KF) or Particle Filter (PF). These predictors use the state space approach to model the object properties which include its position, velocity and acceleration and thus, are well-suited for estimation of the state of any time varying system. Commonly, their estimation are enhanced by assuming that motion is always governed by a Gaussian distribution, based on Brownian motion or constant velocity motion models (Isard & Blake, 1998; Gustafsson et al., 2002; Weng, Kuo, & Tu, 2006; X. Li, Wang, Wang, & Li, 2010). However, the motion assumptions which are commonly imposed, do not apply to nonlinear systems and tends to give poor estimations of state variables that do not follow the Gaussian distribution, such as when tracking abrupt motion. In this

context, abrupt motion refers to inconsistency and irregularity in motion pattern between adjacent frames. This type of motion may occur in situations where; i) the motion of the target object changes at irregular intervals with unknown pattern (e.g. ball or floating object) and ii) the motion of target is unpredictable due to edited clips acquired from different cameras (e.g. camera switching) or iii) partially low-frame rate sequences (e.g. object motion appears jerky), as shown in Fig. 2.1.



(a) Sample scenario where the motion of the object (table tennis ball) is rapid.

(b) Sample scenario of camera switching. The object (person wearing red shirt) appears in the different position of the scene under varying perspective, captured from different cameras.

(c) Sample scenario of low frame rate. The object (tennis player) moves abruptly in the scene, from frame 1 to 2 and 2 to 3.

Figure 2.1: Sample scenarios where abrupt motion occur.

2.2.1 Tracking Abrupt Motion

A variety of tracking algorithms have been proposed to cope with non-linear systems. While they introduce new extension of the estimator, their work is still bounded by the basis of linearising the non-linear models so that the traditional linear estimator such as Kalman Filter can be applied (Seekircher, Abeyruwan, & Visser, 2011; Julier & Uhlmann, 1997). Although they are more lenient towards Gaussian motion constraint, the different sampling schemes used to estimate the state variable by either hierarchical or recursive estimation are still subjected to a certain degree of motion (Van der Merwe, Doucet, de Freitas, & Wan, 2000; C. Yang, Duraiswami, & Davis, 2005; W. Wu, 2008; Maggio & Cavallaro, 2009). In practice, these methods have well-known drawbacks including unstable filters if the assumption of local linearity is violated, potentially unbounded number of parameters to be fine-tuned, and exponentially increasing number of samples for sampling as the dimensions of state space increases. Therefore, they are only able to deal with nonlinearity to a certain degree and tend to produce poor estimation when tracking abrupt motion.

2.2.2 Statistical Approach for Correspondence

Specifically, there is limited work that deals with abrupt motion. In a recent work in (Zuriarrain, Lerasle, Arana, & Devy, 2008), Markov Chain Monte Carlo (MCMC) was used to overcome the computational complexity in PF as the state space increases while dealing with abrupt motion. Although the proposed MCMC copes better with a high dimensional space, it is still subjected to the common problem of requiring a large number of samples when tracking abrupt motion. Thereafter, there are a number of researchers who introduced modifications and refinements to the conventional MCMC. Kwon *et al.* in (Kwon & Lee, 2008), integrate the Wang-Landau algorithm into the MCMC tracking framework to track abrupt motion. Their method alleviates the constant-velocity motion

constraint in MCMC by improvising the sampling efficiency using the proposed Annealed Wang-Landau Monte Carlo (A-WLMC) sampling method. The A-WLMC method increases the flexibility of the proposal density in MCMC by utilising the likelihood and density of states terms for resampling. In the same way, another variation of MCMC known as the Interactive Markov Chain Monte Carlo (I-MCMC) was proposed (Kwon & Lee, 2010), where multiple basic trackers are deployed to track the motion changes of a corresponding object. The basic trackers which consist of different combinations of observation and motion models are then fused into a compound tracker using the I-MCMC framework. The exchange of information between these trackers has been shown to deal with abrupt motion while retaining the number of samples used. In another advancement, an intensely adaptive MCMC, the Intensely Adaptive Markov Chain Monte Carlo (IA-MCMC) sampler (X. Zhou et al., 2012) has been proposed. Their method further reduces the number of samples required when tracking abrupt motion by performing a two-step sampling scheme; the preliminary sampling step discovers the rough landscape of the proposal distribution (common when there is large motion uncertainty - abrupt motion) and the adaptive sampling step refines the sampling space towards the promising regions found by the preliminary sampling step. In another attempt for effective sampling of abrupt motion, (Kwon & Lee, 2013) proposed the N-fold Wang-Landau (NFWL) tracking method that uses the N-fold algorithm to estimate the density of states which will then be used to automatically increase or decrease the variance of the proposal distribution. The NFWL tracking method copes with abrupt changes in both position and scale by dividing the state space into larger number of subregions. The N-fold algorithm was introduced during sampling to cope with the exponentially increasing subregions.

In another variation of tracking approach in (Wong & Dooley, 2011), the template matching approach is adopted. The possible object positions in every frame are obtained by means of an object detection algorithm, which is a brute force method of searching

the search window for a region similar to the object template defined in the previous frame. The estimation is usually based on a similarity measure such as cross correlation. In contrast to the traditional brute force search, they proposed an adaptive control of the region-of-interest in order to limit the search to the vicinity of detected candidate objects (i.e. table tennis ball). Similar work as proposed in (Comaniciu, Ramesh, & Meer, 2003), the mean-shift procedure is performed to optimised the search window. The mean-shift tracker maximises the appearance similarity of the target iteratively to obtain the translation of the target. These improvements of the brute-force template-based methods however, are still highly dependent on the search window size and incur high computational cost. In addition, images of the fast moving balls are normally blurred, which makes object detection more difficult.

2.2.3 Stochastic Optimisation Approach for Correspondence

Recently, Particle Swarm Optimisation (PSO) (Eberhart & Kennedy, 1995; Van den Bergh & Engelbrecht, 2006; X. Zhang et al., 2008; Tong, Fang, & Xu, 2006; Neri, Mininno, & Iacca, 2013; Sun, Zeng, Pan, Xue, & Jin, 2013), a new population based, stochastic optimisation technique, has received more and more attention because of its considerable success. Variations of PSO methods have been proposed and applied to various applications, including tracking of dynamic systems, evolving weights and structure of neural networks, controlling reactive power or voltage, and simulating crowd behaviour for evacuation planning (Poli, 2008; Ahmed & Glasgow, 2012).

The main difference between the conventional PF and PSO approaches is the interactions between particles in the system. In the PSO, particles interact locally with one another and with their environment, using the analogy of the cooperative aspect of social behaviours of animal swarm such as flocks of birds. In the context of visual tracking, the particles in PSO adjust their velocities dynamically according to their historical per-

formance, as well as their neighbours in the search space, towards finding the optimal proposal distribution of the state. The success lies in the experience-sharing behaviour in which the experience of each particle is continuously communicated to part or the whole swarm, leading the overall swarm motion towards the most promising areas detected so far in the search space, to achieve more accurate estimation.

In what constitutes to the closer work to ours, there are several work which have adopted PSO to perform visual tracking (X. Zhang et al., 2008; Thida et al., 2009; W. Li et al., 2009; X. Zhang, Hu, & Maybank, 2010). Mostly, the interactions between particles in the PSO are leveraged to accelerate the convergence in the search space, where tracking is deemed as a search problem in a high dimensional search space. The underlying idea is that a swarm of particles are deployed around the image to look for the best-fit tracking window. The exploration capability of the particles mitigates the problem known to occur in PF, where the experience-sharing in PSO, deals with sample impoverishment and prevent the search from being trapped in local maxima. The variations of PSO methods for tracking in the literatures vary in terms of the fitness functions used, scales and iterations of search, object detectors and convergence criterion. For example, Zheng and Meng in (Zheng & Meng, 2008) introduce normalised-accumulative histogram to generate the fitness function in order to handle the uncertainty of real-world tracking. In another related work by Yang et al. in (J. Yang, Ji, & Liu, 2011), the fitness function, the fitness function of particles is combined with fuzzy clustering technique for better approximation of the true posterior distribution. Each particle is implemented as a local window classifier that search through the search space. It is important to note that similar to other evolutionary methods, PSO does not use gradient information and thus, can be applied effectively to ill-behaved cost functions. Furthermore, it has been observed, through empirical simulations, that the number of particles and iterations required scale weakly with the dimensionality of the solution space (Saisan, Medasani, & Owechko, 2005), making

PSO a possible solution for tracking abrupt motion. However, one major drawback of the traditional PSO is the need to fine-tune parameters as well as the uncertainty in finding the multiple optimal solutions.

2.2.4 Discussion

In summary, the available visual tracking methods, including the statistical and stochastic optimisation approaches are still subjected to known constraints when tracking abrupt motion. This is due to the dynamic nature of the abrupt motion as well as the need to accommodate to scenario changes in real-world applications. Furthermore, motivated by the meta-level question prompted in (Zhu, Vondrick, Ramanan, & Fowlkes, 2012), this study recognise the need for a trade-off between the precision and computational cost of visual tracking algorithms. From the literature, it is observed that there is a trend towards increasing complexity as tracking algorithms progress. More often than not, these refined methods compensate the increased in complexity in a certain aspect of the algorithm by reducing another aspect of it. Chapter 3 investigates and introduces a novel tracking algorithm inspired by both, the statistical and stochastic optimisation approach to deal with the complexity of tracking abrupt motion while maintaining, if not reducing the computational requirement.

2.3 Behaviour Analysis in Crowded Scenes

One particular class of interest in public security involves a large number of people gathering together, such as in public assemblies (e.g. music festivals, religious events), sport competitions, demonstrations (e.g. strikes, protests). The security of public events involving large crowd has always been of high concern to relevant authorities due to the dynamics and degeneration risk. Various examples from historical incidents have shown how things can easily get out of control when mass of people come together during big events. In a crowd, there is high tendency for emotion spirals (e.g. panic, aggression) to

develop to high levels and the consequences are often devastating. One must understand that at large events, where crowds of hundreds or even thousands gather, video monitoring is a daunting task. This alludes briefly to the second trend of video analytics, where there is an increasing demand for analytics during special events and mass gatherings, in particular. Automatic behaviour analysis in crowded scenes is often intended for pre-screening of scenes for better coordination and control of crowd activities.

Generally, behaviour analysis in crowded scenes faces two fundamental challenges in computational complexity and uncertainty. Human, which is the class of object in this context, demonstrates complex spatio-temporal dynamics, and large variations in a highly dynamical and uncertain environments. For instance, people tend to react and response differently with different environment. Thus, segmenting and categorising these behaviours and their related activities are ill-posed. Understanding and interpreting behaviours are not as straightforward. Limited research efforts have been done in developing computer vision algorithms that deal with high density or extremely crowded scenes (Ali & Shah, 2007) due to its complexity. The recent years however, has shown rapid growing interest in the research community (Gong & Xiang, 2011).

The general approach for crowd behaviour analysis and modelling contains the steps of, detecting moving objects, tracking the targeted objects, finally analysing their trajectories to identify the dominant flows, or to model atypical motion patterns. (Thida, Yong, Climent-Pérez, Eng, & Remagnino, 2013; Vishwakarma & Agrawal, 2013). Based on this pipeline, this section discusses the different categories of crowd modelling and the various approaches in extracting crowd properties for subsequent analysis to infer abnormal activities.

2.3.1 Categorisation of Crowd

Thus far, there is yet any agreed, detailed definition of crowd. Although the multiple descriptions of crowds are vague, and relevant to the varying context, they share common characteristics in several aspects: i) size – large gathering of people, ii) density – crowd members should be in a particular area, with sufficient density distribution, iii) time – crowd members should come together at a specific location for a specific purpose over a measurable time period, iv) collectivity – crowd members should share a social identity (i.e. common goal and interest) and act in a coherent manner. (Challenger, Clegg, & Robinson, 2010).

There are three distinct philosophies for modelling a crowd in the literature, where crowd models are defined as microscopic, mesoscopic and macroscopic (Zhan, Monekosso, Remagnino, Velastin, & Xu, 2008). Microscopic model focuses on the individual level while macroscopic deals with the crowd as a whole, and concern the collective observable behaviours emerging from crowd. Mesoscopic on the other hand, combines properties of the two extremes, where the characters of individuals are kept while maintaining a general view of the crowd, entirely.

2.3.2 Properties of Crowd

Various efforts have been done to detect and track objects in order to generate reliable trajectories. The outputs of tracking algorithms can either be used for higher-level analysis using the tracks and mining trends in a bottom-up approach of crowd analysis; or, conversely, the properties (e.g. trajectory, velocity) obtained from tracking algorithms can be further refined by using cues obtained from crowd analysis using the top-down approach (Thida et al., 2013). However, the complexity of most tracking algorithms is very much influenced by the density of crowd, context and environment in which the tracking is performed. Visual tracking of individuals becomes more challenging as the

density increases from pedestrian ↔ group ↔ crowd (Jo, Chug, & Sethi, 2013). While there are ample tracking approaches for the various density of crowd, the focal point of this discussion is on work related to identifying abnormality in crowd.

2.3.2 (a) Microscopic Modeling

In a bottom-up approach as proposed in (Stauffer, 2003), the trajectories are obtained from background subtraction and correspondence-based tracking of the blobs over a period of time. Due to the high possibility of tracking failures in different environments, (Stauffer, 2003) further introduced a connecting mechanism that associates fragmented tracking sequences to improve the tracking correspondence. His method aims to provide the semantics knowledge of the environment (i.e. sources and sinks), based on the assumption that individuals tend to appear and disappear at particular locations that correspond to sources and sinks. Other similar extensions of using refined trajectories to obtain the semantics of the scenes are as discussed in (Makris & Ellis, 2005; X. Wang, Tieu, & Grimson, 2006; Nedrich & Davis, 2010). Although these methods work well, up to a certain extent, in sparse crowd scenes, they are not suitable to deal with dense crowd scenes. Tracking in dense crowd scenes is extremely challenging, given the complexity arise due to the interactions and occlusions between individuals in the crowd. Furthermore, based on trajectory, or spatio-temporal path alone is not sufficient to detect other scenarios of abnormality in crowd. In practice, the trajectory property is often used to infer direct abnormality such as sources and sinks or dominant flows. They are not well-suited to infer higher-level semantics which requires further analysis on the interactions within crowd members and their environment, including unstable flow, counter flow direction or bottleneck.

2.3.2 (b) Macroscopic Modeling

In order to alleviate the need to track individuals in the scene, in the recent years, researchers have proposed holistic approaches that exploit the contextual information to improve flow understanding in crowded scenes (Jacques Junior, Raupp Musse, & Jung, 2010). Rather than computing the trajectories of individuals, holistic approaches build a crowd motion model using the instantaneous motions of the entire scene such as the flow field (Andrade, Blunsden, & Fisher, 2006; M. Hu, Ali, & Shah, 2008). The flow field is then fed into clustering methods such as the Hidden Markov Models to group the coherent motion patterns of a given scene. Likewise, a multi-scale representation of motion features (i.e. direction and speed) extracted from optical flow and low-pass filtering is used to represent crowd motion (Mancas, Riche, Leroy, & Gosselin, 2011). Other proposals such as in (Ali & Shah, 2008) utilises the contextual information to track unstructured crowd scene. (Ali & Shah, 2008) method assumes that all particles, which represent individuals in the crowd, are moving towards a unified direction. The work has demonstrated an improvement in tracking of individuals in the dense crowd scenes. However, the floor field representation is still limited to having one dominant motion in crowd, and thus is not suitable when the crowd motion is random with different groups of individuals moving at different directions. In another variation (Mehran et al., 2009), the interactions of targets are integrated to model the crowd flow. Subsequently, the energy function obtained from the social information and physical constraint in the environment is used to model the normal behaviour of the crowd. Rodriguez *et al.* in (S. K. T. Rodriguez M.and Ali, 2009) employ correlated topic model to model the random motion in unstructured crowded scenes. Similarly in (Solmaz, Moore, & Shah, 2012; Kratz & Nishino, 2010), the learned motion patterns are incorporated into the motion model for accurate prediction of the local spatio-temporal patterns that describe the motion of individuals

in highly dense scenes. Although these methods provide accurate "holistic" tracking of individuals in crowded scenes, they require exhaustive learning and prior information of the scene, which may not be practical in most real world scenarios. Furthermore, the contextual information are often not available in most scenarios.

2.3.3 Abnormality in Crowd

Detecting abnormality in crowded scenes has gained growing research efforts. Automated detection sources and sinks for instance, can aid in providing contextual information for semantic modelling or scene understanding. On the other hand, automated detection of abnormality caused by congestion may help avoid unnecessary overcrowding or clogging for traffic monitoring and crowd control. Discoveries of other abnormality such as instability and counter flow direction may provide cues to alert the authority of potential threats or incidents.

The definition of abnormality has been causing much debate in the research community, due to the subjective nature and complexity of human behaviours. Some researchers consider any deviation from the ordinary observed events as abnormal, whereas others consider rare or outstanding event as atypical. In this study, the term salient is used to refer to interesting regions acquiring attention, or potentially precarious.

Finding interesting regions in a given scene is generally accomplished by firstly learning an activity model of the scene, followed by using the learned model to identify anomalies (Kuettel et al., 2010; Hospedales et al., 2011; X. Tang et al., 2012). For example, the tracking approaches keep track of each individual motion and further apply a statistical model of the trajectories to identify the semantics or geometric structures of the scene, such as the walking paths, sources and sinks. Then, the learned semantics are compared to query trajectories in order to detect anomaly. In a more flexible approach which detects a broader scope of salient region is proposed by Ali *et al.* (Ali & Shah,

2007). Their method utilises the lagrangian particle dynamics to segment regions based on the motion stability. Then, abnormality is discovered by comparing the segmented region with the learned model.

Detection and localisation of salient regions by using spectral analysis is proposed in (Loy, Xiang, & Shaogang, 2012). In contrast to other methods, their method suppresses dominant flows with a focus on the motion flows that deviate from the norm. While their method deals with unstable crowd flow, their experiments were limited to the detection of simulated instability, and not real-time public scenes. In the closest work to this study, Solmaz *et al.* (Solmaz et al., 2012) propose a linear approximation of the dynamical system to categorise the different behaviours of crowd by observing their eigenvalues over an interval of time. Their method shows promising results in detecting five different scenarios of detailed saliency that includes bottlenecks, lane, arch, fountain-head and blocking.

2.3.4 Discussion

A major drawback of methods that require a learning stage is the need to have a large amount of video data for training, which is tedious and tend to be specific to a particular learned scenario. Usually, they are not general enough to be adapted to other domains, or deployment scenarios. In addition, they are restricted to identifying specific causes of abnormality such as wrong direction and do not deal with other complex scenarios of saliency. Meanwhile, a major shortcoming of the deterministic approaches is the incapability to detect instability when there is lack of consistent characteristic flow. Furthermore, based on the notion that human behaviours are indeed complex and diversified, the categorisation of human behaviours into predefined distinct categories, such as in (Solmaz et al., 2012), may be an oversimplification.

2.4 Abnormal Event Detection

Since the first appearance of CCTV cameras in the early 1950s, video surveillance technologies have undertaken several generations of continuous evolutions and the pace becomes faster over the last decade (Xu, 2007). Technological advancements have led to the development of semi-automatic systems, known as the second generation surveillance systems, where algorithms for automatic real-time detection of events, or analytics to aid the user in identifying events acquiring attention are implemented. Most recently, the development of a wide range of analytics beyond the traditional event detectors are explored, leading the way to third generation surveillance systems. This term is sometimes used in the literature to refer systems conceived to deal with a large number of cameras, a geographical spread of resources, many monitoring points, and to mirror the hierarchical and distributed nature of the human process of surveillance (Velastin & Remagnino, 2006). In short, the surveillance space is growing along with the increasing demand for higher level reasoning and scene understanding for effective surveillance. The race is now on providing integrated, customised solutions that cater to public and private needs, and at the same time, are able to handle practical surveillance problems. While great strides have been made in developing more advanced analytics solutions that deals with more complicated events, this section provides a brief overview of the state-of-the-art solutions, which are divided into two broad categories, namely the shape or pattern recognition based and machine learning based approaches (Benezeth, Jodoin, & Saligrama, 2011). For comprehensive review on the various approaches of video analytics solutions including the neural network, topic model and syntactic approaches, please refer to the survey paper by Turaga *et al.* (Turaga, Chellappa, Subrahmanian, & Udrea, 2008).

2.4.1 Rule or Pattern Based Approach

Generally, the rule or pattern recognition based approach requires prior information regarding the object, or abnormal events that are of interest. In fact, most video analytics suppliers such as Bosch, Honeywell, Axis and iOmniscient adopt this approach for implementation. Using this approach, the analytics or event detectors are packaged into modules which are predefined and can be packaged accordingly to suit the needs of various deployment scenarios. Examples of such event detectors include i) perimeter intrusion (Jodoin, Konrad, & Saligrama, 2008), where simple motion descriptor based on the dynamics of luminance and colour profiles are observed over a period of time to detect abnormality, ii) abandoned object detection (Tian, Feris, & Hampapur, 2008), the refined foreground regions are analysed based on user defined parameters to determine abandoned or removed object scenarios, iii) loitering detection (Nam, 2013), where predefined time span are imposed to consistent tracking outputs for detection. These methods usually either perform comparison between the properties (i.e. low-level attributes) of the query object and the learned templates for detection, or apply specific set of rules to the extracted properties to infer an event.

The need to have prior information or rules that define a particular abnormal event is not always applicable, especially given the rare occurrence and unpredictability of abnormal event. Therefore, alternative approaches which impose set of rules or knowledge to infer events usually prior define *normal* patterns. An event is deemed *abnormal* if it is non-conforming or deviates from the defined norm. In addition, these methods are often implemented using the general pipeline of video analytics framework; where objects are firstly detected via motion descriptors, classified and tracked over a time frame, and finally the resulting trajectories are matched to the defined rules to discriminate between a normal and abnormal event (Buxton, 2003). Despite the fact that trajectory-based analysis

has been proven successful in many applications, these methods are specific, and do not cope well with complex scenarios of abnormal events (Morris & Trivedi, 2008). Since they follow the general pipeline from pre-processing to low-level feature extraction to high-level interpretation, these methods are prone to error propagation. For instance, the erroneous estimation of motion will lead to incorrect tracking, which will subsequently result in false detection of events. Furthermore, rule or pattern based recognition are often specific to certain defined scenarios, and thus, are rigid and cannot be easily extended.

2.4.2 Machine Learning Based Approach

In order to overcome the limitations and inflexibility of rule based recognition, a great diversity of methods based on learning approach have been introduced. Flexible models have evolved in the machine learning community, and adopted into the computer vision arena over the years. These models encompass a wide class or parametric models such as the Gaussian Mixtures (GM), Hidden Markov Model (HMM) and Bayesian Belief Networks (BBN) (L. Wang, Cheng, Zhao, & Pietikäinen, 2011). Typically, learning based approach consists of a matching procedure that compares a measured sequence to the pre-learned models or labelled sequences that represent events; and need to be learnt by the system via a learning stage. In contrast to the rule-based approach, learning-based algorithms are more flexible and deal with the lack of information on abnormal events more effectively, since they do not require prior knowledge. Examples of analytics which applied learning models include i) crowd dispersal detection, where the interaction force in crowd is estimated using the social force model and mapped into the feature space to obtain the dynamics of the interaction; this is then followed by a learning stage using the bag of words approach to infer abnormal event (Mehran et al., 2009), ii) loitering detection, where a linear discriminant approach is used to classify and recognise pedestrian by correlating their appearance feature; time stamps collected with the snapshots in the corre-

sponding database are then used to infer loitering event (Bird, Masoud, Papanikolopoulos, & Isaacs, 2005), iii) running detection, where hierarchical classifiers using HMM and its variation are used to discriminate between normal and abnormal events (Aliakbarpour et al., 2011). Due to the need to determine the basis prior to classification, learning approach often requires training stage on a set of example data before it can be used to discriminate events. In some models such as the BBN, large numbers of training data or extensive fine tuning are required. A great deal of refinements on the basic learning models have been proposed, such as Probabilistic Topic Model (PTM) which requires less computational cost and are less sensitive to noise in comparison to the standard BBN. However, the high learning cost required by most learning methods hinders them from dealing with incremental learning; which is still a topic of debate at present.

2.4.3 Discussion

There are a number of established providers of analytics thus far. CROMATICA (Khoudour et al., 1997) - Crowd Monitoring with Telematic and Communication Assistance combined video-analysis based technologies and wireless data transfer to improve surveillance in public transport systems. Their method deals with multiple events such as intrusion and unattended object but is limited to an indoor environment. PRISMATICA (Velastin et al., 2005) - Pro-active Integrated systems for Security Management by Technological Institutional and Communication Assistance is a distributed system with automated event detection to improve the safety in public transport. The system components were tested in a real world environment and achieved satisfactory results. Although their method deals with a certain degree of crowded scene, it is limited to an indoor environment. Fuentes and Velastin (Fuentes & Velastin, 2004) proposed a framework that utilises low-level descriptions such as the centroid position of blobs to infer events. An extension of this work to include not only the low-level features, but also the high-level

descriptions was discussed in detail, in (Schwerdt et al., 2005). This project, which is also known as the EAGLE project, shows satisfactory evaluation results using the Challenge of Real-time Event Detection Solutions (CREDS) dataset. Similarly, Black *et al.* (Black, Velastin, & Boghossian, 2005) evaluated their proposed real-time surveillance system for metropolitan railways in the United Kingdom and Italy using the CREDS dataset. A more recent work by in (Fernández-Caballero, Castillo, & Rodríguez-Sánchez, 2012), atomised or divided low-level human actions into smaller components and used these components as grammars to infer an event. Again, their method is limited to indoor environment although they cope well with crowded scenes.

In summary, current analytics solutions often act separately to detect multiple events in different scenarios. For example, systems that perform loitering detection or/and abnormal trajectory in a given scene is based on two separate modules that work independently. Thus, they are usually not flexible or general enough to allow detections of different events at one time. There is still an open challenge for a solution that deals with multiple events and provides flexibility in handling different environments (e.g. indoor and outdoor) to meet the rising demand from the public and private sectors.

2.5 Summary

In summary, there are still challenges and issues in conjunction with the three main modules in the general pipeline of computer vision solutions in video analytics as shown in Fig. 2.2.

Tracking Abrupt Motion: Generally, tracking is often simplified by imposing constraints on the motion of objects, where the motion is assumed smooth with no abrupt change. They are further constrained by supposing that the motion is of constant velocity and acceleration based on a priori information (F. Yan, Christmas, & Kittler, 2005; Maggio & Cavallaro, 2009; Adam, Rivlin, & Shimshoni, 2006; Kwon

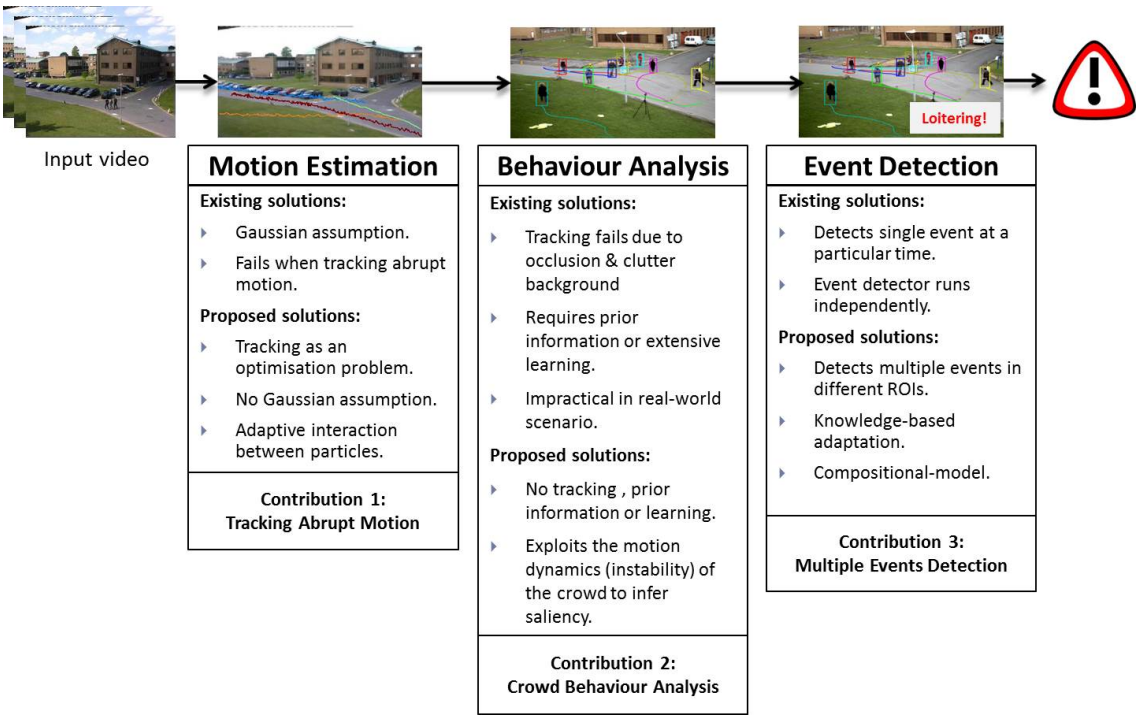


Figure 2.2: The summary of the issues in the three main modules of the general pipeline of computer vision solutions for video analytics, that serves as the motivations of this thesis towards achieving the respective objectives.

& Lee, 2008; Wong & Dooley, 2011; X. Zhang et al., 2008). These assumptions tend to fail in the case of abrupt motion, which is fairly common given the growing number of cameras deployed today. Therefore, the first contribution of this thesis aims to develop a visual tracking algorithm that deals with abrupt motion, while striving for an optimised trade-off between accuracy and processing cost.

Crowd Behaviour Analysis: Visual tracking of objects in dense crowd scenes such as rallies can be daunting due to the large variations of crowd. Thus tracking methods are not well-suited in dense crowd scenes where hundreds or even thousands gather. Meanwhile, holistic approaches that alleviate the need to track each individual in crowded scenes often requires learning stage to detect abnormal events (Kuettel et al., 2010; Hospedales et al., 2011; X. Tang et al., 2012). A major drawback of these methods is the need to have a large amount of video data for training during the learning stage. Hence, the second contribution proposes an alternative

solution that deals with dense crowd scenes. A new research direction that identifies and localises interesting regions by exploiting the motion dynamics of crowd is proposed.

Multiple Events Detection: There are various works in the industry and academia that provide analytics solutions (Albusac et al., 2014; Chen et al., 2011; Khoudour et al., 1997; Velastin et al., 2005). However, most of these systems are focused at detecting singular events at a particular region. Meanwhile, systems that provide multiple events detection are subjected to handling specific scenarios such as railway, and require extensive fine-tuning or learning when deployed in other environment such as an open market. Therefore, the third part of this study introduces a framework that deals with multiple events, on different regions-of-interest (ROI), at a particular time, while utilising the low-level features of a given scene, for a broader degree of scene understanding.

CHAPTER 3

ABRUPT MOTION TRACKING

Visual tracking is one of the most important and challenging research topics in computer vision. Its importance stems from the fact that it is pertinent to the tasks of motion based recognition, automated surveillance, video indexing, human-computer interaction, vehicle navigation and video analytics (H. Yang et al., 2011; Yilmaz et al., 2006). While considerable research exist in relation to visual tracking, only a handful deals with abrupt motion (Wong & Dooley, 2011; Kwon & Lee, 2013; H. Liu & Sun, 2012; X. Zhou et al., 2012). Abrupt motion can be defined as situations where the object motion changes between adjacent frames with unknown patterns in scenarios such as i) partially low-frame rate, ii) switching of camera views in a topology network or iii) irregular motion of the object as shown in Fig. 3.1 (Kwon & Lee, 2008). These scenarios are extremely common given the enormous number of cameras deployed in public scenes.

Thus, in this chapter, the focus is on addressing the problem by casting tracking of abrupt motion as an optimisation problem. A novel abrupt motion tracker which is inspired by swarm intelligence, known as the Refined Swarm-based Abrupt Motion Tracking (SwATrack) is proposed. Unlike existing swarm-based filtering methods, the SwATrack presents an optimised sampling strategy to trade-off between the exploration and exploitation of the state space, in search for the optimal proposal distribution. A fine-tuning mechanism, known as the Dynamic Acceleration Parameters (DAP) that allows on-the-fly tuning of the best mean and variance of the distribution for sampling is also introduced. By combining these two sampling strategies within the Particle Swarm Optimisation (PSO) framework represents a novel method to address abrupt motion. To the best of the author's knowledge, this has never been done before. Experimental results

with a consolidated dataset using benchmarked sequences are presented. The quantitative and qualitative results demonstrate the effectiveness of the proposed method in terms of dataset unbiased, object size invariant and fast recovery in tracking the abrupt motions. Finally, a comprehensive discussion on the findings and comparisons with the state-of-the-art solutions conclude this chapter.

This chapter is structured as the following: a brief introduction on abrupt motion and the conventional Bayesian tracking framework are described in Section 3.1. Section 3.2 explains the standard Particle Swarm Optimisation algorithm as well as its limitation in dealing with abrupt motion. This is followed by a detailed description of the proposed SwATrack tracking solution in Section 3.3. Experimental results are reported and discussed in Section 3.4. Specifically, the effectiveness of the proposed approach in tracking abrupt motion is evaluated using benchmarked videos which exhibit various challenging scenarios of abrupt motion. Further evaluations on the research questions raised in this chapter is discussed in Section 3.5. Finally, the conclusions are drawn in Section 3.6.

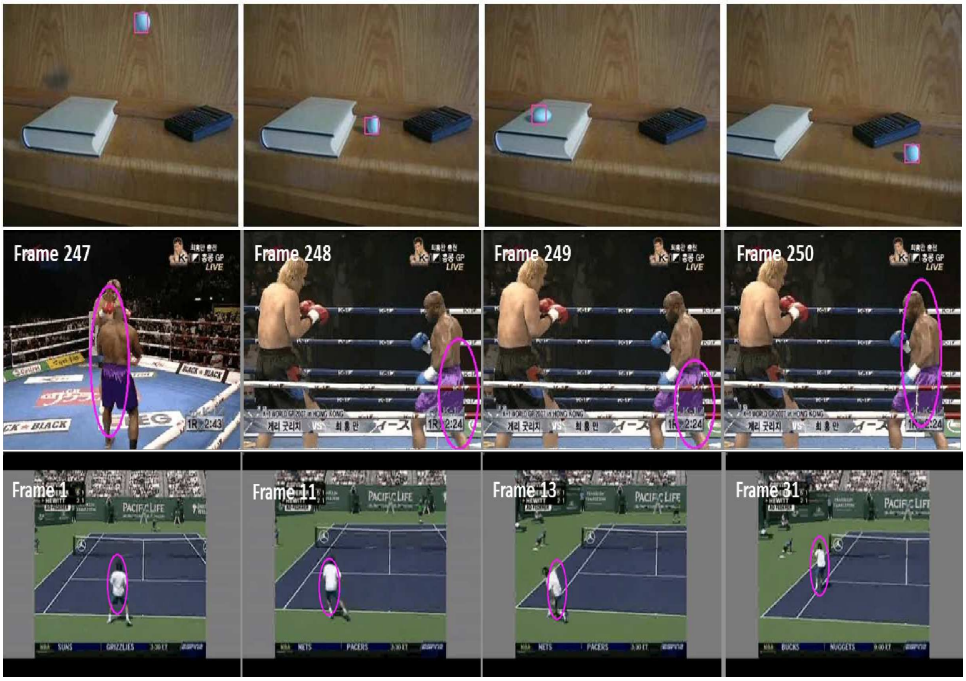


Figure 3.1: Example of the abrupt motion in different scenarios. Top: Inconsistent speed. Middle: Camera switching. Bottom: Low frame-rate videos.

3.1 Abrupt Motion

The task of motion estimation is usually implemented by utilising sampling-based predictors such as Kalman Filter (KF) (Welch & Bishop, 1995; Wan & Van Der Merwe, 2000; Oussalah & Schutter, 2000), Particle Filter (PF) (Isard & Blake, 1998; Arulampalam et al., 2002; H. Liu & Sun, 2012; Chan et al., 2008; Chan & Liu, 2009), or linear regression techniques (Ellis et al., 2011). These predictors are commonly enhanced by assuming that motion is always governed by a Gaussian distribution based on the Brownian motion or constant velocity motion models (Yilmaz et al., 2006; Cifuentes, Sturzel, Jurie, & Brostow, 2012).

While this assumption holds true to a certain degree for smooth motion, it tends to fail in the case of abrupt motion such as inconsistent speed (e.g. the movement of ball in sport events), camera switching (tracking of subject in a camera topology) and low frame-rate videos, as illustrated in Fig. 3.1. The main reason is that the state equation cannot cope with the unexpected dynamic movement, e.g. sudden or sharp changes of the camera/object motion in adjacent frames. These sampling-based solutions also suffer from the well-known local trap problem and particle degeneracy problem. In order to handle these problems, one of the earliest work (Y. Li, Ai, Yamashita, Lao, & Kawade, 2008) considered tracking in low frame rate videos. Their work considers tracking in low frame rate as abrupt motion, and proposed a cascaded PF to solve this problem. This is then followed by a number of sampling strategies (Kwon & Lee, 2008, 2010, 2013; X. Zhang, Hu, Wang, et al., 2010; X. Zhou et al., 2012; Xia, Deng, Li, & Geng, 2013; F. Wang & Lu, 2012), which are incorporated into the standard Markov Chain Monte Carlo (MCMC) tracking framework. Their method alleviates the constant velocity motion constraint in MCMC by improvising the sampling efficiency.

Although the aforementioned works have shown satisfactory results in tracking abrupt motion, it is observed that there is a clear trend towards increasing complexity of the Bayesian filtering framework. Proposed methods have become more complicated to cope with more difficult tracking scenarios. Often these sophisticated methods compensate the increased in complexity by trading-off performance in some other areas. For example, the increased number of subregions for sampling to cope with the variation of abrupt motion is compensated by using a smaller number of samples to reduce, if not maintaining, the computational cost incurred. Therefore, this chapter is motivated by the meta-level question on *whether these complex and sophisticated methods are really necessary?*.

3.1.1 Bayesian Tracking Framework

Visual tracking is often formulated as a graphical model and involves a searching process for inferring the motion of an object known as the state, x_t, from uncertain and ambiguous observations, z_t, at a given time, t. Generally, it consists of two essential steps: prediction and update. Given all available observations, $z_{1:t-1} = z_1, ...z_{t-1}$, from time $t = 0 : t-1$, the prediction stage applies the probabilistic transition model, $p(x_t \mid x_{t-1})$, to predict the posterior distribution, $p(x_t \mid z_{1:t-1})$, at time t as follows:

$$p(x_t \mid z_{1:t-1}) = \int p(x_t \mid x_{1:t-1}) p(x_{t-1} \mid z_{1:t-1}) dx_{t-1} \tag{3.1}$$

When the observation, z_t, is available at time t, the state is then updated using the Bayesian rule:

$$p(x_t \mid z_{1:t}) = \frac{p(z_t \mid x_t) p(x_t \mid z_{1:t-1})}{p(z_t \mid z_{1:t-1})} \tag{3.2}$$

where $p(z_t \mid x_t)$ is the observation model, also known as the likelihood. This process is also known as Bayesian filtering, optimal filtering or stochastic filtering.

The transitions from $p(x_{t-1} \mid z_{1:t-1})$ to $p(x_t \mid z_{1:t-1})$ is often analytically intractable, and in some cases, the posterior cannot be evaluated simply in closed form due to the complexity of its observation model, $p(z_t \mid x_t)$. Therefore, tracking algorithms resort to methods that are based on recursive approximations or iterative sampling techniques. In the PF approach as proposed in (Isard & Blake, 1998), the posterior, $p(x_t \mid z_{1:t})$ is approximated with a Dirac function, δ, using a finite set of N particles, $X_t = \{x_t^1, x_t^2, ...x_t^N\}$. To accomplish this, candidate particles are sampled from an appropriate importance distribution, $\pi(x_t \mid z_{1:t})$, that approximates the posterior, $x_t^n \sim \pi(x_t \mid x_{1:t-1}^n, z_t)$ and the approximation approaches the true posterior density as $N \to 1$. The weight of each candidate particle can be computed according to the following importance ratio:

$$w_t^n \;\alpha\; w_{t-1}^n \frac{p(z_t \mid x_t^n) p(x_t^n \mid x_{t-1}^n)}{\pi(x_t^n \mid x_{1:t-1}^n, z_t)} \tag{3.3}$$

Finally, the posterior filtered density, $p(x_t \mid z_{1:t})$ can be approximated by:

$$p(x_t \mid z_{1:t}) \approx \sum_{n=1}^{N} w_t^n \delta(x_t - x_t^n) \tag{3.4}$$

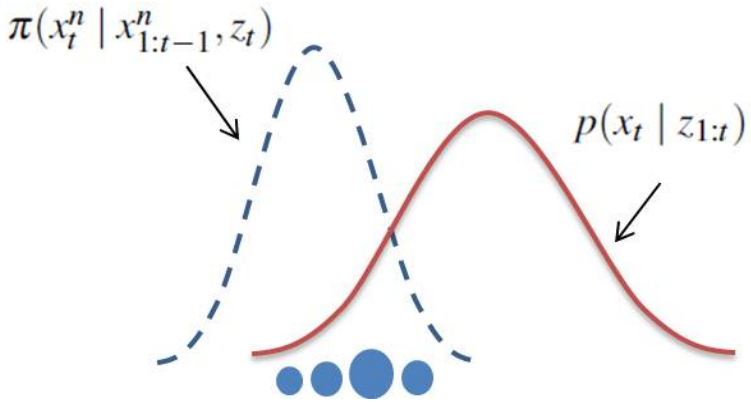
A summary of the standard PF algorithm is described in Algorithm 1.

Algorithm 1 : PF Algorithm

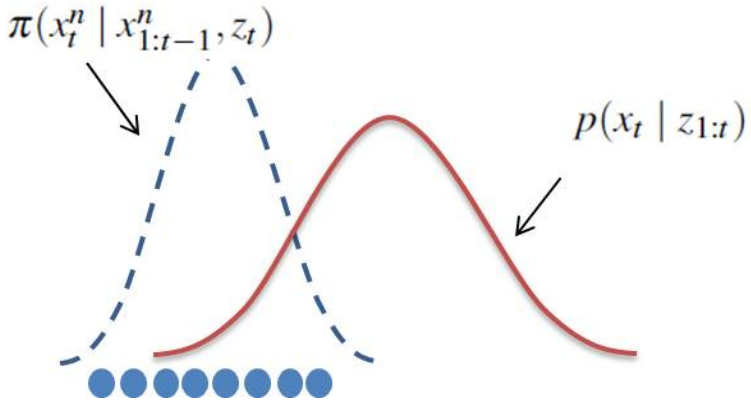
Initialisation:
$\{x_t^n = \emptyset\}$ $n = 1, ..., N$
$\eta = 0$
for n=1:N **do**
  Select the n^{th} sample, $x_{t-1}^n \in X_{t-1}$
  Sample x_{t-1}^n from $\pi(x_t \mid x_{1:t-1}^n, z_t)$
  Estimate the importance weight, $w_t^n \;\alpha\; w_{t-1}^n \frac{p(z_t \mid x_t^n) p(x_t^n \mid x_{t-1}^n)}{\pi(x_t^n \mid x_{1:t-1}^n, z_t)}$
  Update normalisation factor, $\eta = \eta + w_t^n$
  Insert $X_t = \{x_t^n, w_t^n\}$
end for
for n=1:N **do**
  Normalise weight, $w_t^n = w_t^n / \eta$
end for

3.1.2 Optimal Proposal Distribution

In general, the quality of the posterior distribution through iterative sampling methods is highly influenced by the quality of the proposal distribution, $\pi(x_t \mid z_{1:t})$, since it concerns sampling the particles in the relevant area where the posterior is significant. If the proposal distribution is similar to the posterior, the imbalance of weights amongst the particles can be reduced, and therefore, high accuracy and high computational efficiency can be achieved. Otherwise, only a few particles would have high likelihood, resulting in a large variance between the particles' weights and erroneous estimation. This situation is known as the particle degeneracy problem as shown in Fig. 3.2a. Also, if the proposal distribution is drawn from the tail of the actual posterior, a phenomenon known as trapped in local optima happens as shown in Fig. 3.2b.



(a) Particle degeneracy.



(b) Trapped in local optima.

Figure 3.2: Known problem of sampling-based tracking methods such as particle filter tracking and its variation.

Many techniques have been proposed to design efficient proposal distributions (Doucet & Johansen, 2011). In particular, the use of standard suboptimal filtering techniques such as the Extended Kalman Filter (EKF) (Ljung, 1979; Ruck, Rogers, Kabrisky, Maybeck, & Oxley, 1992; Welch & Bishop, 1995; PomÃ¡rico-Franquiz, Khan, & Shmaliy, 2014) and Unscented Kalman Filter (UKF) (Wan & Van Der Merwe, 2000; Rui & Chen, 2001; Van Der Merwe & Wan, 2001; G. Liu, Tang, Huang, Liu, & Sun, 2007; Holmes, Klein, & Murray, 2009) to obtain proposal distributions are very popular in the literature. Besides that, there are also methods that apply local optimisation techniques to design the proposal distribution, centred around the mode of the actual posterior, $p(x_t \mid z_{1:t})$ (M. K. Pitt & Shephard, 1999; M. J. Pitt & Shephard, 2001; F. Yan et al., 2005).

While ideally it is best to perform sampling on the posterior, in reality this is not possible; as the posterior is not known and is the one to be estimated. Thus, the posterior is estimated by sampling particles from known proposal distribution. In the simplest scenario, the dynamic transition model is usually taken as the proposal distribution for sampling, $\pi(x_t \mid z_{1:t}, z_t) = p(x_t \mid x_{1:t-1})$. The distribution is commonly governed by the Gaussian assumption, where $\pi(x_t \mid x_{1:t-1}, \mathcal{N})$ and $\mathcal{N}(\theta, \sigma^2)$. Often the mean, θ and variance σ^2 are fixed. In constant-velocity dynamic models, the transition model is governed by Gaussian distribution with the addition that the displacement of the state is relative to its previous time-step. However, the non-linear Gaussian transition model does not take into account current observation data, and therefore is still prone to have most of the particles located in the low likelihood area. Furthermore, in abrupt motion tracking, the posterior is often of a larger state space with rugged landscape, and cannot be modelled by Gaussian with fixed mean and variance or with uniform sampling. Theoretically, a larger variance would help to cover abrupt motion. However, the increase in proposal variance would incur additional computational cost.

Motivated by the meta-level question prompted in (Zhu et al., 2012) on *whether there is a need to have more training data or better models for object detection*, this chapter raises similar question in the domain of visual tracking; *will continued progress in visual tracking be driven by the increased complexity of tracking algorithms?* As indicated in the earlier section, often these sophisticated methods compensate the increased in complexity in a certain aspect of the algorithm by reducing another aspect of it. Furthermore, according to (Cifuentes et al., 2012), different scenarios require different dynamic models. If a particular *motion models only work sometimes*, on a particular scenario, then *how far should the increased in complexity of tracking algorithms be*, to deal with the challenges of real-time scenarios?

In order to address the common issues of sampling-based tracking methods in dealing with abrupt motion while prioritising the need for accurate, yet efficient tracking algorithm, this chapter proposes an optimal approximation of the proposal distribution, known as the SwATrack algorithm. In the proposed framework, abrupt motion tracking is deemed as an optimisation problem, where the approximation of the distribution incorporates the latest observation and sharing of information between particles. A stochastic optimisation approach known as the PSO algorithm, which exploits the emergent behaviour of a swarm of particles is adapted to guide the proposal distribution towards the actual posterior. The optimal distribution is described as, $q(x_t \mid x^n_{1:t-1}, Q)$, where $Q(\hat{\theta}, \hat{\sigma}^2)$ is the output of the proposed SwATrack algorithm. There are several advantages of the proposed solution. First, the proposed method alleviates the Gaussian assumption for sampling by utilising the global solution (emergence behaviour) found by the swarm of particles in the SwATrack to automatically adjust the variance of the proposal distribution. This allows a more flexible and efficient proposal distribution that deals with both, smooth and abrupt motion. Second, the communication and synergy between particles provide a way to avoid the local optima and reach the global maximum. This is made possible

since each particle in the SwATrack swarm has its own velocity and they communicate with each other to direct the search into optimal regions, a notion that is not present in other methods such as the PF. Third, the convergence of the swarm takes into account the observations, thus allowing a faster convergence with limited number of samples.

3.2 Particle Swarm Optimisation Revisit

PSO - a population-based stochastic optimisation technique was developed by Kennedy and Eberhart in 1995 (Eberhart & Kennedy, 1995). It was inspired by the social behaviour of a flock of birds. Briefly, let us assume a n-dimensional search space, $S \in \mathscr{R}^n$ and a swarm comprising of N particles. Each particle represents a candidate solution to the search problem and is associated with a fitness function (cost function), $f : S \rightarrow \mathscr{R}$. At every kth iteration, each particle is represented as $\{x_k^n\}_{n=1,\ldots,N}$, where $k = 1,2,\ldots K$. Each particle, x_k^n has its own position, $p(x_k^n)$, velocity, $v(x_k^n)$, and a corresponding fitness value (cost), $f(x_k^n)$. Each particle will remember its personal best solution encountered throughout kth iterations, $pBest_k^n = x_l^n$, where $x_l^n = \arg\max(f(x_k^n))$. l represents the kth iteration, which has the best solution for the kth particle. The position of the personal best is denoted as $p(pBest_k^n)$, while its corresponding fitness value is $f(pBest_k^n)$. Additionally, for every kth iteration, the particle with the best fitness value will be chosen as the global best and is denoted as the index of the particle, g, at kth iteration. $gBest_k = x_k^g$, where $x_k^g = \arg\max(f(x_k^n))$. The position of the personal best is denoted as $p(gBest_k)$, while its corresponding fitness value is $f(gBest_k)$.

The summary of the conventional PSO algorithm is shown in Algorithm 2. In Eq. 3.5, the parameters ω, c_1 and c_2 are positive acceleration constants used to scale the influence of the inertia, cognitive and social components respectively; r_1, $r_2 \in [0,1]$ are uniformly distributed random numbers to randomise the search exploration.

Algorithm 2 : Standard PSO

Initialisation, at iteration k=0

- Initialise a population of N particles, $\{x_k^n\}_{n=1,\ldots,N}$ with positions, $p(x_k^n)$, at random within the search space, S.
- Initialise the velocities, $v(x_k^n)$ at random within $[1,-1]$.
- Evaluate the fitness value of each particle, $f(p(x_k^n))$ and identify their personal best, $pBest_k^n$. Update $p(pBest_k^n)$ and $f(pBest_k^n)$.
- Identify the global best gth particle. Update $gBest_k$, $p(gBest_k)$ and $f(gBest_k)$.

for $k = 1$ to K **do**
  for $n = 1$ to N **do**
    Compute the new velocity according to:

$$v(x_{k+1}^n) = [(\omega\, v(x_k^n)) + (c_1\, r_1\, (p(pBest_k^n) - p(x_k^n))) + (c_2\, r_2\, (p(gBest_k) - p(x_k^n))] \tag{3.5}$$

    Update the position according to:

$$p(x_{k+1}^n) = p(x_k^n) + v(x_{k+1}^n) \tag{3.6}$$

    Check for out of bound:

$$p(x_{k+1}^n) \in S \tag{3.7}$$

    Update personal best variables; $pBest_k^n$, $p(pBest_k^n)$, and $f(pBest_k^n)$.
    Update global best variables; g, $gBest_k$, $p(gBest_k)$, and $f(gBest_k)$.
    Check for **Convergence**

    if Convergence == TRUE **then**
      Terminate iteration
    else
      Continue
    end if

  end for
end for

3.2.1 Limitations of the conventional PSO in Tracking Abrupt Motion

The conventional PSO is not able to track abrupt motion effectively, due to several reasons as follows:

Constant Acceleration Parameters: The parameter c_1 controls the influence of the cognitive component, $(c_1 * r_1 * (p(pBest_k^n) - p(x_k^n)))$, that is, it represents the individual memory of particles (personal best solution). A higher value of this parameter indicates a bias towards the cognitive component and vice versa. On the other hand, the parameter c_2 controls the influence of the social component, $(c_2 * r_2 * (p(gBest_k) - p(x_k^n))$, that is, it indicates the joint effort of all particles to optimise a particular fitness function, f.

The main drawback of the conventional PSO is the lack of a reasonable mechanism to effectively handle the acceleration parameters (ω, c and r); which are often set to constant variables (Clerc & Kennedy, 2002; Epitropakis, Plagianakos, & Vrahatis, 2012; H. Wang, Sun, Li, & Rahnamayan, 2013). For example, many applications of the PSO and its variant set these values to, $c_1 = c_2 = 2.0$, which gives the stochastic factor a mean of 1.0 and giving equal importance to both the cognitive and social components (Eberhart & Kennedy, 1995). This limits the search space and therefore cannot deal with abrupt motion, where the search for distribution of the proposal is not known. Therefore, it is essential to have dynamic acceleration parameters that are able to cope better with the unexpected dynamics of abrupt motion.

Trade-off between Exploration and Exploitation: The inertia weight, ω plays an important role, directing the exploratory behaviour of the swarms. A high value of inertia accentuates the influence of the previous velocity information and forces the swarm to explore a wider search space; while a decreasing inertia weight re-

duces the influence of the previous velocity and exploit a smaller search space. Often, the inertia value that controls the influence of the previous velocity is set to $\omega \in [0.8, 1.2]$ (Y. Shi & Eberhart, 1998). Recently, decaying inertia weight, $\omega = 0.9 \rightarrow 0.4$ have been proposed and tested, with the aim of favouring global search at the start of the algorithm and local search towards the end (Poli, 2008; Rana, Jasola, & Kumar, 2011). While these settings have been shown to work well in other optimisation problems (Poli, 2008; Rana et al., 2011), one must note that it is not applicable to tracking abrupt motion as the dynamic change is often unknown. Therefore, a solution that is able to handle the trade-off between the exploration and exploitation is crucial.

3.3 Proposed Tracking Framework: SwATrack

In this section, the SwATrack tracking framework to track target with abrupt motion is proposed. The SwATrack is a variant of the conventional PSO. Particularly, this section discusses in detail on how the effective combination of Dynamic Acceleration Parameters (DAP) and Exploration Factor $\mathscr{E}$ in the proposed SwATrack framework allows effective tracking of abrupt motion.

3.3.1 Dynamic Acceleration Parameters (DAP)

Since PSO is an iterative solution, efficient convergence is an important issue towards a real-time abrupt motion estimation system. However, the strict threshold of the conventional PSO velocity computation as in Eq. 3.5 will always lead to particles converging to a common state estimate (the global solution). One reason is that the velocity update equation uses a decreasing inertia value which indirectly forces the exploration of particles to decrease over the iterations. On the other hand, an increasing inertia value will lead to swarm explosion in some scenarios.

To overcome this, a mechanism to self-tune the acceleration parameters that utilises

the average velocity information of all particles in the swarm is introduced; the DAP mechanism. Firstly, the acceleration parameters are normalised so that they can be compared fairly with respect to the estimated velocity, $normalise(w, c_1, c_2) = 1.0$. The DAP mechanism takes into account the observation data, by incorporating the quality of estimation (likelihood) to refine the acceleration parameters dynamically, $f(C)$. The basic notion is that when an object moves consistently in a particular direction, $C = 1$, the inertia, w and cognitive weight, c_1 values are increased to allow resistance to any changes in its state of motion in the subsequent frames. Otherwise when $C = 0$, the social weight c_2 is increased by a step size, m to reduce its resistance to motion changes as Eq. 3.8. The increase of the social weight allows global influence and exploration of the search space, which is relevant when the motion of a target is dynamic. The exploitation within nearby regions is equitable when an object is moving with consistent motion.

The parameter C is estimated by computing the frequency of change, $f(C)$, in the quantised motion direction of the object; $C = 1$ represents consistent motion with minimal change of direction, while $C = 0$ represents inconsistent or dynamic motion. In this study, the motion velocity is categorised into 8 quantised directions, within an interval of 5 frames to determine its consistency.

$$f(C) = \begin{cases} c_1 = c_1 + m; \;\; c_2 = c_2 - m; \;\; \omega = \omega + m; \quad C = 1 \\ c_1 = c_1 - m; \;\; c_2 = c_2 + m; \;\; \omega = \omega - m; \quad C = 0 \\ {}^{*} \text{ subject to } \quad normalise(\omega, c_1, c_2) = 1.0 \end{cases} \tag{3.8}$$

3.3.2 Exploration Factor ($\mathscr{E}$)

The normalisation of DAP to 1.0 will restrict the overall exploration of the state to a certain degree. Hence, the exploration factor, $\mathscr{E}$ which serves as a multiplying factor to increase or decrease the exploration is introduced to alleviate the aforementioned

limitation. The exploration factor, $\mathscr{E}$ is defined as the parameters that adaptively:

1. increase the *exploration* with high variance, or

2. increase the *exploitation* with low variance.

By utilising these exploitation and exploration capabilities, the SwATrack framework is able to escape from being trapped in a common state (local optima). Thus, allowing the SwATrack to deal with smooth and abrupt motion more effectively. At every kth iteration, the quality of the estimated position upon convergence (global best) is evaluated using its fitness value. $f(gBest_k)_{k\to K} = 1$ indicates high likelihood whereas $f(gBest_k)_{k\to K} = 0$ indicates low likelihood or no similarity between an estimation and target.

When $f(gBest_k) \leq T_{MinF}$, where T_{MinF} is a threshold, it clearly indicates that there is low resemblance between the estimation (which is derived from the observation) and the target; most likely the proposal distribution may not be the actual distribution of the posterior. Thus in this scenario, $\mathscr{E}$ is increased with the maximum number of iterations, K by an empirically determined step size, l. This drives the swarm of particles to explore the region beyond the current local maxima (increase exploration). However, when an object has left the scene, K tend to increase continuously and may cause swarm explosion. Thus, to avoid searching beyond the search space, a boundary search is imposed, where $K \in S$.

$$\mathscr{E} \propto f(gBest_k) \tag{3.9}$$

In another scenario, where $f(gBest_k) \geq T_{MinF}$, $\mathscr{E}$ is decreased along K; constraining the search around the current local maximum (exploitation). In a straightforward manner, it is always best to drive particles at its maximum velocity to provide a reasonable bound in order to cope with the maximum motion change. However, this is not reasonable for real-time applications as it incurs unnecessary computational cost especially when

the motion is not abrupt. Thus, by introducing the adaptive scheme to automatically adjust the exploration and exploitation behaviour of the swarm, SwATrack is able to deal with smooth and abrupt motion, while keeping the computational cost at its minimal. Also, since the particles in SwATrack exchange information with one another, a minimal number of particles is sufficient for sampling. In summary, the $\mathscr{E}$ is proportional to the fitness function, or in this context, the quality of estimation. The threshold, T_{MinF} is dependent on the fitness function or cost function used and may vary from one application to another. In this study, the normalised Bhattacharyya distant measure is used as the fitness value to measure the quality of the estimation; where 1 represents the highest similarity between an estimation and target and 0 represents no similarity. Based on the sequences used for experiment, $T_{MinF} = 0.7$ is applied.

3.3.3 Novel Velocity Model

With the introduction of DAP and $\mathscr{E}$, the novel velocity model, $\grave{v}$ in the proposed SwATrack framework can be written as:

$$\grave{v}(x_{k+1}^n) = \mathscr{E}_k[(\omega\, \grave{v}(x_k^n)) + (c_1\, r_1\, (p(pBest_k^n) - p(x_k^n))) + (c_2\, r_2\, (p(gBest_k) - p(x_k^n))] \tag{3.10}$$

where, $\mathscr{E}_k$ is the exploration factor at iteration k, and c, r, ω are the acceleration parameters with the condition $normalise(\omega, c_1, c_2) = 1.0$. The normalised condition applied to the acceleration allows on the fly tuning of these parameters according to the quality of the fitness function. The fitness function used here is represented by the normalised distant measure between the appearance model of an estimation and the object-of-interest. The fitness value of a particle, $f(x_k^n)$ measures how well an estimation of the object's position matches the actual object-of-interest by taking into account the observation data; where

1.0 represents the highest similarity between an estimation and target and 0 represents no similarity. At every kth iteration, each particle varies its velocity according to Eq. 3.10 and move its position in the search space according to:

$$p(x_{k+1}^n) = p(x_k^n) + \hat{v}(x_{k+1}^n) \tag{3.11}$$

Note that the motion of each particle is directed towards the promising region found by the global best, $gBest_k$ from the previous iteration, $k = k-1$. The summary of proposed SwATrack algorithm is shown in Algorithm 3.

3.4 Experimental Results and Discussion

In this section, the feasibility and robustness of the proposed SwATrack in handling abrupt motion tracking is evaluated using a machine with a configuration of Intel core-i7, 1GHz with 8GB Random Access Memory (RAM). The proposed SwATrack was implemented in C++ and OpenCV library.

3.4.1 Experiment Setup

The object-of-interest is assumed to be priori and hence, the 2D position of the target in the first frame is initialised manually using the prior information. Automatic initialisation of target is a challenging research topic by itself, and thus is not in the scope of this study. The object is represented by its appearance model, which comprises HSV histogram with uniform binning; 32 bins. The normalised Bhattacharyya distant measure is used as the fitness value (cost function) to measure the quality of the estimation; where 1 represents the highest similarity between an estimation and target and 0 represents no similarity. Here, the initial values for SwATrack are $\mathscr{E} = 25$, $\omega = 0.4$, $c_1 = 0.3$, $c_2 = 0.3$, $K = 30$, $N = 15$, $m = l = 5$, respectively. The initial values are set based on the set of test sequences used in the experiments and the optimal settings may vary from one sequence

Algorithm 3 : Proposed SwATrack

Initialisation, at iteration k=0

- Initialise a population of N particles, $\{x_k^n\}_{n=1,\ldots,N}$ with positions, $p(x_k^n)$,at random within the search space, S.
- Initialise the velocities, $v(x_k^n)$ at random within $[1,-1]$.
- Evaluate the fitness value of each particle, $f(p(x_k^n))$ and identify their personal best, $pBest_k^n$. Update $p(pBest_k^n)$ and $f(pBest_k^n)$.
- Identify the global best gth particle. Update $gBest_k$, $p(gBest_k)$ and $f(gBest_k)$.

for $k = 1$ to K **do**
for $n = 1$ to N **do**
Compute the new velocity according to:

$$\grave{v}(x_{k+1}^n) = \mathscr{E}_k[(\omega\,\grave{v}(x_k^n)) + (c_1\,r_1\,(p(pBest_k^n) - p(x_k^n))) + (c_2\,r_2\,(p(gBest_k) - p(x_k^n))]$$

where,

$$normalise(\omega, c_1, c_2) = 1$$

if $f(gBest_k) \leq T_{MinF}$ **then**
$\mathscr{E} = \mathscr{E} + l$, $K = K + l$
else
$\mathscr{E} = \mathscr{E} - l$, $K = K - l$
end if

if $C = 1$ **then**
$c_1 = c_1 + m$, $c_2 = c_2 - m$, $\omega = \omega + m$
else
$c_1 = c_1 - m$, $c_2 = c_2 + m$, $\omega = \omega - m$
end if

Update the position according to:

$$p(x_{k+1}^n) = p(x_k^n) + \grave{v}(x_{k+1}^n)$$

Check for out of bound:

$$p(x_{k+1}^n) \in S \tag{3.12}$$

Update personal best variables; $pBest_k^n$, $p(pBest_k^n)$, and $f(pBest_k^n)$.
Update global best variables; g, $gBest_k$, $p(gBest_k)$, and $f(gBest_k)$.
Check for **Convergence**

if Convergence == TRUE **then**
Terminate iteration
else
Continue
end if

end for
end for

to another.

A comparison between the proposed SwATrack and a variety of state-of-the-art tracking solutions is performed. The benchmarked solutions include the conventional PSO tracking, PF (F. Yan et al., 2005; Maggio & Cavallaro, 2009), Ball Detection Method (BDM) (Wong & Dooley, 2011), Fragment-based Tracking (FragTrack) (Adam et al., 2006), Annealed Wang-Landau Monte Carlo (A-WLMC) (Kwon & Lee, 2008) and Compressive Tracking (CT) (X. Zhang et al., 2008). A thorough evaluation which consists of both, the detection accuracy (%) and processing time (milliseconds per frame) is done. In all experiments, the parameters of the state-of-the-art algorithms (i.e. top performing trackers) are set to fine-tuned settings as proposed by the respective authors accordingly (F. Yan et al., 2005; Maggio & Cavallaro, 2009; Adam et al., 2006; Kwon & Lee, 2008; Wong & Dooley, 2011; X. Zhang et al., 2008). In this study, the conventional PSO and PF (F. Yan et al., 2005; Maggio & Cavallaro, 2009) were re-implemented using C++ and OpenCV library, while the rest are using the publicly available framework by their respective authors. Only the BDM tracker in(Wong & Dooley, 2011) was implemented in a Matlab environment, while the others are in C++ and OpenCV library.

3.4.2 Dataset

The proposed SwATrack was evaluated using a newly introduced abrupt motion dataset - namely the **M**alaya **A**brupt **Mo**tion Dataset (MaMO) dataset. This dataset comprises 12 videos as illustrated in Fig. 3.3 and Table 3.1, which are compiled from the various dataset used by the state-of-the-art tracking solutions. These sequences are arranged according to the different challenging scenarios as described in the following:

a) Rapid Motion of Small Object: There are 5 video sequences in this scenario to test the effectiveness of the proposed method in terms of tracking small object (e.g. table tennis ball) that exhibits fast motion. $TableT_1$ is the SIF Table Ten-

nis sequence - a widely used dataset in the area of computer vision, especially for evaluation of detection and tracking methods (Wong & Dooley, 2011). This sequence has complex, highly textured background and exhibit camera movement with some occlusion between the ball and the player's arm. $TableT_2$ is a sample training video from the International Table Tennis Federation (ITTF) video library which is created to expose players, coaches and umpires to issues related to service action. Although this sequence is positioned to provide the umpire's point of view of a service, it is very challenging as the size of the tennis ball is very small; about 8x8 pixels to 15x15 pixels for an image resolution of 352 x 240. The video comprises of 90 frames including 10 frames in which severe occlusion happens, where the ball is hidden by the player's arm. $TableT_3$ is a match obtained from a publicly available source. In this sequence, the tennis ball is relatively large as it features a close-up view of the player. However, there are several frames where the ball appears to be blurred due to the low frame rate and abrupt motion of the tennis ball. $TableT_4$ and $TableT_5$ were captured at a higher frame rate, thus the spatial displacement of the ball from one frame to another appears to be smaller (less abrupt) and the ball is clearer. This is to test the ability of proposed method to handle normal visual tracking scenario. Since the ground truth for these data were not provided, the ground truth of the object-of-interest, in each sequence is manually labelled for evaluation. The ground truth is described as bounding box information, $X(x,y,w,h) =$ (positions in the x-dimension, y-dimension, width, height).

b) Switching Camera: In general, the sampling-based methods often assume a large variance in the proposal density to deal with abrupt motion. However, a large variance tends to decrease the tracking accuracy when tracking smooth motion. Thus, the scenario of tracking using sequences obtained from switching between

multiple cameras is included to evaluate the tracking methods in dealing with both, the abrupt and smooth motion. This category comprises 3 videos which includes the the *Youngki*, *Boxing* and $Malaya_1$ sequences. The *Youngki* and *Boxing* can be found at (Kwon & Lee, 2013), where they consist of frames edited from changes of camera shots between multiple cameras, where the hand-over between cameras are aimed at tracking a particular object throughout the scene. Due to the object's handover between multiple cameras, the object appears to have drastic change in position between adjacent frames during the switching period as well as the scale. Otherwise, the object exhibits smooth motion as it is captured by a single camera. The $Malaya_1$ sequence is created by combining the frames in the *Boxing* and *Youngki* sequences in an alternative manner. This combination is done to introduce definite tracking error when tracking the boxer in the *Boxing* sequence; since the boxer is missing in the *Youngki* sequence. The simulation of inaccurate tracking scenario is to test the robustness of tracking methods in not only tracking abrupt motion, but recovery from inaccurate tracking.

c) Partially low frame rate: In another example of abrupt motion, a scenario of tracking in partially low frame rate is simulated. The *Tennis* video (Kwon & Lee, 2013) comprises of down-sampled data to mimic abrupt change caused by low frame rate. The frames are down sampled from a video with more than 700 original frames, by keeping one frame in every 25 frames. The rapid motion of the tennis player from one frame to another due to the down-sampling made tracking extremely difficult. Down-sampling is done to simulate abrupt motion during low-frame rates.

d) Inconsistent Speed: 2 video sequences were obtained from the YouTube, where each sequence comprises an object which moves with inconsistent speed through-

out the sequence. The first video, $Malaya_2$, aims to track a synthetic ball which moves randomly across the sequence with inconsistent speed, whilst the second video, $Malaya_3$ tracks a soccer ball which is being juggled in a free-style manner in a moving scene with a highly textured background (grass).

e) Multiple targets: This is to demonstrate the capability of the proposed system to track multiple targets; whilst most of the existing solutions are focused on single target. A synthetic video, $Malaya_4$ that consists of two simulated balls moving at random speed is created for this purpose.

In general, most of the video sequences in *MAMo* dataset are well diversified as most of them contain a mixture of both the smooth and abrupt motion. It is less likely for an object-of-interest to move with abrupt motion at all time, unless the video is captured at low frame rate as exhibit by the *Tennis* sequence, in particular. The *MAMo* dataset is publicly available along with their corresponding ground truth information[1].

Table 3.1: Summary of the Malaya Abrupt Motion *(MAMo)* dataset.

	Category	Name of the Video Sequences	Number of Videos
a)	Rapid Motion of Small Object	$TableT_{1-5}$	5
b)	Switching Camera	*Youngki*, *Boxing* and $Malaya_1$	3
c)	Partially low-frame rate	*Tennis*	1
d)	Inconsistent Speed	$Malaya_{2-3}$	2
e)	Multiple targets	$Malaya_4$	1
	TOTAL		**1**2

3.4.3 Quantitative Result

3.4.3 (a) Experiment 1: Detection Rate

Detection rate refers to the correct number and placement of the objects in the scene. For this purpose, the ground truth of *j*th object is denoted as GT_j, and the output from the tracking algorithms of *j*th object is denoted as, ξ_j. The ground truth and tracker output of

[1]http://web.fsktm.um.edu.my/ cschan/project3.htm

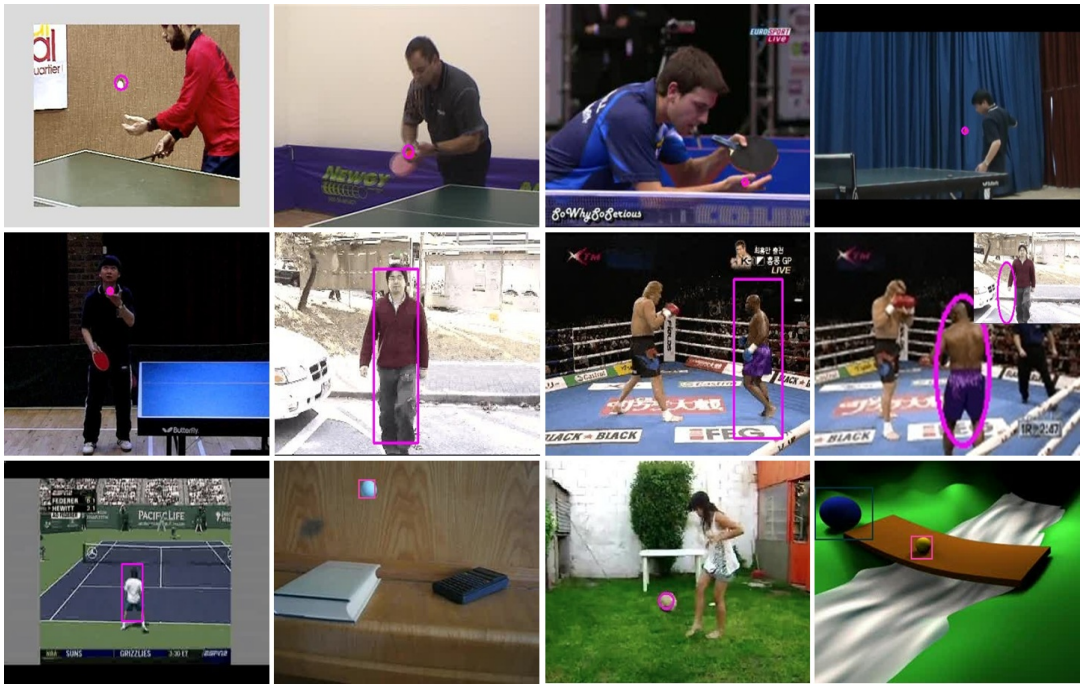


Figure 3.3: Sample shots of the newly introduced Malaya Abrupt Motion (MAMo) dataset. This collection of data comprises 12 videos which exhibit the various challenging scenarios of abrupt motion. Top row, from left to right: TableT$_1$,TableT$_2$,TableT$_3$,TableT$_4$; Second row, from left to right: TableT$_5$, *Youngki*, *Boxing*, *Malaya*$_1$; Bottom row, from left to right: *Tennis*, *Malaya*$_2$, *Malaya*$_3$, *Malaya*$_4$.

each *n*th object as bounding box information is described as, $X^j(x^j,y^j,w^j,h^j$ = *x-position, y-position, width, height)*. The coverage metric determines if a GT is being tracked, or if an ξ is tracking accurately. In (K. Smith, Gatica-Perez, Odobez, & Ba, 2005), it is shown that the *F-measure*, F, suited this task as the measure is 1.0 when the estimate, ξ_j overlaps perfectly with the ground truth, GT_j. Two fundamental measures known as *precision* and *recall* are used to determine the *F-measure*.

Recall: Recall measures how much of the GT is covered by the ξ, and takes the value of 0 if there is no overlap and 1.0 if the estimated position fully overlap with the actual locality of the target. Given a ground truth, GT_j, and a tracking estimate, ξ_j, the *recall*, $\Re_j$ is expressed as:

$$\Re_j = \frac{|\xi_j \cap GT_j|}{|GT_j|} \tag{3.13}$$

Precision: Precision measures how much of the ξ covers the *GT* takes the value of 0 if there is no overlap and 1 if they are fully overlapped. The *precision*, $\wp_j$ is expressed as:

$$\wp_j = \frac{|\xi_j \cap GT_j|}{|\xi_j|} \tag{3.14}$$

F-measure: The *F-measure*, F_j is expressed as:

$$F_j = \frac{2\Re_j \wp_j}{\Re_j + \wp_j} \tag{3.15}$$

Coverage Test: In this experiment, the *F-measure* is employed according to the score measurement of the known PASCAL challenge (Everingham, Gool, Williams, Winn, & Zisserman, 2010). That is, if the F_j of jth object is larger than 0.5, the estimation is considered as correctly tracked in the frame. Table 3.2-3.3 demonstrate the detection accuracy of the benchmarked tracking algorithms for all 8 test sequences. Overall, the experimental results show that the average tracking accuracy of the proposed method surpasses most of the state-of-the art tracking methods with an average detection accuracy of 91.39%. For all 6 test sequences ($TableT_1$, $TableT_2$, $TableT_5$, *Youngki* and *Tennis*), the SwATrack generates the best tracking results amongst the rest and ranked second best for sequence $TableT_4$ and *Boxing*, respectively.

Methods that are not built based on sophisticated motion model such as the FragTrack (Adam et al., 2006) performs poorly, overall with an average accuracy of 37.19%. Their method, which employs a refined appearance model that adapts to the changes of the object, copes well with partial occlusion. However, it is still dependent on the search radius and thus fails when tracking abrupt motion, where the

object tends to be outside the search window. PF on the other hand, achieves a detection accuracy of 85.6%. This is expected, since the PF algorithm is constrained to a fixed Gaussian motion model. Once PF has lost track of the object, it has the tendency to continue searching for the object in the wrong region; leading to error propagation and inability to recover from incorrect tracking such as shown in Fig. 3.4. Fig. 3.4a demonstrates sample shots of an abrupt motion scenario, where the PF tracker exhibits the state of being trapped in local optima. At frame 449-451, the PF tracker continues to locate the object within the assumed Gaussian distribution when the object has in actual fact, moved abruptly to the other corner of the image. On the other hand, the proposed SwATrack copes better with abrupt motion and does not get trapped in local optima; since the exploitation and exploration is self-adjusted based on the fitness function, and is shown in Fig. 3.4b. Thus, as shown in Fig. 3.4b, the SwATrack is able to track the object accurately although the motion is highly abrupt. Similarly, the inability of MCMC and its variants, the A-WLMC (Kwon & Lee, 2008) and Intensely Adaptive Markov Chain Monte Carlo (IA-MCMC) (X. Zhou et al., 2012) tracking methods in handling abrupt motion is shown in Fig. 3.6.

Dataset Unbias: The problem of dataset bias was highlighted in (Torralba & Efros, 2011) where the paper argue that *'Is it valid to expect that when training on one dataset and testing on another, there is a big drop in performance?'*. Motivated by this, a similar scenario in the tracking domain is replicated, and it is observed that although the A-WLMC method (Kwon & Lee, 2008) performs well in $TableT_4$ and *Youngki* sequence, they do not produce consistent results when tested across the other datasets as shown in Table 3.2-3.3. For example, it can be seen that the accuracy of A-WLMC changes drastically from one tracking scenario to another. The

average detection accuracy for *TableT* video sequences is fairly low at, 14.28% while for the *Tennis*, *Boxing* and *Youngki* sequences, it performs remarkably well with an average accuracy of 93.33%. This provides an indication that the A-WLMC solution (Kwon & Lee, 2008) may suffer from dataset bias problem, as it seems to only work well in their proposed dataset, but performed poorly when it is employed on different video sequences. Perhaps this is due to the motion model employed by these tracking methods that only works well on certain scenarios, alluding to the notion in (Cifuentes et al., 2012) that *different motion requires different motion models*. This is indeed not the case for our proposed SwATrack, which is more flexible and non-bias in handling different scenarios or datasets. The overall detection rates for SwATrack are 85.02% and 97.76%, respectively. For all video sequences that exhibit different challenging conditions, e.g. rapid motion ($TableT_{1-5}$), camera switching (*Youngki* and *Boxing*), low-frame rate (*Tennis*), the SwATrack has shown its ability to cope with the various scenarios of abrupt motion.

Size Invariance: A further investigation on the dataset bias problem is performed and the results demonstrate that there is an influence of the object size to the detection rate. For instance, the A-WLMC algorithm (Kwon & Lee, 2008) performs poorly for sequences in which the resolution of the object-of-interest is relatively small, such as in the *TableT* video sequences and performs surprisingly well when the object is large such as in the *Youngki*, *Boxing* and *Tennis* sequences. This indicates the need to have better representation of the object for a more accurate acceptance and rejection of estimations in the MCMC algorithm.

3.4.3 (b) Experiment 2: Computational Cost

Fig. 3.5 demonstrates the comparison results between the proposed method and the state-of-the-art tracking solutions in terms of time complexity. It is observed that the

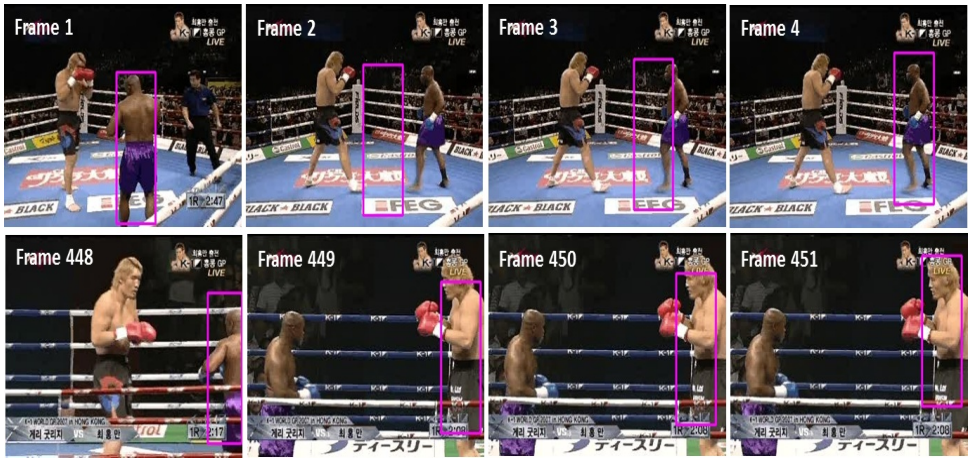
Table 3.2: Experiment results - Comparison of the Detection Rate (in %)

	PSO	PF	BDM	FragTrack	A-WLMC	CT	SwATrack
TableT1	70.1	58.4	68.3	64.9	47.2	72.3	**87.8**
TableT2	83.1	69.8	53.4	24.1	3.2	4.3	**93.1**
TableT3	58.2	52.1	67.3	55.3	8.7	24.5	**74.1**
TableT4	59.6	47.3	73.2	57.2	6.9	**98.2**	97.3
TableT5	60.3	34.5	64.2	9.7	5.4	36.3	**72.8**
Average	66.26	52.42	65.28	42.24	14.28	47.12	**85.02**

Table 3.3: Experiment results - Comparison of the Detection Rate (in %)

	PSO	PF	FragTrack	A-WLMC	SwATrack
Tennis	87.3	67.3	20.6	95.1	98.3
Youngki	87.1	47.2	27.5	86.8	98.7
Boxing	82.4	16.3	48.3	98.1	96.3
Average	85.6	43.6	32.13	93.33	**97.76**

SwATrack algorithm requires the least processing time with an average of 63 milliseconds per frame. On the contrary, the MCMC-based solutions which include the A-WLMC (Kwon & Lee, 2008) and PF (F. Yan et al., 2005; Maggio & Cavallaro, 2009) require higher processing time. This is likely due to the inherent correlation between the MCMC samplers which is known to suffer from slow convergence when an object has not been tracked accurately. In the experiments, it is observed that in scenarios where the MCMC requires high processing time, the accuracy of the MCMC is minimal. The increase in computational cost is due to the increase of search space when the observation model is unlikely representing the object. Note that the optimal number of samples deployed in the PF and MCMC throughout the sequences has been selected empirically; where it ranges from 150 to 1000 particles in PF, 600 to 1000 particles in MCMC with 600 iterations while the SwATrack uses 10-50 particles (15$\times$ in reduction) with 5-70 iterations. Intuitively, an increase in the number of samples would lead to an increase in computational cost as each particle would need to be evaluated against the appearance observation; explaining the minimal processing time required by the proposed SwATrack.



(a) Sample detections from PF tracking.

(b) Sample detections from SwATrack tracking.

Figure 3.4: Sample output to demonstrate the incorrect tracking state, which is caused by trapped in local optima. The aim of this sequence is to track the person in dark skin and purple short. From Frame 449-451 (a), PF lost track of the object due to sampling from incorrect distribution during abrupt motion. Thus, it can be observed that PF continues to track the object inaccurately once it has lost track of the object. On the other hand, the results in (b) demonstrate the capability of the SwATrack tracker in dealing with the non-linear and non-Gaussian motion of the object (Best view in colour).

As shown in Table 3.2-3.3, in which the SwATrack detection rate is ranked second. It is observed that although the CT (X. Zhang et al., 2008) and A-WLMC (Kwon & Lee, 2008) achieved better accuracy, their average processing time are threefold as compared to the SwATrack. This is due to the need to increase the subregions for sampling when the state space increases in the A-WLMC algorithm (Kwon & Lee, 2008). On the contrary, the SwATrack adaptively increases and decreases its proposal variance for a more effective use of the samples. Thus the processing time required is much lower as compared to the other methods. The advantage of the dynamic mechanism is reflected when

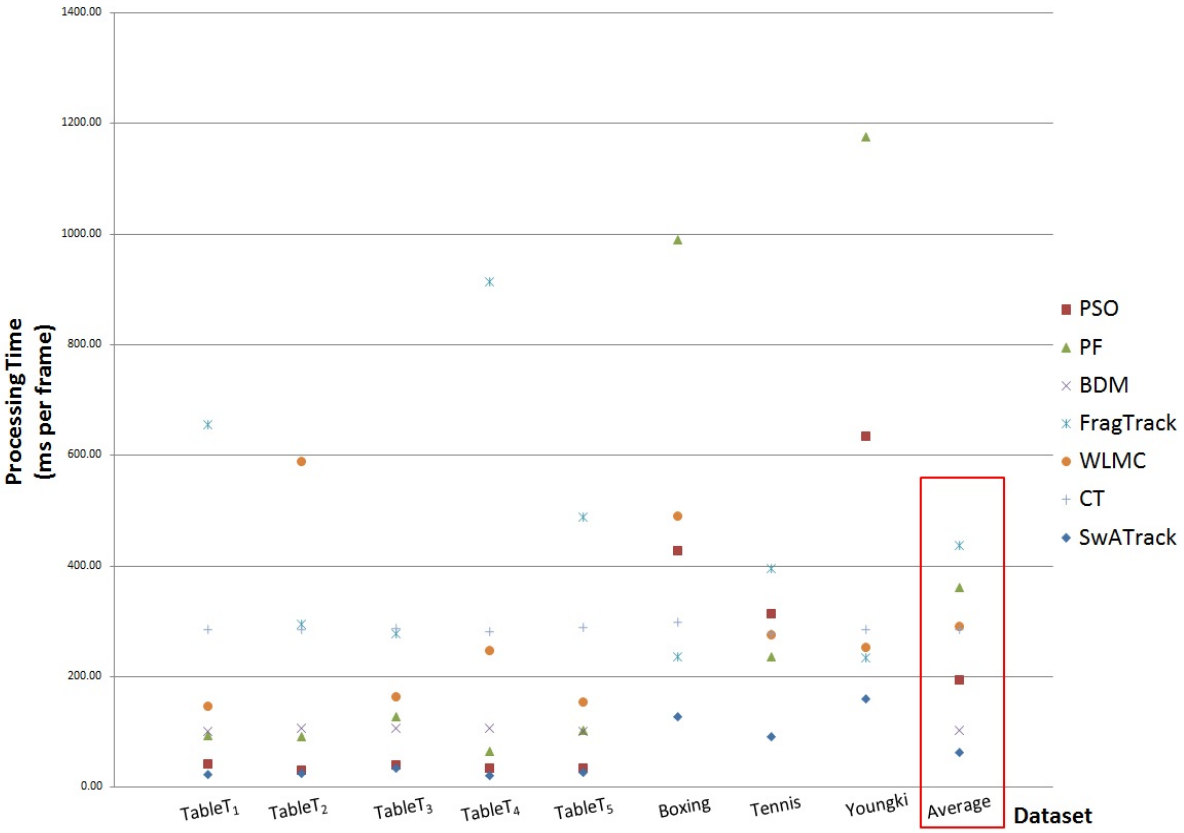


Figure 3.5: Time Complexity. This figure illustrates the comparison in terms of processing time (milliseconds per frame) between the proposed SwATrack, conventional PSO, PF, BDM, FragTrack, A-WLMC (Kwon & Lee, 2008) and CT.

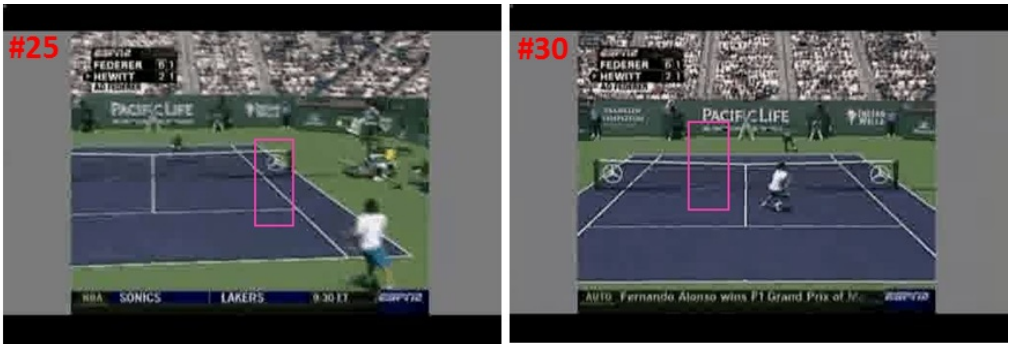
comparing the processing time of the SwATrack to conventional PSO (average of 195.20 milliseconds per frame); where the processing time of the PSO is three times greater than that of the SwATrack. In summary, the experimental results have demonstrated the capability of the proposed system to cope with the variety of scenarios which exhibit highly abrupt motion. The adaptation of a stochastic optimisation method into tracking abrupt motion has been observed to incur a slight increase in the processing cost, yet at the same time is able to have fair tracking accuracy as compared to the more sophisticated methods. Thus, the preliminary results at this stage, gives a promising indication that sophisticated tracking methods may not be necessary after all.

3.4.4 Qualitative Result

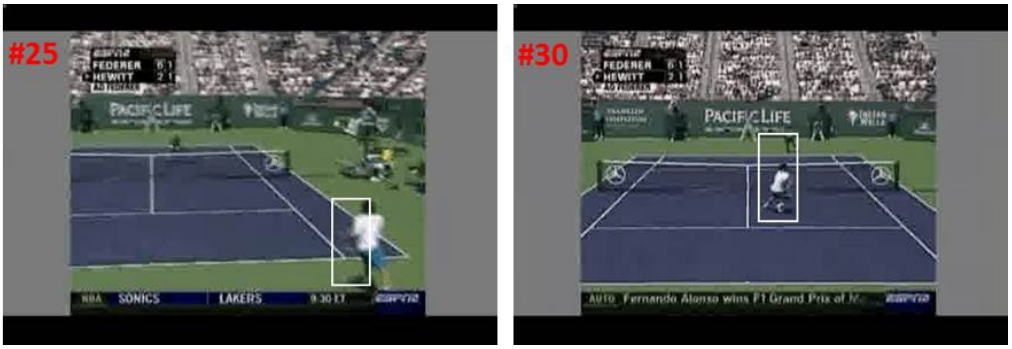
Partially low Frame Rate: This sequence aims to track a tennis player in a low-frame rate video, which has been down-sampled from a 700 frames sequence by keeping

one frame in every 20 frames. Here, the object (player) exhibits frequent abrupt changes which violate the smooth motion and constant velocity assumptions. Thus, motion that is governed by Gaussian distribution based on the Brownian or constant-velocity motion models will not work in this case. Fig. 3.6 shows sample shots to compare the performance between the PF tracking (500 samples), A-WLMC (600 samples) (Kwon & Lee, 2013), IA-MCMC (300 samples) (X. Zhou et al., 2012) and SwATrack (50 samples). It is observed that the tracking accuracy of the SwATrack is better than the PF and A-WLMC (Kwon & Lee, 2013) even by using fewer samples. While the performance of the SwATrack is comparable to the IA-MCMC (X. Zhou et al., 2012), the SwATrack requires fewer samples and thus requires less processing requirement. These results further validated that the proposed SwATrack is able to track moving object accurately and effectively, regardless of the varieties of change in the object's motion.

Local Minimum Problem: This experiment aims to test the capability of SwATrack to recover from incorrect tracking. This experiment would in particular, evaluate the efficacy of the proposed DAP and $\mathscr{E}$ in handling abrupt motion. Fig. 3.7 shows the result for *Youngki* and *Boxing* sequences, which exhibits abrupt motion in a camera switching scenario. Due to the object's handover between multiple cameras, the object appears to have a drastic change in position between adjacent frames during the switching period as well as the scale. Otherwise, the object exhibits smooth motion as it is captured by a single camera. The switching happens repeatedly when an object moves out of a particular camera view, into the field of view of another camera. The assortment of both smooth and abrupt motion would test the capability of the proposed DAP in increasing its exploitation during smooth motion and increasing its exploration during abrupt motion. In Fig. 3.7a, it is shown that during the switch from frame 247 and 248, the SwATrack appears to have inaccurate tracking of the object; as the estimated position which is highlighted by the ellipse does not overlap accurately with the object. However, due to the flexibility of the

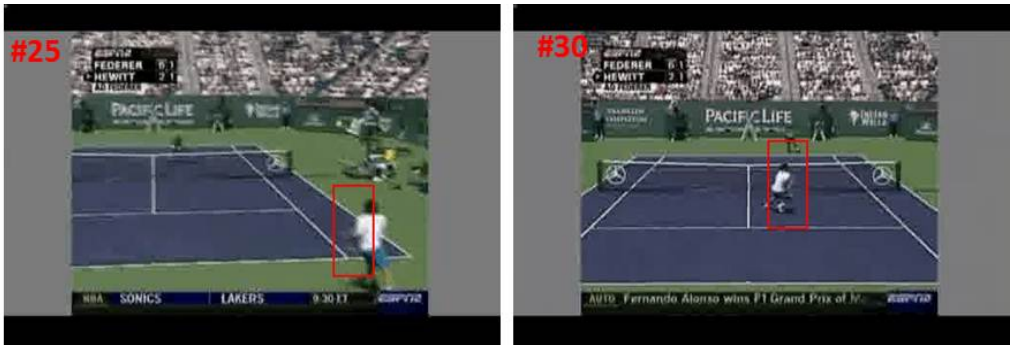


(a) Sample detections of PF.



(b) Sample detections of SwATrack.

(c) Sample detections of A-WLMC.



(d) Sample detections of IA-MCMC.

Figure 3.6: A comparison between PF, SwATrack, A-WLMC (Kwon & Lee, 2013) and IA-MCMC (X. Zhou et al., 2012). It is observed that the SwATrack tracking gives a more accurate fit of the object's locality.

proposed DAP and $\mathscr{E}$ which allows self-adjustment of the exploitation and exploration of the swarm based on the fitness function, the SwATrack algorithm is able to recover from the inaccurate tracking within minimal number of frames. Similar behaviour is observed in the *Youngki* sequence, as shown in Fig. 3.7a, which further validated our notion.

In another experiment on an even more challenging scenario, the $Malaya_1$ sequence is used for evaluation. In the $Malaya_1$ sequence, the frames in the *Boxing* sequence are combined in an alternative manner with the frames from the *Youngki* sequence; incorrect tracking is most likely to happen due to the frequent changes of the object's locality between two adjacent frames. In this combined sequence, the object-of-interest appears and disappears from one frame to another interchangeably due to the merge between two different video sequences, as shown in Fig. 3.8. From the qualitative results shown in Fig. 3.8b, it is shown that the A-WLMC tracking (Kwon & Lee, 2008) is not robust and does not cope well with inaccurate tracking. When the object-of-interest disappears from the scene (i.e. Frame 77), the A-WLMC gives an erroneous estimation of the object. In the subsequent frame, where the object re-appears, the A-WLMC has difficulty recovering from its tracking such as shown in Frame 78 where the estimation does not fit the actual position of the object accurately. In the subsequent frames, the A-WLMC tend to continuously missed tracked of the object. Although the sampling efficiency in the A-WLMC adopts a more efficient proposal distribution as compared to the standard PF, it is still subjected to a certain degree of trapped in local optima. Furthermore, the A-WLMC utilises the information of historical samples for intensive adaptation, thus requiring more frames information to recover from inaccurate tracking. The proposed SwATrack on the other hand, is observed to work well in this $Malaya_1$ video sequence, where minimal frame is required to recover from erroneous tracking. As shown in Fig. 3.8c, the SwATrack is able to track the object accurately when the object appears or re-appears in the scene (as shown in the even frame number). This is made possible due to the information exchange and

cooperation between particles in a swarm that provide a way to escape the local optima and reach the global maximum; leading to and optimised proposal distribution.

(a) Sample of SwATrack on *Boxing* sequence.

(b) Sample of SwATrack on *Youngki* sequence.

Figure 3.7: Sample outputs to demonstrate the flexibility of the proposed SwATrack to recover from incorrect tracking. It can be noticed that the SwATrack only requires minimal frames (1-2frames) to escape from local optima and achieve global maximum.

Swarm Explosion Problem: In the conventional PSO algorithm, the lack of a mechanism to control the acceleration parameters and the dependency on randomness in the system fosters the danger of swarm explosion and divergence. When swarm explosion or divergence happens, the velocities and positions of each particle are steered towards infinity and thus, preventing convergence. In the context of our study, swarm explosion and divergence are very likely. This is due to the tendency of the swarm to increase its exploration in order to deal with the abrupt change in an object locality. Thus, in this combination sequences (similar sequence as shown in Fig. 3.8) where the boxer disappears and reappears in the scene from one frame to another, it is shown that the conventional PSO fails to track the abrupt motion of the boxer accurately as shown in 3.9a. When the object disappears from the scene (since the boxer is missing in the *Youngki* sequence), the swarm tends to increase its exploration and is most likely to steer towards infinity; explosion happens. If this happens, the swarm lose track of the object and is most likely to continue searching from an inaccurate distribution leading to continuous incorrect track-

ing of the object. However, in the proposed SwATrack, recovery from incorrect tracking is made possible by the Dynamic Acceleration Parameters (DAP) and Exploration Factor $\mathscr{E}$ mechanisms, which prevent the particles from steering towards infinity by expanding and constricting the velocity of particles. See Fig. 3.9.

(a) Sample shots of the dataset that is obtained by combining frames from two different sequences. The object enclosed in the ellipse is the object to be tracked.

(b) Sample detections by the A-WLMC tracking. A-WLMC tend to tracked the object inaccurately once it has lost or missed tracked of the object as shown from Frame 79 onwards.

(c) Sample detections by the SwATrack tracking. In Frame 77, since the object-of-interest does not appear in the frame, inaccurate tracking happens. However, the SwATrack is able to recover its tracking at the following frame, Frame 78.

Figure 3.8: Sample outputs to demonstrate the capability to recover from incorrect tracking.

Invariant to Object Size: Further investigation to test the effectiveness of proposed SwATrack, PF (F. Yan et al., 2005; Maggio & Cavallaro, 2009) and A-WLMC (Kwon & Lee, 2013) is performed on resized sequences of similar set of datasets. This is to simulate the scenario in which the object size is smaller. Thus, the initial frame size of 360x240 is reduced into half, to 180x120 pixels. Observations on the results show that the SwATrack is the least sensitive towards the size of object-of-interest, while the detection accuracy of the A-WLMC decreases as the size of object gets smaller. This is due to the robustness of the optimised sampling in SwATrack as compared to the least robust method of rejection

(a) Sample detections by the conventional PSO tracking. PSO tracker tend to tracked the object inaccurately once it has lost or missed tracked of the object as shown from Frame 103 onwards.

(b) Sample detections by the SwATrack tracking. In Frame 103, since the object-of-interest does not appear in the frame, inaccurate tracking happens. However, the SwATrack is able to recover its tracking at the following frame, Frame 104.

Figure 3.9: Sample outputs to demonstrate the inaccurate tracking in conventional PSO due to swarm explosion, and the capability of the proposed SwATrack to track object accurately.

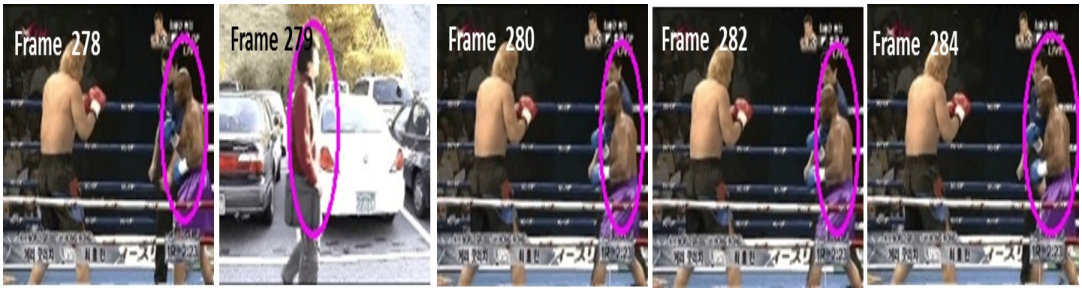
and acceptance as proposed in the A-WLMC. The overall detection accuracy of the proposed SwATrack remains at an average of 90% regardless of the object's size whereas the detection accuracy of PF and A-WLMC decrease significantly by more than 25% when the object's size decreases. Sample output is as shown in Fig. 3.10. Finally, an evaluation on videos obtained from the Youtube ($Malaya_{2-4}$) is performed. The qualitative results are as depicted in Fig. 3.11. It is observed that the SwATrack is able to track the abrupt motion of the balls efficiently, as well as the capability of the proposed system to track multiple objects; two simulated balls move at random. From the best of our knowledge, most of the existing solutions (Kwon & Lee, 2013; X. Zhou et al., 2012) are focused on single object.

3.5 Discussion

3.5.1 Can an increase in the complexity of tracking algorithms enhance the results of tracking abrupt motion?

Motivated by the meta-level question prompted in (Zhu et al., 2012) on *whether there is a need to have more training data or better models for object detection*, this chapter

(a) Sample detections from A-WLMC on reduced image size. A-WLMC has a high tendency to lose track of the object when it moves abruptly, and demonstrate continuous inaccurate tracking such as shown in Frame 279-284. Note that for similar frames, the A-WLMC tracker is able to track the object accurately when the image size is larger. Number of iterations = 600, particles = 600.



(b) Sample detections from SwATrack on reduced image size. SwATrack produces consistent tracking as compared to PF and A-WLMC, regardless of the size of object. Number of iterations = 30, particles =20.

Figure 3.10: Qualitative Results: Comparison between the A-WLMC and our proposed SwATrack in terms of reduced object size.

Figure 3.11: Sample of SwATrack on tracking the object(s) in $Malaya_{2-4}$ video sequences.

raises similar question in the domain of visual tracking. This section is motivated by the research question, on *whether the continued progress in visual tracking will be driven by the increased complexity of tracking algorithms?*

Intuitively, an increase in the number of samples in sampling-based tracking methods such as the PF and MCMC would increase the tracking accuracy. One may also argue that the additional computational cost incurred in the iterative nature of the proposed SwATrack and MCMC would complement the higher number of particles required by the PF. Thus, in order to investigate if these intuitions hold true, an experiment using an increasing number of samples and iterations *(Sampling-based vs. Iterative-based solutions)* is conducted. The behaviours of both the PF and SwATrack in terms of their tracking accuracy and processing time with the increase in complexity is then observed. PF is chosen in this testing as it bears close resemblance to the proposed SwATrack algorithm in which a swarm of particles are deployed for tracking. The experiments in this section are performed on the $TableT_{1-5}$ video sequences.

3.5.1 (a) Number of Samples vs Accuracy and Processing Time

Particle Filter: In the PF algorithm, a variety of values are set for the number of samples or particles (i.e. 50, 100, $\cdots$, 2000) used throughout the sequence to determine the statistical relationship between number of samples and performance. The performance is then measured by the detection accuracy (%) and the processing time (in milliseconds per frame). The average performance across all five $TableT$ video sequences is as shown in Fig. 3.12a. The detection accuracy and performance of the PF algorithm with different parameter settings are shown in Fig. 3.13-3.17(a).

The results had demonstrated that the amount of particles used in the PF is correlated to the detection accuracy, that is, the growth in the number of particles leads to an increase in the accuracy. Similarly, the average time taken also increases greatly as the number of

particles used in PF grows. This alludes the fact that as the number of particles increases, the estimation processes which include object representation, prediction and update also multiply. However, it is observed that the PF reaches a plateau when reaching the optimal accuracy. After which, any increase in the number of particles will either have a decrease in accuracy or no significant improvement. Fig. 3.12 shows that the detection accuracy decreases after the optimal solution, which is given when the number of particles is 600. This finding instigates the underlying assumption that the increase of number of particles will lead to an increase in the accuracy. Again, the question of whether complex (in this context the complexity is proportional to the number of particles deployed) tracking methods are really necessary is raised. Also, the best parameter configurations may differ from one sequence to another due to the different motion behaviour portrayed by the different object in each sequence, respectively. For example, in Fig. 3.13(a), the optimal setting is 250 particles which produces detection accuracy of 55% and takes 1.78 seconds of processing time. Meanwhile, the second sequence has a different optimal setting of 150 particles as shown in Fig. 3.14(a). This advocates the notion as in (Cifuentes et al., 2012) that *motion models indeed only work for sometimes*.

SwATrack: Similarly, a test using various parameter settings is conducted on the proposed SwATrack algorithm and the average results are demonstrated in Fig. 3.12b. Meanwhile, Fig. 3.13-3.17(b) illustrate the results for $TableT_1$ and $TableT_5$. In addition to the number of particles used in PF tracking, the proposed SwATrack has an additional influencing parameter, the maximum number of iterations. A variation on the number of particles against the number of iterations is set for fair evaluation. As illustrated in the left y-axis of the chart (bottom graph) of Fig. 3.12b, the average processing time increases as the number of iterations increase. This is similar to the behaviour of the PF. However, the processing time increases up to a maximum value; after which any increase in the iterations would not make much difference in its processing time. Notice that the processing

time for large number of iterations (55 & 70) overlaps with one another, demonstrating minimal increase in processing time as the number of iterations grows. This is due to the optimisation capability of the proposed SwATrack in terminating its search upon convergence, regardless of the defined number of iterations. This is particularly useful in ensuring efficient search for the optimal solution, with minimal number of particles. As for the detection accuracy, it is shown that in general, the average accuracy of the proposed SwATrack is higher than PF, with an average accuracy of 92.1% in the first sequence as shown in Fig. 3.13(b). The sudden decrease in accuracy for SwATrack tracking during iteration = 70, as shown in Fig. 3.14(b), is hypothesised to be due to the erratic generation of random values in the C++ implementation. This behaviour is not observed in other sequences, where their detection accuracy are consistent across frames. Thus far, an average result for each test case is computed over 10 runs to ensure the reliability of the results. Unbiased results without outliers can be obtained with a higher number of runs. In summary, the results further validate our findings that the proposed SwATrack is able to achieve better accuracy as compared to the PF, whilst requiring only about 10% of the amount of samples used in the PF, with minimal number of iterations. This is made possible by an iterative search for the optimal proposal distribution, incorporating available observations rather than making strict assumptions on the motion of an object. Thus, the findings from this study create prospects for a new paradigm of object tracking. Again, similar question is raised; *if there is a need to make complex existing tracking methods by fusing different models and algorithms to improve tracking efficiency? Would simple optimisation methods be sufficient?*

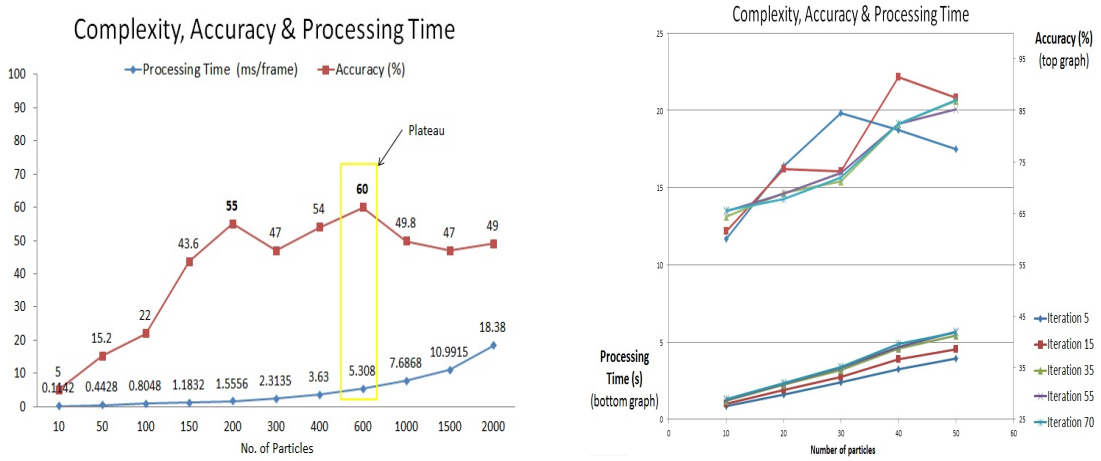
3.5.1 (b) Sampling Strategy

A sampling strategy test is performed in order to further evaluate the robustness of the proposed algorithm as well as to understand the behaviours of other algorithms when

tracking abrupt motion. In this test, a scenario of receiving inputs from the sensors with a lower frame rate is simulated. This is done by down-sampling the number of frames from the test sequence. Assuming that the actual data are obtained at normal rate of 25 frames per second to a lower rate of 5 frames per second. The down-sample frames of $TableT1-4$ sequences are denoted as $DoS-TableT1-4$. The $TableT5$ sequence is excluded in this test, as the object appears to be out of the scene in the early frames of this video, and thus the down-sampled sequence will comprise minimal number of frames in which the object appears in the scene.

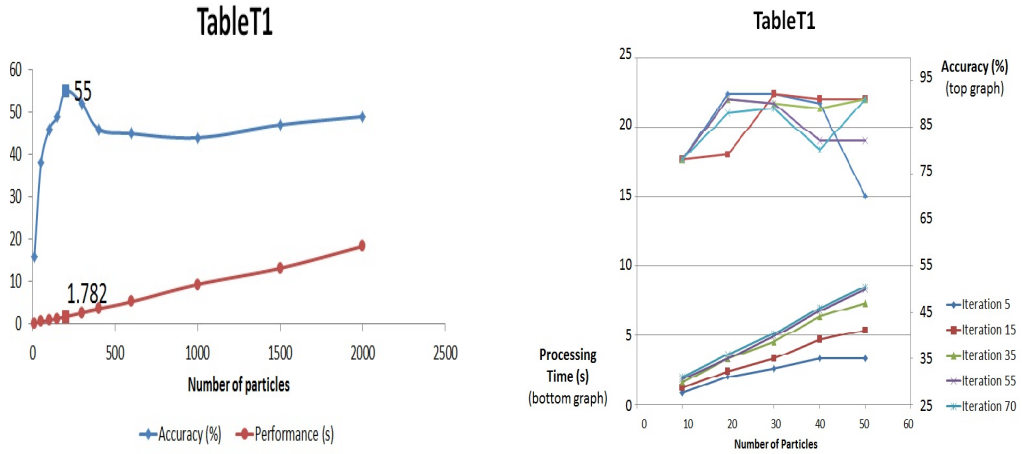
Fig. 3.18 demonstrates the detection accuracy between the proposed SwATrack and PF for all four sequences by down-sampling each sequence to simulate the 5 frames per second scenario. Note that the detection accuracy is determined by comparing the ground truth of the sampled frames only. It is observed that in general the proposed SwATrack has better detection accuracy as compared to PF in both situations; with and without sampling. The average detection accuracy of the SwATrack for the complete sequences is approximately 95.5% whereas the average for the PF is approximately 62.5%. During sampling, the average detection of accuracy of the SwATrack is approximately 77.25% whereas the PF is approximately, 32.75%. The experimental results show that the detection accuracy of the PF drops drastically when the frame rate decreases. This is because, in the low frame rate videos, the object tends to have abrupt motion and thus, methods that assume the Gaussian distribution in its dynamic motion model such as the PF, fail in such cases. The changes between the detection accuracy on both, the complete and sampled sequence is as indicated in red in Fig. 3.18. SwATrack on the other hand, copes better with low frame rate with an average accuracy of more than 70% although there is a decrease in its efficiency. This is because, the proposed SwATrack algorithm allows iterative adjustment of the exploration and exploitation of the swarm in search for the optimal motion model without making assumptions on the object's motion. This alludes

to the notion that the proposed SwATrack is able to deal with another scenario of abrupt motion, where the frame rate is low.



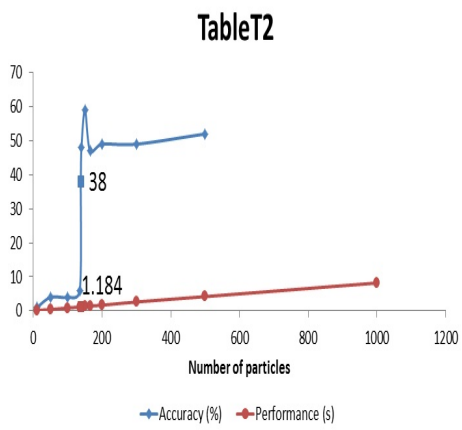
(a) PF (b) SwATrack

Figure 3.12: A comparison in terms of accuracy vs different number of samples and accuracy vs different number of samples and iteration.

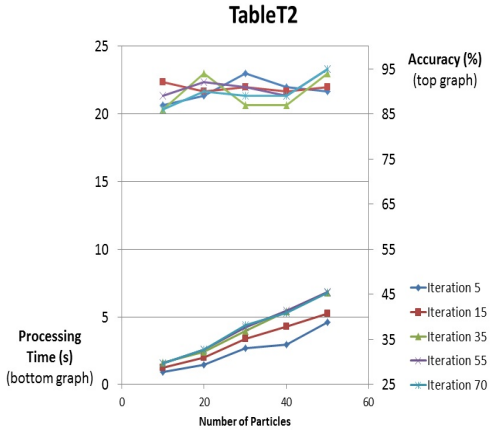


(a) PF (b) SwATrack

Figure 3.13: $TableT_1$: The accuracy and performance of PF and SwATrack with different parameter settings.

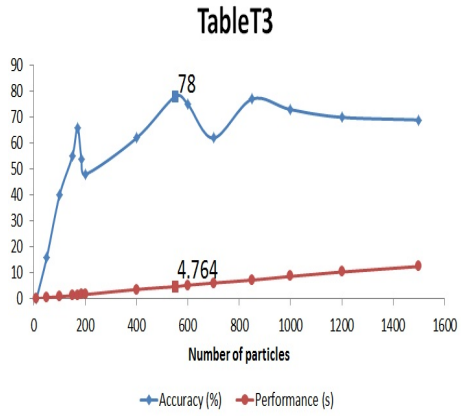


(a) PF

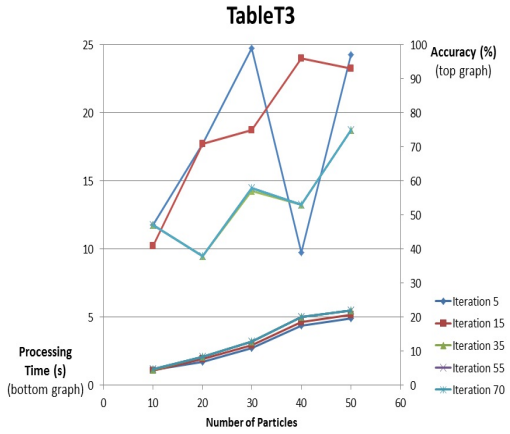


(b) SwATrack

Figure 3.14: *TableT*$_2$: The accuracy and performance of PF and SwATrack with different parameter settings.



(a) PF

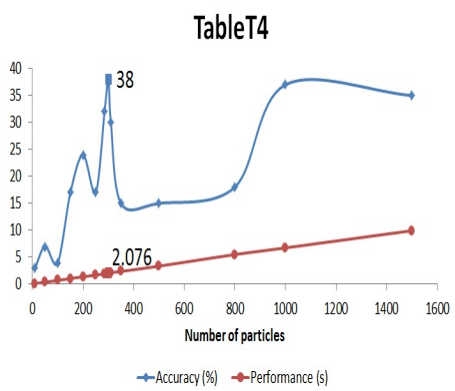


(b) SwATrack

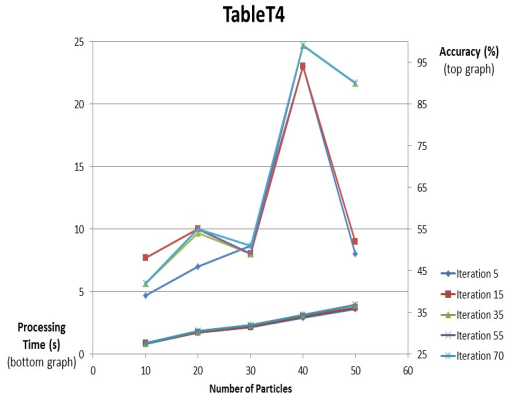
Figure 3.15: *TableT*$_3$: The accuracy and performance of PF and SwATrack with different parameter settings.

3.6 Summary

This chapter introduced a novel swarm intelligence-based tracker for visual tracking that deals with abrupt motion efficiently. The proposed SwATrack optimised the search for the optimal distribution without making assumptions or need to learn the motion model before-hand. Furthermore, the introduction of an adaptive mechanism that detects and responds to the changes in the search environment to allow on-the-fly tuning of the parameters allows a more flexible and effective solution. Unlike the conventional sampling-based tracking solutions which require a large number of particles for accurate

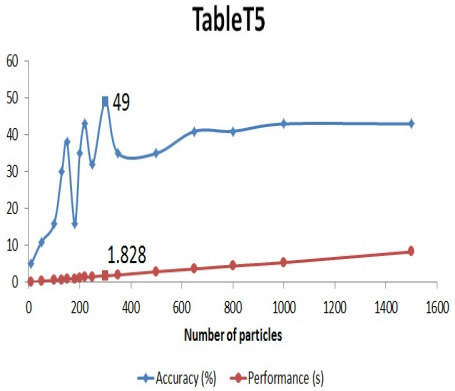


(a) PF

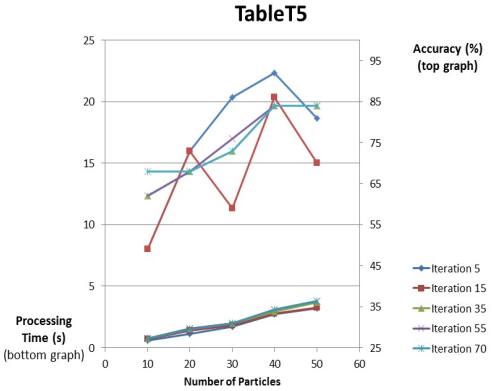


(b) SwATrack

Figure 3.16: *TableT*$_4$: The accuracy and performance of PF and SwATrack with different parameter settings.



(a) PF



(b) SwATrack

Figure 3.17: *TableT*$_5$: The accuracy and performance of PF and SwATrack with different parameter settings.

tracking, the sharing of information between particles in the SwATrack allows accurate tracking while keeping the number of samples at its minimal. To the best of the author's knowledge, this has never been done before. A new dataset - the Malaya Abrupt Motion *(MAMo)* dataset comprising 12 videos which are consolidated from benchmarked sequences, along with their ground truth tracking information is provided for future reference. A variation of experiments have been performed and the results have shown that the proposed SwATrack improves the accuracy of tracking abrupt motion while significantly reduces the computational overheads, since it requires less than 20% of the samples used by the PF.

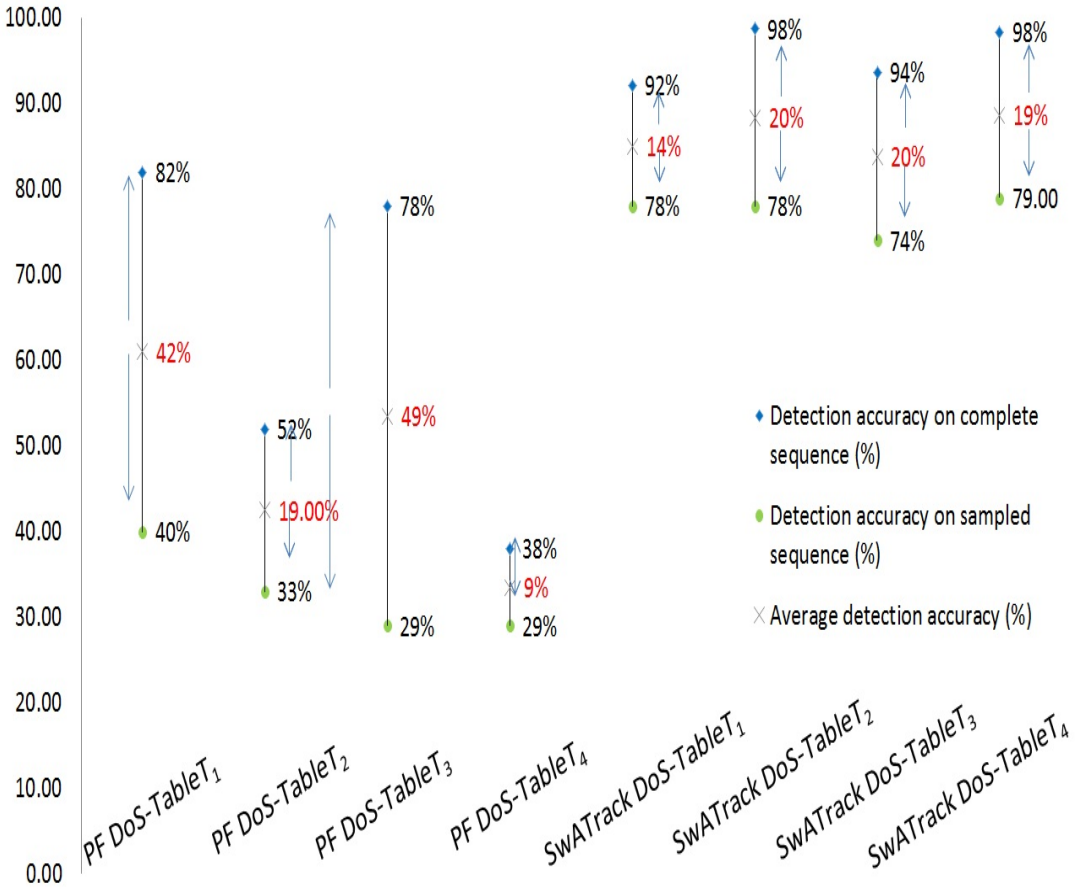


Figure 3.18: The detection accuracy of SwATrack against PF during sampling for $DoS - TableT_{1-4}$.

CHAPTER 4

CROWD BEHAVIOUR ANALYSIS

At large events such as rallies and marathons, where crowds of hundreds or even thousands gather, video monitoring is extremely challenging and complex. Identifying interesting regions in the crowded scenes, that could ultimately lead to unfavourable events, is a vital cue to direct the attention of security personnel (Schultz, 2008). Table 4.1 illustrates some cases of crowd disasters at mass gathering events. Carnage in crowd happens for a variety of reasons and have seen a two-fold increase in the past two decades (James et al., 2010; Ngai et al., 2013). In the developing world, crowd disaster often happens during religions festivals, such as the crush that occurred in Cambodia during the water festival. Meanwhile, in the developed world, soccer games and concerts are the most likely events to cause deadly crowds. Some examples of such disasters include the Love Parade and Hillsborough Stadium incidents. The aftermath investigations surrounding most of the crowd disasters conclude that there were missed opportunities for using technology to detect the abnormality of the crowd, which leads to such incidents (Klontz & Jain, 2013).

Therefore, this chapter proposes a framework that identifies and localises interesting regions in the crowded scenes. The focus is on exploiting the motion dynamics of the crowd to infer irregularity or abnormality in the regions that acquire attention. Unlike most of the existing solutions where a learning mechanism is required to group similar motion dynamics into clusters, this work alleviates the need for a learned model. This work suppresses the dominant flow information, while focusing on the unstable motion. In particular, the motion of individuals is assumed to follow the regular or dominant flow and the social conventions of the crowd dynamics. With this, interesting regions can be

considered as the extrema (instability) in the underlying crowd motion dynamics in the scene. Instead of disregarding the instability as noise, a simple yet effective idea of amplifying regions of unstable motion to infer the abnormality is proposed. This is then extended by projecting the low-level motion field into global similarity structure to allow the discovery of intrinsic motion structure that deals with more subtle scenario of abnormal crowd activities. In the context of this work, regions deemed abnormal or acquiring attention are denoted as salient, and refer to areas with high motion dynamics or unstable. In summary, the proposed method contributes in such a way that the amplification of unstable motion, along with the global similarity structure allows the discovery of abnormality in the crowded scenes. In contrast to existing literature, the presented solution requires **no pedestrian tracking**, **no prior information** of the scene or **extensive learning** to identify abnormality. Thus, it can be adapted to the environment over time and is more practical for real-world applications to prevent, or at least mitigate crowd disasters, by identifying bottlenecks, congestion and local irregular motion which may help in avoiding stampede and crowd crushes (Helbing & Mukerji, 2012).

This chapter is structured as the following: a brief introduction of the salient region detection in the crowded scenes is described in Section 4.1. Section 4.2 describes the proposed salient region detection in detail. This is followed by discussion on the experimental results in Section 4.3. The extended framework, which projects the local motion dynamics into global structure is described in detail in Section 4.4 and followed by the experimental results and findings in Section 4.5. Section 4.6 concludes this chapter.

Table 4.1: Examples of crowd disasters at mass events.

Date	**Event - Place**	**Description**	**Casualties**	**Reference**
Jan 1971	Ibrox disaster (football match) - Glasgow, UK	Crush between fans entering and exiting.	66 deaths, 140 injured	(Popplewell, 1986)
Feb 1981	Nightclub fire - Ireland	Fire was started deliberately in the alcove.	48 deaths, 128 injured	(Tribunal of Inquiry on the Fire at the Stardust, 1981)
Apr 1989	Hillsborough disaster (football match)- Sheffield, UK	Crush due to overcrowding surge against barrier.	96 deaths, 766 injured	(Taylor, 1989)
Jul 1990	The Hajj disaster – Mecca, Saudi Arabia	Crush caused by lack of directional flow of pilgrims and crowd control in the tunnel.	1426 deaths, no data is available for injured	(Alamri, 2014)
Jan 1991	Orkney stadium disaster - South Africa	Crush when fans panic and try to escape from brawls that break out in the grandstand.	40 deaths, 50 injured	(Darby & Mellor, 2005)
Jan 1993	New year's eve stampede – Lan Kwai Fong, Hong Kong	Slip and fall which leads to more and more people deprived of footing and fell; piling on top of another.	21 deaths, no data is available for injured	(K. K. Wu, Tang, & Leung, 2011)
May 1994	The Hajj disaster – Mecca, Saudi Arabia	Progressive crowd collapse caused by the sheer number of pilgrimages.	266 deaths, 98 injured	(Gad-el Hak, 2008)
Jul 2001	Akashi pedestrian bridge accident - Akashi Japan	Crush due to sudden panic during fireworks display.	11 deaths, 247 injured	(Yokota, Ishiyama, Yamada, & Yamauchi, 2002)

Date	Event - Place	Description	Casualties	Reference
Feb 2004	Miyun lantern festival disaster - Beijing China	Crush when a spectator stumbled on an overcrowded bridge and in the confusion people were crushed in an oncoming throng.	37 deaths, 24 injured	(Zhen, Mao, & Yuan, 2008)
Feb 2006	Philsports stadium – Manilla Philippines	Sudden surged forward with tremendous speed and force when the entrance gate was flung open, coupled with steep decline and uneven surface of the road which leads to dominoes effect.	74 deaths, 627 injured	(M. Lee, 2012)
Nov 2008	Wallmart black friday shopping - New York, United States	Tension grew as the opening time for the store approaches, where the density of crowd increases rapidly and was out of control.	1 death, no data is available for injured	(Ripley, 2008)
Jul 2010	Love Parade disaster - Duisburg, Germany	Crush due to unauthorised entry to the tunnel; entering fans converge with the exits.	21 death, 510 injured	(Helbing & Mukerji, 2012)
Nov 2010	Khmer water festival - Phnom Penh, Cambodia	Crush caused by bottleneck on the bridge and sudden panic in crowd.	347 death, >755 injured	(Hsu & Burkle, 2012)
Apr 2013	Boston marathon bombing - Massachusetts, United States	Two pressure cooker bombs exploded near the finishing line, where the crowd of spectators gather. The suspect was later identified and found to have abandoned the bag containing the bombs nearby.	3 death, 264 injured	(Starbird, Maddock, Orand, Achterman, & Mason, 2014)

4.1 Salient Region Detection

Research studies and aftermath investigations on the earlier crowd disasters or mass gathering incidents have shown significant progress over the recent years. Expert opinions such as in (Still, 2000; Helbing & Mukerji, 2012; Krausz, 2012) have reported contradicting causes of the crowd disasters. They rebutted the common misconceptions that these disasters are caused by *stampede*, *mass panic* and *trampling*. Instead, their research findings discover other observations that trigger *stampede*, *mass panic* and *trampling*, which then lead to crowd disasters. They recommended the use of these observations to aid situational awareness in crowd. Amongst the recommended observations that can help to assess the level of criticality in a crowded scene include: perturbation in the crowd density, stop-and-go waves, congestion (jams of people forming and growing), bottleneck, unauthorised entry, slip and fall or crawling activity and crowd turbulence.

Motivated by this, this chapter identifies salient regions which correspond to bottleneck, occlusion, instability, crowding, local irregular motion and sources and sinks. The automatic detection of salient regions is vital in assisting the authority to assess the criticality of crowd scenarios. In addition, it eliminates the need to manually annotate these regions for many surveillance applications, especially in the area of crowd control and analysis.

4.2 Proposed Salient Region Detection Framework

In this section, the salient region detection framework, which deals with crowded scenes is proposed. This section discusses in detail the implementation of the motion stability between a point and its neighbouring, as well as the projection of the motion flow into global similarity structure to represent the crowd motion dynamics. Furthermore, evaluations on various scenarios of the salient regions are demonstrated and discussed.

4.2.1 Motion Flow Representation

Each point in a given frame can be described as $p = (x, y; t)$ where $x = 1, ..., X$, and $y = 1, ..., Y$. X and Y refer to the width and height of the frames respectively, t is the frame under investigation, up to the maximum T frames, in a given sequence. Firstly, the velocity field of each point, $V(p) = (u_p, v_p)$ is estimated by employing the dense optical flow algorithm in (C. Liu, Freeman, Adelson, & Weiss, 2008). The velocity field of each point, $V(p)$, comprises of the horizontal, u, and vertical, v, velocity components.

The velocity components for each point are accumulated and an average velocity is calculated within an interval of time, comprising of τ frames. The mean optical flow at frame, t, is denoted as $(\overline{u}, \overline{v})$.

$$\overline{V} = \{\overline{u}, \overline{v}\} = \{\frac{1}{\tau}\sum_{t}^{t+\tau} u_p, \frac{1}{\tau}\sum_{t}^{t+\tau} v_p\} \tag{4.1}$$

Fig. 4.1 illustrates the proposed time instance analysis which is performed to obtain the smooth and consistent fields, where inconsistent velocity components (noise) are often reduced if not removed during the averaging step. Fig. 4.2a shows the velocity field of a cropped region at frame $t = 5$ whereas Fig. 4.2b shows the mean velocity field of the same region for frames, $t = 1 : 25$, where $\tau = 25$.

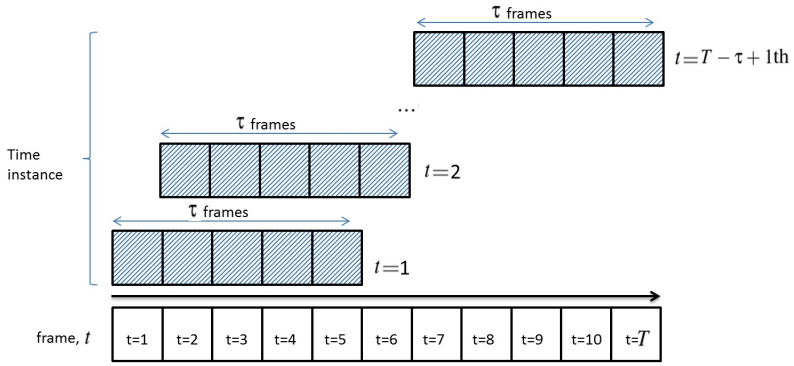


Figure 4.1: Graphical representation of the window-based analysis.

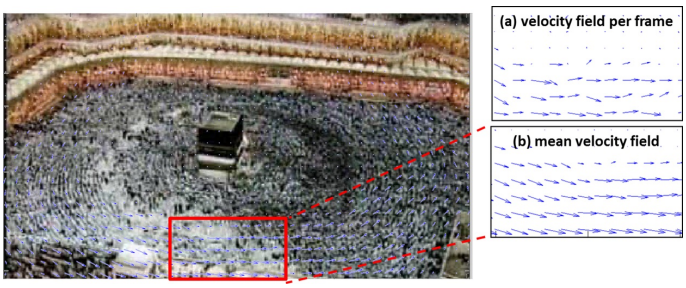


Figure 4.2: Flow field corresponding to the Hajj sequence.

The mean velocity field appears to be a good indicator of the global flow of individuals in a crowd (dominant flow), but may not be sensitive enough to capture the actual interactions and motion flows that deviate from the norm. Therefore, a particle advection process is implemented to keep track of the velocity changes in each point, p along its mean velocity field, (u,v). The basic idea of particle advection is to approximate the 'transport' quantity by a set of particles as proposed in (Moore, Ali, Mehran, & Shah, 2011).

$$\frac{d\vec{x}_p}{dt} = u_p(x_p,t;x_0,t_0) \tag{4.2}$$

$$\frac{d\vec{y}_p}{dt} = v_p(y_p,t;y_0,t_0) \tag{4.3}$$

$$\text{subject to } \; p = p_0 \;\; \text{ at } \;\; t = t_0 \tag{4.4}$$

The suffix, p, indicates the motion of a particular particle or point, where (x_0,y_0) represents the initial position of point, p_0, at time, t_0, while (x_p,y_p) denotes its position at time, t. Assuming that the initial position of p_0 is the mean velocity fields, $(\overline{u},\overline{v})$, the dynamic system can be deemed as an initial value problem. Thus, the pathlines which trace the points from their x_0 and y_0 positions at time, t_0 to their positions, x_t and y_t at time, t can be solved using the classical Runge Kutta method, as discussed in (Kennedy

& Carpenter, 2003). Unlike the conventional optical flow representation that captures the velocity of a pixel in two consecutive frames, the advected flow field captures the velocity of a particle in τ frames. The trace of particles over time forms a pathline which allows quantification of the motion dynamics, which is derived later from the separation coefficients between particles.

The proposed framework implemented the Jacobian method as in (Haller, 2000) to measure the separation between particle's pathlines which are seeded spatially close to a point, p, within a time instance, τ. The Jacobian of the flow map is computed by the partial derivatives of $d\vec{x}$ and $d\vec{y}$, where:

$$\nabla F_t(p) = \begin{bmatrix} \frac{\partial d\vec{x}_p}{\partial x_p} & \frac{\partial d\vec{x}_p}{\partial y_p} \\ \frac{\partial d\vec{y}_p}{\partial x_p} & \frac{\partial d\vec{y}_p}{\partial y_p} \end{bmatrix} \tag{4.5}$$

4.2.2 Stability Analysis

From the physical perspective, the partial derivatives of $d\vec{x}$ and $d\vec{y}$ give an indication of the slope of the tangent plane in the x and y directions; an indication of the rate of change in the flow with respect to x and y. According to the theory of linear stability analysis, the square root of the largest eigenvalue, $\lambda_t(p)$ of $F_t(p)^T F_t(p)$ indicates the maximum displacement, if the particle's seeding location is shifted by one unit as it satisfies the condition that $ln\lambda_t(p) > 0$. In the context of this study, a large eigenvalue indicates that the query point is unstable, and vice versa for a small eigenvalue. Note that in contrast to existing solutions (Ali & Shah, 2007; Weiss & Adelson, 1996; J. Yan & Pollefeys, 2006), where high motion dynamics are regarded as noise and thus removed; the proposed method exploits these unstable regions to infer salient regions. Fig. 4.3 illustrates the main difference between the two. The blue dots denote stable motion, where the spatio-temporal relationships and velocity changes between a point and

its spatially close neighbours are minimal. The stable motion flow often constitutes to coherent motion in existing works and thus unstable motion is disregarded and deemed as noise during segmentation. On the other hand, the proposed method analyses regions with unstable motion activities, represented by the red dots. The red dots demonstrate dynamic changes between a point and its spatially close neighbours from one frame to another. In a dense crowd scene, the motions of individuals tend to follow the regular or dominant flow (coherent) of a particular region. This is due to the physical constraints of the environment (i.e. path, junction) and the social conventions of crowd dynamics. Therefore, irregularities or abnormalities in the scene are identified when the motion dynamics of individuals differ from its close neighbours. The dynamics of a point within its spatially close neighbouring points is estimated by its stability and is represented as a map, $\phi = (x, y; t)$, where x=1,...X, and y = 1,...Y:

$$\phi_t = \frac{1}{|\tau|} log\sqrt{\lambda_t(p)} \tag{4.6}$$

This is then followed by the flow magnification of regions with high motion instability to synthesise the signal, while removing the insignificant ones using equation:

$$\hat{\phi}_t = \begin{cases} \beta.\phi_t, & \text{if } \phi_t \geq \alpha \\ (1-\beta).\phi_t, & \text{otherwise} \end{cases} \tag{4.7}$$

where, β is the magnification factor and α is the segmentation threshold. Fig. 4.4 demonstrates the output of the flow dynamics before and after the magnification, as in Eq. 4.6 and Eq. 4.7 respectively.

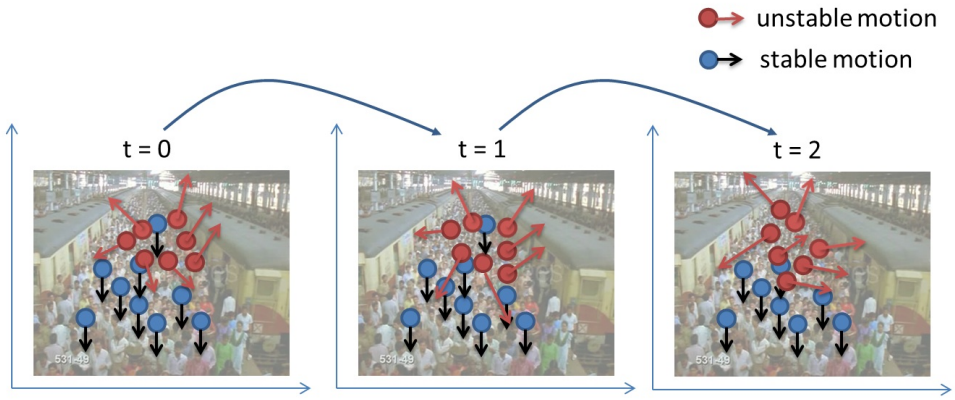


Figure 4.3: Illustration of stable and unstable motion dynamics. Best viewed in colour.

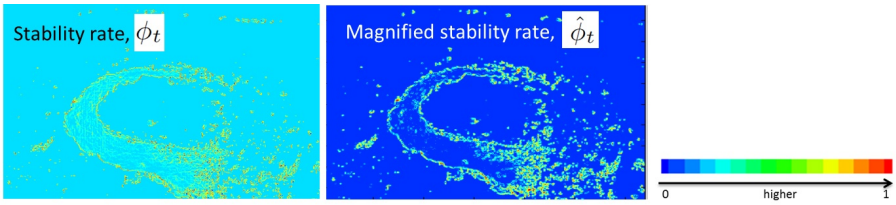


Figure 4.4: Illustration of the stability rate and magnified stability rate on the Hajj sequence. Note that the magnified stability rate amplified regions unstable regions while removing stable regions. Furthermore, the magnified stability rate shows a wider distribution between the stable and unstable points. Best viewed in colour.

4.2.3 Two Stages Segmentation

This chapter proposes two stages of segmentation that combines the output of fine and coarse segmentation obtained from the local and global segmentation steps. Briefly, the fine segmentation retains only the regions with high motion dynamics as seen in Fig. 4.6a. The coarse segmentation on the other hand, segments regions into clusters of coherent motion as shown in Fig. 4.6b. Note that the coarse segmentation is similar to most work in the literature, on motion segmentation. Nevertheless, the proposed two stages of segmentation are necessary in this context to allow accurate detections of the salient region. Otherwise, the output of the local segmentation alone is highly sensitive to noise, resulting in false detections. Meanwhile, the output of the coarse segmentation alone has a high tendency to cluster regions based on motion coherency, which is not the focus of this chapter; where the aim is to detect salient regions which corresponds to potential ab-

normality. In addition, the proposed two stages of segmentation alleviate the need to fine tune the best threshold values to be used for the segmentation of salient regions.

4.2.3 (a) Local Segmentation

In this study, the salient regions are observed to exhibit high values of $\hat{\phi}_t$, due to the high motion dynamics between points in a spatially close region. For example, individuals in a crowd tend to slow down when they approach an exit; increasing the motion instability in the exit region. In a regular flow, individuals tend to move with the crowd and thus exhibiting low motion dynamics (stable). Thus, a strict threshold is firstly applied to the stability rate to retain regions with high motion dynamics only, where $\Omega_t^L := \hat{\phi}_t \geq \alpha$. The parameter setting of α indicates the percentage of the stability rate to be retained. A higher value of α would retain regions with extremely high motion dynamics whereas a lower value of α would be more lenient in preserving regions with less motion dynamics. An opening morphology operation is then performed on the map to retain only the larger regions while removing the outliers. The sample output of the local flow segmentation (fine), Ω_t^L, is a local map that highlights the regions with high motion dynamics and is shown in Fig. 4.6a.

4.2.3 (b) Global Segmentation

Subsequently, a lenient threshold is applied to the stability rate for global segmentation. Most of the time, only the background or regions with minimal motion field are removed; $\Omega_t^G := \phi_t \geq \gamma$ and $\gamma \to 0$. After thresholding, the Ω_t^G comprises boundaries that are watershed-like, with many local optima. Thus, a variation of the watershed segmentation algorithm as introduced in (J. Shi & Malik, 2000) is adopted to cluster regions that are similar. The key idea is the use of the motion dynamics, Ω_t^G, as the feature similarity criterion to be optimised during clustering. A graph, $G = (V,E)$ is constructed by taking each pixel as a node and the edge weight, w_{ij} between node i and j as the product of a

feature similarity, Ω and spatial proximity, X terms:

$$W^{ij} = e^{\frac{||\Omega(i)-\Omega(j)||^2}{\sigma(\Omega)^2}} * \begin{cases} e^{\frac{||X(i)-X(j)||^2}{\sigma(X)^2}}, & \text{if } ||\, X(i)-X(j)\,|| \leq r \\ 0, & \text{otherwise,} \end{cases} \tag{4.8}$$

where $X(i)$ is the spatial location of node i, and $\Omega(i)$ is the corresponding separation field. The weight reflects the likelihood that the motion of the two pixels are coherent. Note that the weight, $W^{ij} = 0$, for any pairs of nodes, i and j that are more than r pixels apart. Fig. 4.6b shows sample shots of the coarse segmentation; each segment is filled with the quantised flow directions obtained from the mean flow field estimated earlier, in Eq. 4.1.

Finally, the coarse segmentation output is combined with the fine segmentation output to obtain the salient regions, Ω_t. The union operator is applied for this purpose, where $\Omega_t = \Omega_t^G \cap \Omega_t^L$. The separation of the task of segmentation into two stages alleviates the need for exhaustive fine tuning of the segmentation thresholds, α and γ. Fig. 4.6 demonstrates the pipeline of the two stages of segmentation and sample salient regions in the Hajj sequence. Since the proposed method is based on time instance, the two stages of segmentation, coupled with the magnification of unstable regions will eventually reduce the influence of stable regions overtime for more accurate detection.A summary of the overall algorithm is as shown in Fig. 4.5.

4.3 Experimental Results and Discussion

This section presents the detection results on real world crowded scenes. The proposed framework is developed in the Matlab environment and evaluated using an Intel® Core™ i7-3770 processor running on Windows 7.

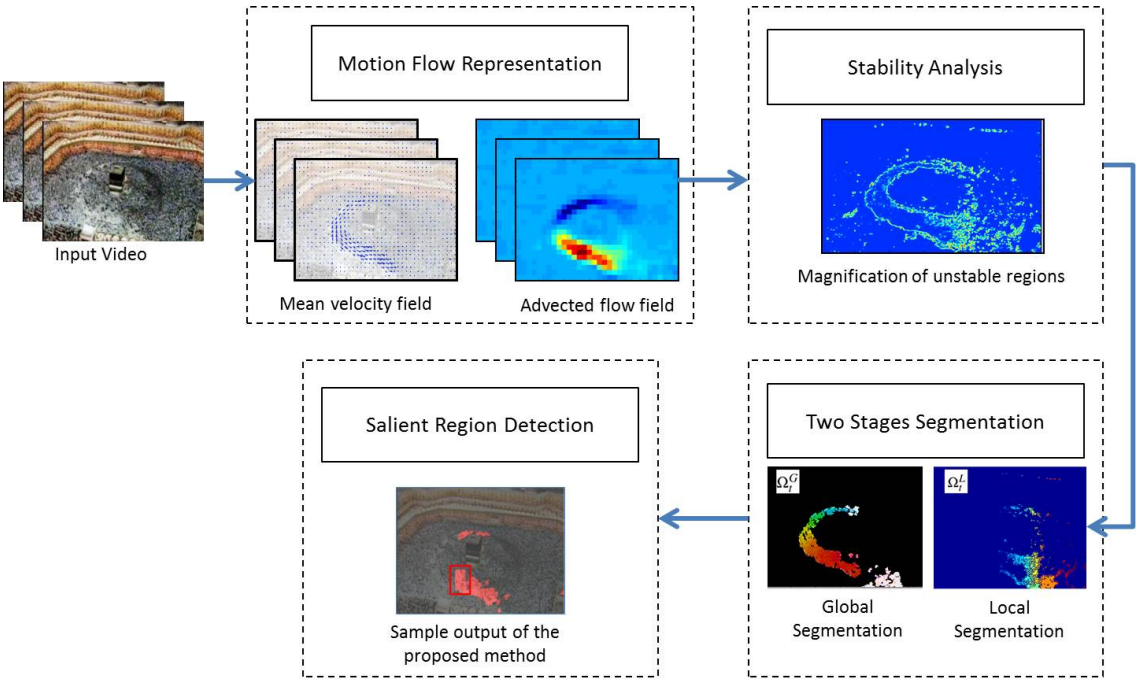
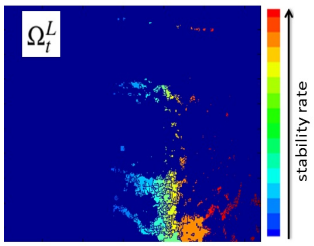
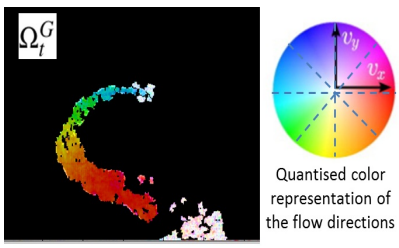


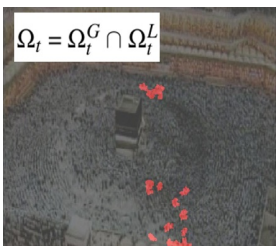
Figure 4.5: The framework of the proposed salient region detection method using two-stages segmentation.



(a) Sample output of the local segmentation.



(b) Sample output of the global segmentation.



(c) Sample output of the two stages segmentation, where the regions highlighted in red represents salient regions.

Figure 4.6: Sample output of the two stages segmentation process on the Hajj sequence.

4.3.1 Experiment Setup and Dataset

A set of 8 test sequences, comprising a variety of dense crowd scenes were used for evaluation and are illustrated in Fig. 4.7. These sequences are obtained from related literature and selected based on their relevancy in the context of exhibiting salient regions (Ali & Shah, 2007, 2008; Solmaz et al., 2012). Note that while there are numerous sequences on crowded scenes, most of the time, they are either depicting normal scenario or subtle salient region which is difficult to be noticed by the human eye. Thus, they are

omitted from this evaluation.

The first sequence is obtained from the National Geographic documentary, 'Inside Mecca', portraying the Hajj scene. Meanwhile, the second till fourth sequences depict the different marathon scenes obtained from real events. The fifth sequence is captured from a train station where a large crowd of passengers alighted from the train and moving towards the exit. The sixth sequence demonstrates a school of fish swirling and is added to demonstrate the capability of the proposed system to deal with non-human crowd which exhibits high motion dynamics. Finally, the last two sequences were of the same sequence as the earlier *(Hajj and Marathon1)*, but with the addition of synthetic noise to simulate changes in the motion dynamics (corrupted). In this experiment, the parameters used across all sequences are: $\tau = 25$ (at 25fps), $\alpha = 0.5$, $\beta = 0.1$, and $\gamma \to 0$. These values are defined based on the dataset used in this sequence. In practice, τ should depend on the rate of change of the flow field, with a higher rate of change of resulting in smaller time scales and vice versa.

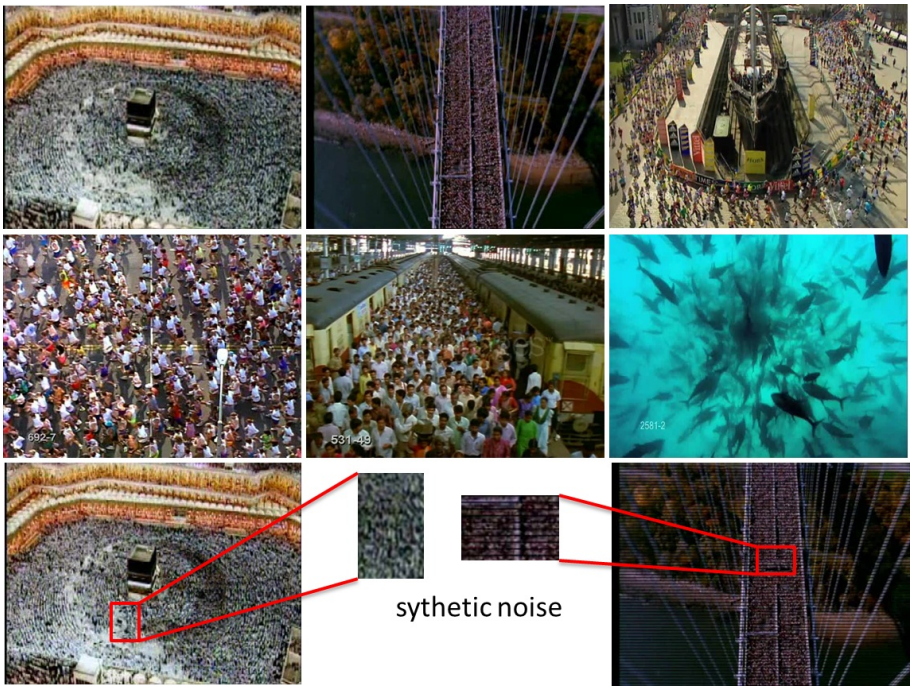


Figure 4.7: Sample shots of the dataset. Top row, from left to right, *Hajj, Marathon1 and Marathon2*; Second row, from left to right, *Marathon3, Train Station and School of Fish*; Bottom row, from left to right, *Corrupted Hajj and Corrupted Marathon1*.

4.3.2 Qualitative Result

The evaluation is categorised into 3 broad categories of triggers which may lead to crowd disasters including, instability or irregular motion, bottleneck and occlusion.

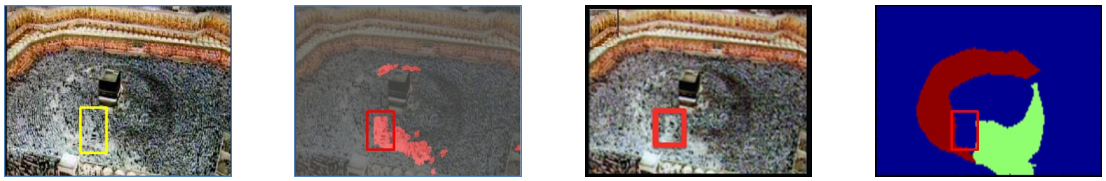
4.3.2 (a) Instability Detection

The corrupted test sequences, *Corrupted Hajj* and *Corrupted Marathon1* are used to test the capability of the proposed framework in detecting instability. It is worth noting that manual annotation of the various triggers of crowd disasters in real world scenarios is an open issue due to its ambiguous nature. Moreover, in an extremely crowded scene, it is very challenging to identify abnormality by the naked eyes. Often, the operator will have to go through the video sequences time and again in order to pick up the subtle cues. Thus, synthetic noise was injected into these two sequences to simulate the ground truth unstable regions as enclosed in the yellow bounding boxes in Fig. 4.8a and Fig. 4.9a. The synthetic instabilities were inserted into the original videos by flipping and rotating the flow of a random location into the specified ground truth position. A comparison between the proposed framework, Loy *et al.* (Loy et al., 2012) and Ali *et al.* (Ali & Shah, 2007) is performed.

As shown in Fig. 4.8 and 4.9, all three methods are able to identify the unstable regions accurately. However, the results show that in addition to the synthetic noise, the proposed method is able to identify other regions that exhibit high motion dynamics as highlighted by the red regions. In order to evaluate if the detected regions provide correct cues to potential triggers to crowd disasters, three informed and trained operators are employed to manually detect the salient regions. The manually detected salient regions serve as ground truth information for benchmarking purpose. After a thorough investigation as well as going through the original sequence time and again, the operators noticed that the detected areas by the system, indeed, correspond to the exit and turning point around the

Kaaba as shown in Fig. 4.8b. This is most likely due to the structure of the scene, or physical constraints of the Kaaba (situated at the centre of the scene); where the motion of the crowd slows down during the turning. Such saliency is difficult to be noticed by the naked eyes. In fact, the three operators noted this scenario only when they are asked to pay extra attention on the detected regions.

Again, it is important to emphasise the open issue in the current literature on identifying the ground truth of potential triggers to crowd disasters. This is due to the subjective nature of salient regions and the complexity of what constitutes to abnormality. Thus, this chapter argues that it is unfair to deem these detections as false positives. Instead, this chapter deliberates if these detections can assist in investigating and understanding the non-obvious motion dynamics of a scene.



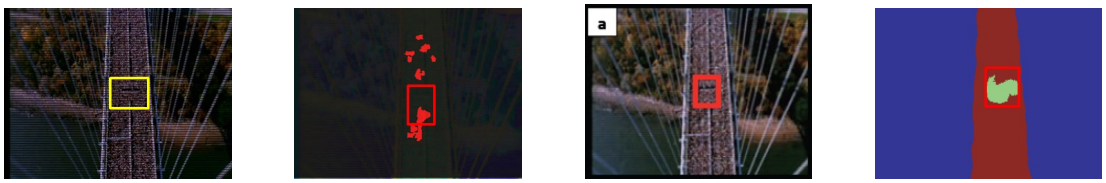
(a) Ground truth unstable region (yellow box).

(b) Output of the proposed method.

(c) Output of (Loy et al., 2012).

(d) Output of (Ali & Shah, 2007)

Figure 4.8: Comparison of unstable region detection using the *Corrupted Hajj* sequence. Best viewed in colour.



(a) Ground truth unstable region (yellow box).

(b) Output of the proposed method.

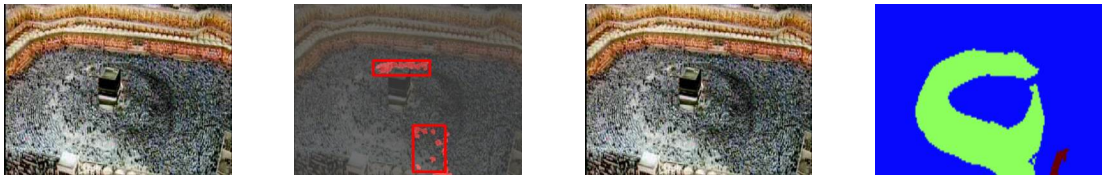
(c) Output of (Loy et al., 2012).

(d) Output of (Ali & Shah, 2007)

Figure 4.9: Comparison of unstable region detection using the *Corrupted Marathon1* sequence. Best viewed in colour.

4.3.2 (b) Bottleneck or 'Stop and Go' Detection

In the following, the original sequence of *Hajj* and *Marathon1* are used to detect bottleneck. The results in Fig. 4.10 and Fig. 4.11 show that the state-of-the-art methods, (Loy et al., 2012; Ali & Shah, 2007) are not able to deal with such abnormalities and do not have any detection on these sequences. On the contrary, our method is able to detect bottleneck caused by turning and exit areas such as shown in Fig. 4.10b and Fig. 4.11b. The detections of bottleneck has tremendous potential as an indication of impending danger such as stampede or overcrowding taking place, due to the stop-and-go waves or sudden build up in the crowd motion. This study reveals new insights into identifying abnormality in crowd motion. Potentially, the detected regions can be used to perform intervention for better crowd and congestion control.



(a) Hajj sequence without synthetic noise.

(b) Output of the proposed method.

(c) No detection from (Loy et al., 2012).

(d) No detection from (Ali & Shah, 2007).

Figure 4.10: Comparison of unstable region detection using the *Hajj* sequence, where the salient region is not obvious. Best viewed in colour.

(a) Hajj sequence without synthetic noise.

(b) Output of the proposed method.

(c) No detection from (Loy et al., 2012).

(d) No detection from (Ali & Shah, 2007)

Figure 4.11: Comparison of unstable region detection using the *Marathon1* sequence, where the salient region is not obvious. Best viewed in colour.

Further evaluation on the *Marathon2, Train Station* and *School of Fish* sequences demonstrate the capability of the proposed method to detect bottleneck in different scenarios of crowd. The detected salient regions are as enclosed in the red bounding box in

Fig. 4.12 and Fig. 4.13. It is observed that the detected regions in the *Marathon2* sequence correspond to obvious bottleneck near the entry and exit regions, where the crowd enters and disappears from the scene. Meanwhile, validating the detected regions for the *Train* sequence is not as straightforward. However, throughout the video sequence, the proposed framework identified the central region within the crowd as exhibiting high motion dynamics. This is most likely due to the bottleneck that happens as the crowd moves through a constraint pathway. Future investigation would include putting a barrier or obstruction in the detected salient regions in such scenes, to explore the probability of reducing traffic congestion or crowd evacuation that may occur due to bottlenecks.

Figure 4.12: Sample bottleneck regions in the *Train* and *Marathon2* sequences.

Fig. 4.13 demonstrates the detection results using the proposed method with another high density crowd scenario; the school of fish. In this sequence, an aggregation of schooling fish can be seen in the school fish sequence where the school of fish swirl in what looked like tightly choreographed manoeuvres towards the central of the scene. As shown in Fig. 4.13, the proposed framework is able to detect regions near the central of the swirl as salient regions. It is observed that the salient regions enclose a wider space of the central region across the frames before growing towards the right side of the image space. This is because the movement of the aggregation which gets more dynamic across the frames as more individual members are attracted towards the central point, 'whirpool' and later moving towards the right direction. The capability of the proposed system in detecting the 'whirpool' region would be useful in studies which are related to understanding the collective motion of fish.

Figure 4.13: Another scenario of high motion dynamics caused by the collective motion of fish.

4.3.2 (c) Occlusion Detection

In another variation, the *Marathon3* sequence with a visible occlusion or barrier is used. In this sequence, there is a huge street light appearing in the middle of a dense marathon scene. The proposed method is able to detect the regions near the street light as salient. See Fig. 4.14. The results demonstrate the capability of the proposed solution to not only detect instability and bottleneck, but also other regions that demonstrate high motion dynamics such occlusion. The automatic detection of salient regions is crucial in many surveillance applications, such as evacuation planning, traffic analysis and anomaly detection.

4.3.2 (d) Configuration Test

In order to investigate the influence of the parameters to the detection accuracy, this evaluation varies their settings. Fig. 4.15 demonstrates sample output for the *Hajj* sequence on the different frames using $\alpha = 0.5$, while other parameters are set to their default. It can be seen that while the detected salient regions differ slightly from one frame to another, the regions which correspond to the ground truth remains consistent throughout the frames. This shows that the proposed method which requires no learning stage can adapt to the environment and thus allows detection of changing salient regions. Meanwhile, existing work which requires learning stage would not be able to detect changing salient regions. The capability of dealing with changing salient regions is important in most real world applications, since the motion activity differs from one scene to another,

and from one time to another (i.e. traffic increases during peak hours).

Fig. 4.16 shows sample outputs using different values of α. A wider area of salient region is detected when the $\alpha \to 1$ while a $\alpha \to 0$ generates less salient regions. An empirical value of $\alpha = 0.5$ has been found to be satisfactory across all sequences. Finally, Fig 4.17 illustrates the sample outputs using the proposed two stages segmentation, local and global segmentation individually. It is observed that the global segmentation tend to cluster regions into coherent motion (i.e dominant motion) while the local segmentation is sensitive to noise and is prone to false alarms. The proposed two stages segmentation on the other hand, exploits the advantages from the coarse and fine segmentation, resulting in a better trade-off between the two outputs for a more accurate salient region detection.

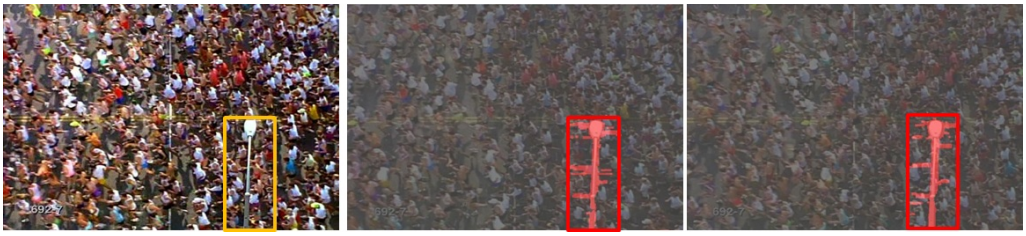
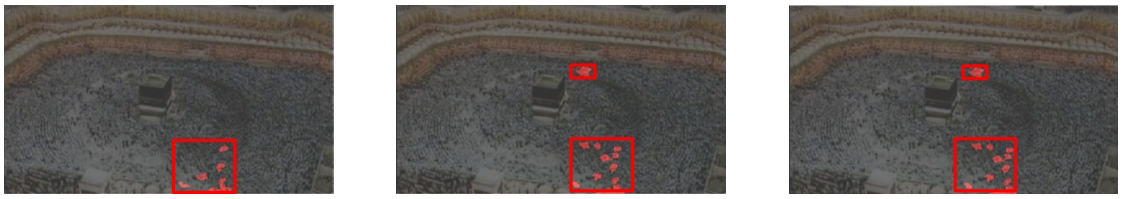
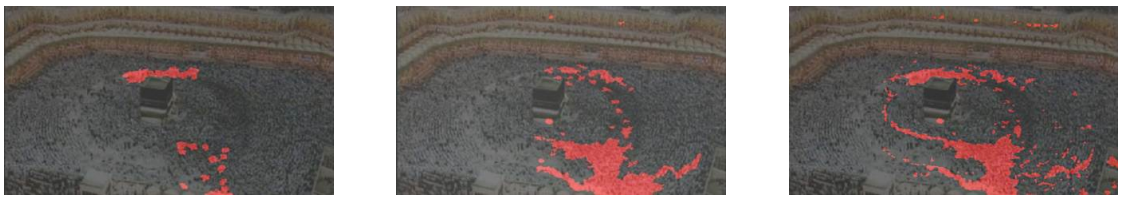


Figure 4.14: Sample occlusion region in the *Marathon3* sequences.



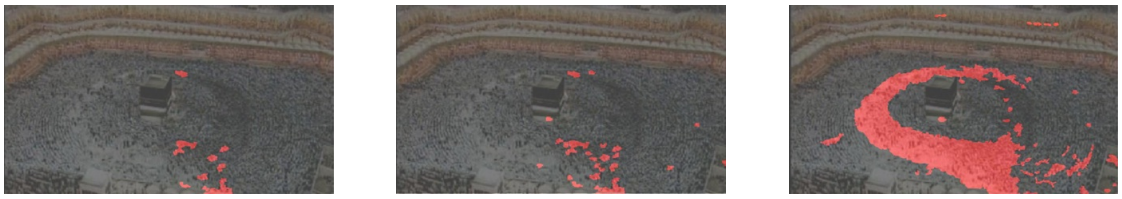
(a) $\alpha = 0.5$, at frame 5. (b) $\alpha = 0.5$, at frame 15. (c) $\alpha = 0.5$, at frame 25.

Figure 4.15: The detected salient region is consistent throughout the frames, and changes according to the motion dynamics of the crowd. Best viewed in colour.



(a) $\alpha = 0.5$, at frame 8. (b) $\alpha = 0.7$, at frame 8. (c) $\alpha = 0.9$, at frame 8.

Figure 4.16: The detected salient region grows as, $\alpha \to 0$. Best viewed in colour.



(a) $\Omega_t = \Omega_t^G \cap \Omega_t^L$, at frame 20. (b) $\Omega_t = \Omega_t^L$, at frame 20. (c) $\Omega_t = \Omega_t^G$, at frame 20.

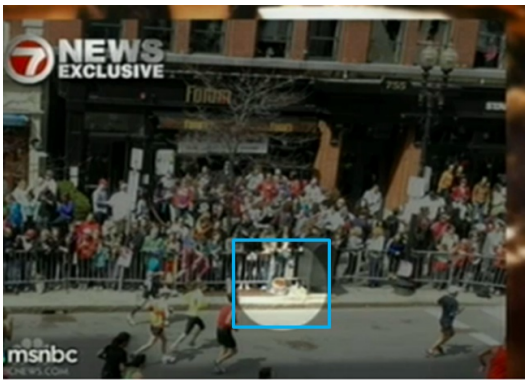
Figure 4.17: The detected salient region grows as, $\alpha \to 0$. Best viewed in colour.

4.4 Extended Framework

The crowd behaviour analysis framework proposed earlier, has the advantage of detecting salient regions caused by high motion dynamics such as bottlenecks. However, the stability rate which is derived locally does not deal with subtle motion changes. In the context of this study, subtle motion refers to local irregular motion which is usually caused by an individual or a small group of individuals moving against the dominant flow. For example, in the recent *Boston Marathon* bombing incident where two bombs exploded near the finish line and killed 3 people and injuring 264 others (Malone & McCool, 2013). Surveillance videos have been found to show the suspect, carrying backpack and walking nonchalantly in the area, before leaving the backpack containing the explosives as shown in Fig. 4.18a. In another example, Fig.4.18b illustrates the movement of individual against the dominant flow, where crowd of spectators are seated in the stadium. Therefore, the proposed framework is extended to include higher level representation of the stability feature to allow the discovery of the intrinsic manifold of the motion dynamics, which could not be captured by the low-level representation.

4.4.1 Global Motion Flow Representation

The global motion flow representation is a projection of the low-level stability feature extracted from the earlier framework, $\hat{\phi}_t$. The difference between the local stability, $\hat{\phi}_t$, of each point, p, with every other point, $p\prime$, is computed to represent the motion dynamics in a higher dimensional manifold. The global stability structure comprises $S_t \in I^{X \times Y}$, where



(a) The highlighted area is widely believed to be the site of where the bomb went off.

(b) The irregular motion of individuals against the crowd, who are seated is as highlighted in the blue bounding box.

Figure 4.18: Sample scenarios of local irregular motion.

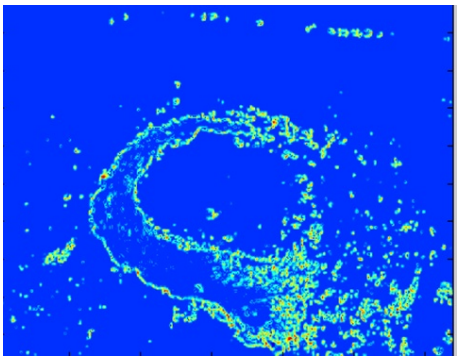
X and Y refer to the width and height of the frames respectively. Each point, S_i in the global structure captures the distance correlation, C, between every pair of pixels, $p(x,y;t)$ and $p\prime(x,y;t)$. Fig. 4.19 illustrates the local and global stability structure in two and three dimensional representations. It can be seen that the three-dimensional embedding of the global similarity structure obtained using multi-dimensional scaling, provides additional information on the intrinsic manifold of the motion dynamics. The combination of both feature representations therefore, allows discovery of a broader scenarios of saliency.

$$S_i(p,p\prime) = C_p(p\prime) \tag{4.9}$$

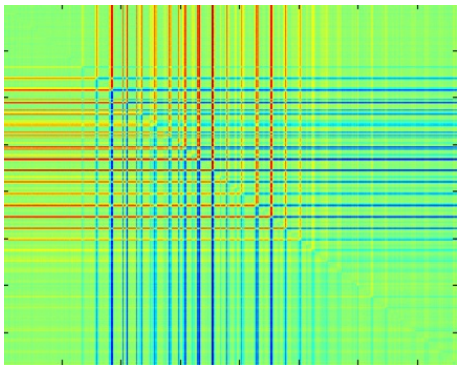
$$C_p(p\prime) = \sqrt{(p-p\prime)^2} \tag{4.10}$$

4.4.2 Ranking Manifold

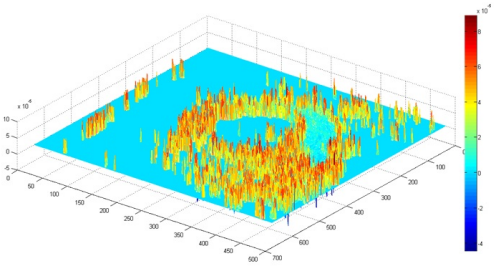
The aforementioned proposed two stages segmentation is not suitable for the S_t manifold, as the global stability structure lacks spatial information and is of a higher dimension. Thus, the ranking of data manifold as proposed in (J. Zhou D.and Weston, Gretton, Bousquet, & Schölkopf, 2004) is adopted to detect subtle salient regions using the ranking



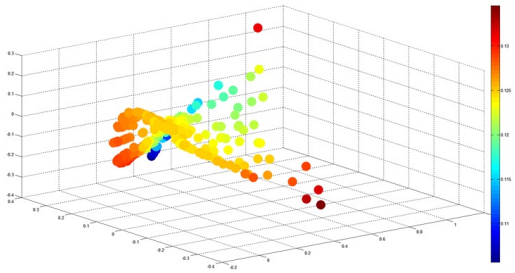
(a) Local stability structure (2D), ϕ_t.



(b) Global stability structure (2D), S_t.



(c) Local stability structure (3D), ϕ_t.



(d) Global stability structure (3D), S_t, obtained using multi-dimensional scaling.

Figure 4.19: Sample feature representation for the Hajj sequence. Note that the spatial information is absent in the global structure representation.

points with respect to the intrinsic manifold structure uncovered by the global structure. As opposed to related work such as in (M. Rodriguez, Sivic, Laptev, & Audibert, 2011), no learning stage is required for this purpose.

Each point in S_t is represented in the form of a weighted k-Nearest Neighbour (kNN) undirected network graph, $G = (V, E)$. Each vertex in the graph represents a data point, S_i. Two vertices are connected by an edge, E weighted by a pairwise affinity matrix, W_{ij}, and is defined as:

$$W_{ij} = exp\left(-d^2(r_i, r_j)/2\sigma^2\right) \tag{4.11}$$

where the parameter σ determines the width of the neighbourhood, $i \neq j$ and $W_{ii} = 0$, to avoid bias self reinforcement during the manifold ranking (J. Zhou D.and Weston et al., 2004). The distance metric, d, denotes the Euclidean distance. Given the affinity

matrix, W_{ij}, the connected graph, G can then be represented by using the symmetrical normalisation matrix (Laplacian), $L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where D is the diagonal matrix with $D_{ii} = \sum_j W_{ij}$.

Let f denotes a ranking function which assigns to each point, S_i a ranking value of f_i. f can be viewed as a vector, $f = [f_1, \ldots, f_n]^T$. The global stability structure, S_t can be represented as $S = \{S_1, S_2, \ldots S_m, S_{m+1}, \ldots S_n\}$, where $m, n = [X \times Y]$. The first m points are the queries (random), while the rest are the points to be ranked according to their relevance to the queries. A vector, y, is defined to store the label assignment of feature instances, where $y = [y_1, \ldots, y_n]^T$, in which $y_i = 1$ if S_i is a query, and $y_i = 0$ otherwise.

The connected graph, W_{ij}, is weighted and symmetrically normalised iteratively, where each point spread their ranking score to their neighbours via the weighted network. The spread process is repeated until convergence. The ranking score is computed using:

$$f(t+1) = \alpha L f(t) + (1-\alpha) y, \quad 0 \leq \alpha \leq 1 \tag{4.12}$$

where, L is the Laplacian graph, and α is the scaling parameter in the range of $[0, 1]$.

By performing ranking, the extrema can be detected as data points with the highest and lowest rank scores, deviating from the query points. Such extrema suggest salient regions caused by subtle motion change which correspond to local irregular motion. A summary of the overall algorithm is as shown in Fig. 4.20.

4.5 Extended Experimental Results and Discussion

Similar to the local motion structure framework, the proposed framework which utilises global motion structure is developed in the Matlab environment and evaluated using an Intel® Core™ i7-3770 processor running on Windows 7. This section presents further evaluation for salient region detection in crowd, including the subtle abnormality.

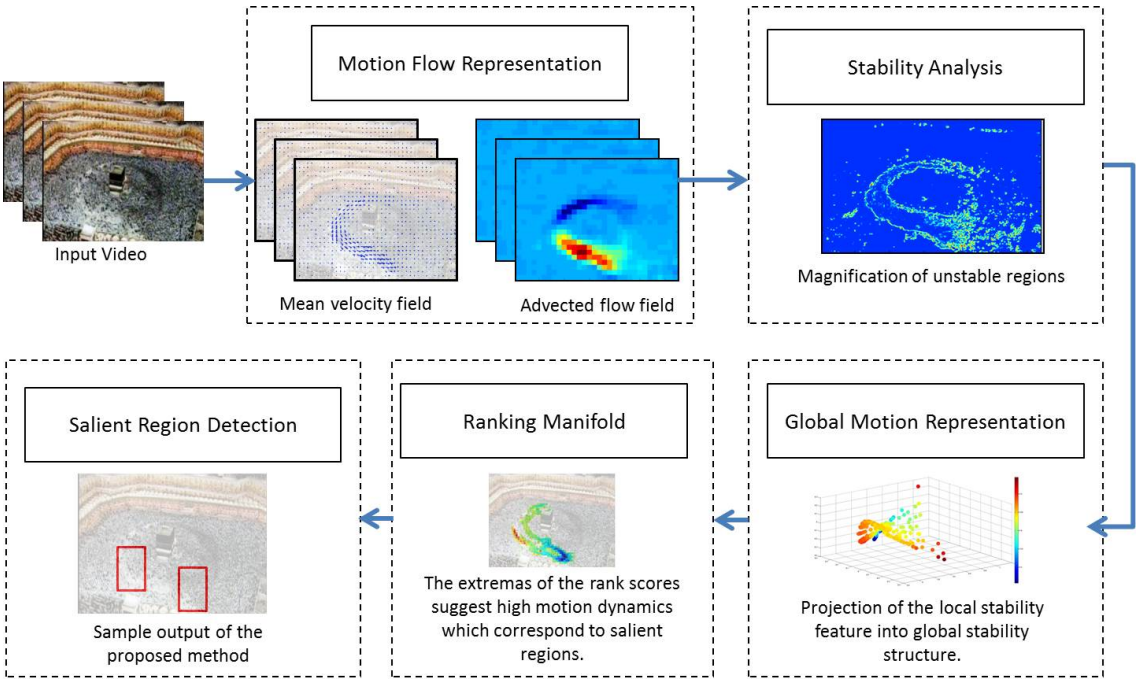


Figure 4.20: The framework of the proposed salient region detection method, using global similarity structure.

4.5.1 Experiment Setup and Dataset

For evaluation a set of 30 datasets obtained from benchmarked crowd dataset is used (M. Rodriguez et al., 2011; Loy et al., 2012; Solmaz et al., 2012). The sequences are diversified, representing dense crowd in public spaces in various scenarios such as pilgrimage, station, marathon, rallies and stadium. Each sequence have different field of views, resolutions, and exhibit a multitude of motion behaviours that covers both the obvious and subtle instability. Sample shots of the sequences used are shown in Fig. 4.21.

4.5.2 Qualitative Result

The qualitative evaluation on detecting subtle saliency is categorised into 2 categories of triggers comprising *local irregular motion* and *find Boston bomber*. Although in general, the two triggers are closely related and are not discriminated by the proposed framework, this section splits the two for easier understanding of the solution in real world applications.

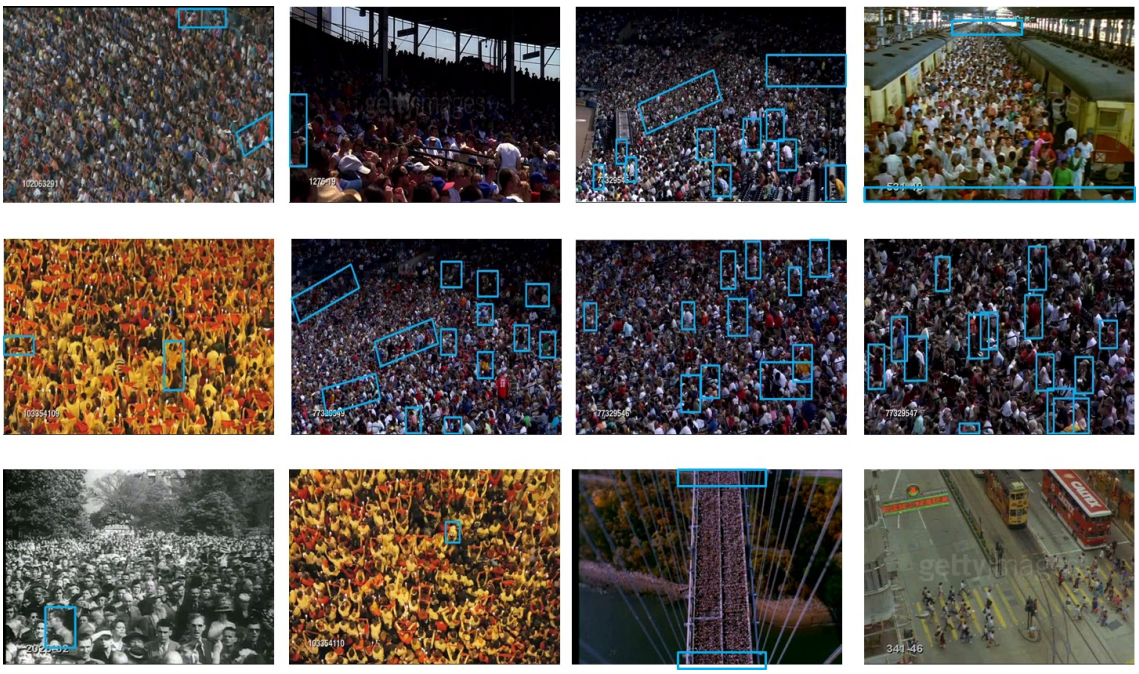
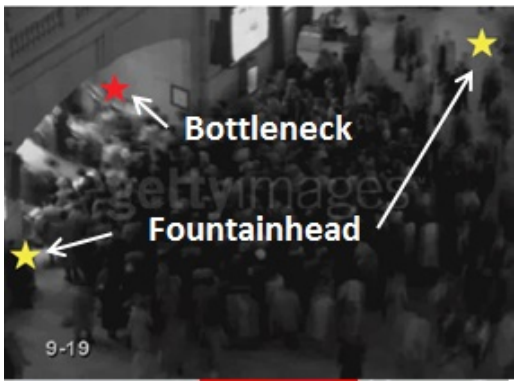


Figure 4.21: Sample test sequences comprising of the different scenarios of dense crowd. The blue bounding box depicts the ground truth salient regions, which exhibit local irregular motion in particular.

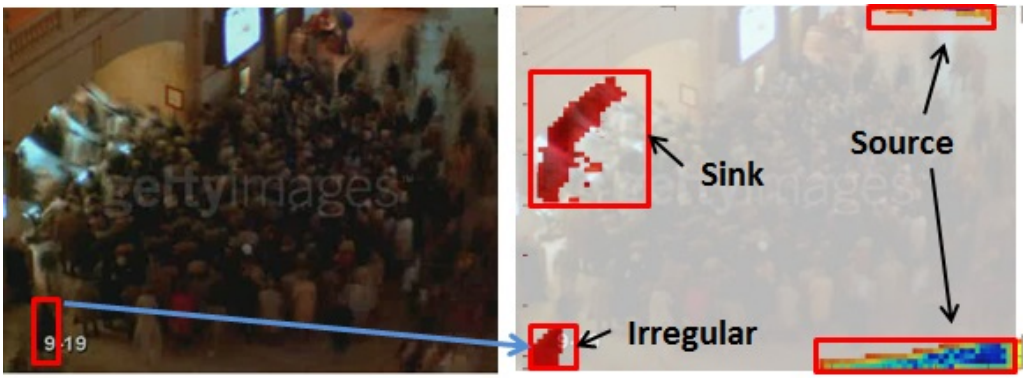
4.5.2 (a) Local Irregular Motion

A comparison is performed between the proposed work and Solmaz *et al.* in (Solmaz et al., 2012) using the sequence obtained from an underground station as depicted in Fig. 4.22. In this sequence, there is obvious source and sink regions which are denoted as bottleneck and fountain-head, respectively in (Solmaz et al., 2012). The results demonstrate the capability of the global stability structure framework to detect similar salient regions as in (Solmaz et al., 2012), with the addition of another source region at the bottom right of the scene. The proposed global stability structure is able to detect such subtle motion of someone walking into the scene from the bottom left corner of the scene. This is not the case in (Solmaz et al., 2012), where their detection does not highlight accurately the locality of the triggering event. It is worth noting that while the proposed methods are able to detect the different scenarios of abnormality caused by high motion dynamics, the salient regions are not characterised into different categories. Instead, these detections are deemed abnormal and are intended to provide cues for better understanding of the crowd.

(a) Original image.



(b) Output of (Solmaz et al., 2012).



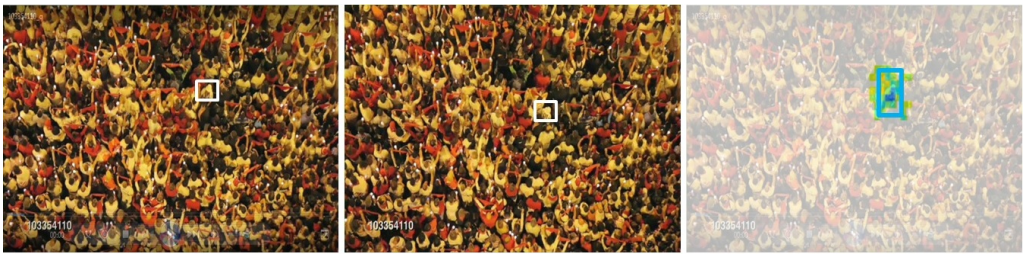
(c) Output of the proposed method.

Figure 4.22: Comparison of salient region detection using the *Marathon1* sequence, where the salient region is not obvious. Best viewed in colour.

4.5.2 (b) Find Boston Bomber

Further investigation on sequences that resembles the *Boston Marathon* bombing incident which exhibit an individual moving against the dominant crowd flow such as shown in Fig. 4.23 is performed. This scenario is to mimic the Boston Marathon Person Finder page launched by Google, which aims to identify individuals that seem suspicious. To the best of the author's knowledge, most of the conventional solutions fail to detect this type of anomaly, which is not obvious (Sarafraz, 2013; Solmaz et al., 2012; M. Rodriguez et al., 2011; Loy et al., 2012; Ali & Shah, 2007). On the contrary, the proposed global stability structure method which represents the motion flow in a higher dimensional space, consistently detects such anomaly as illustrated in Fig. 4.23.

(a) The proposed framework is able to detect an individual walking across the scene, when the rest of the crowd is seated.



(b) The individual maneuvering through an extremely crowded scene is detected as salient.

Figure 4.23: Find Boston Bomber: Example results of abnormality caused by local irregular motion. The ground truth is enclosed in the white bounding box in the first two columns, while the detected salient regions are as highlighted in the blue bounding box on the right most column. Best viewed in colour.

4.5.3 Quantitative Result

At present, most of the related works merely provide qualitative results and since this field of work is still at its infancy, the implementations are not shared publicly; leading to difficulties in performing a quantitative comparison for evaluation. As such, this section summarised the detection rate of the proposed method, against the manually generated ground truth for all the sequences obtained from the datasets which are publicly available. The anomalies are determined as per video basis and the *F-measure* according to the score measurement of the known PASCAL challenge (Everingham et al., 2010) is applied. That is, if the detected region overlaps the ground truth by more than 50%, then the estimation is considered as correctly identified salient region. The different scenarios of abnormality are categorised as crowding, sources and sinks and local irregular motion as shown in Table 4.2. The detections are further categorised into 3 categories, including *bottleneck*, *sources and sinks* and *local irregular motion* as can be perceived naturally by the human eye, as shown in Table 4.2. Bottleneck in this context is defined as poten-

tial build up in density or crowding that are typically affected by the physical structure of the environment. For example, near junctions where the crowd density builds up and thus, preventing smooth motion amongst individuals. Sources and sinks refer to regions where individuals in a crowd enter or leave the scene. Finally, the local irregular motion is triggered by flow instability of individuals or a small groups manoeuvring against the dominant flow in the scene. Fig. 4.24 illustrates sample outputs of the proposed framework on the various scenarios of abnormality in crowded scenes.

Table 4.2: Summary of the abnormal detection results.

Anomalies	Total # of Anomaly	# of True Detection	# of Missed Detection	# of False Detection
Bottleneck	20	23	3	0
Sources and Sinks	17	26	9	0
Local Irregular Motion	43	47	2	6

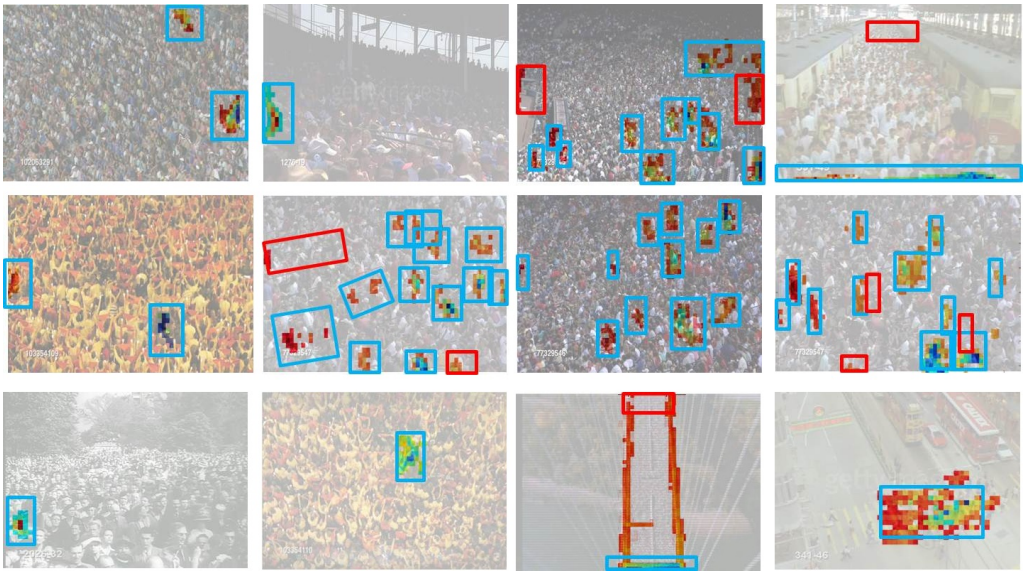


Figure 4.24: Sample output of the proposed algorithm. Red denotes false positive and false negative detections, while blue bounding box represents true positive. Best viewed in colour.

4.5.4 Comparison Result

In this section, results from the two proposed framework and its extension are compared and discussed. Fig. 4.25a and Fig. 4.25b illustrates sample comparison results. It is observed that the global projection of motion field enables extraction of intrinsic motion dynamics, thus allowing detection of subtle motion change caused by local irregular motion. This type of motion, such as an individual moving against the crowd flow, is not dealt with by the local representation of motion. Another drawback of the local representation, which the global representation could overcome, is the inability to detect saliency which appears near the boundary of the image. This is due to the advection process in the local framework which removes the boundary of the image to avoid out of bound estimation of the motion flow. As shown in Fig. 4.25c, the global framework is able to detect the sink region near the boundary, where the crowd disappear from the scene, while the local framework fails to detect such saliency. Note that the both frameworks are able to detect the potential bottleneck at the central area of the scene.

4.6 Summary

This section discusses the implementation of different frameworks for salient region detection in crowded scenes, where one is the extension of the other. The proposed methods eliminate the need to track each object individually, prior information or extensive learning to identify anomalies by observing the flow activities in a given scene for inference. In addition, the projection of the low-level motion flow into global similarity structure has been shown to be an effective indicator of subtle motion dynamics and irregularities in the crowded scenes. Experimental results show that the proposed frameworks are not only able to detect salient regions that correspond to the instability, bottleneck, or occlusion, but also local irregular motion which is subtle and difficult to be noticed by the naked eyes. Preliminary results demonstrate that the detections from the two pro-

(a) From left to right: Ground truth saliency; No output from the local framework; Output of the extended framework.

(b) From left to right: Ground truth saliency; No output from the local framework; Output of the extended framework.

(c) From left to right: Ground truth saliency; Output of the local framework; Output of the extended framework.

Figure 4.25: Comparison result between the proposed framework and its extension, where their detections complement each other for the various scenarios of saliency. Red denotes missed detection, while blue represents ground truth and true positives. Best viewed in colour.

posed frameworks are complementary and thus, are worthy of future integration work for a broader scope of salient region detection. It is acknowledged that at its present point, the experiment lack of comprehensive dataset. Furthermore, manual annotation of the ground truth is very challenging as it is not always that an entry or exit is made obvious by a door or gate. It is often very subjective to ascertain that a detected region is indeed an entry or exit region and this is made even more challenging in identifying bottlenecks. However, the promising preliminary results obtained are definitely worthy of future investigation since it is able to detect regions that are otherwise oblivious to the human operator. Future work would include a more comprehensive testing, a wider selection of datasets and further analysis on the detected salient regions to infer higher level semantics.

CHAPTER 5

MULTIPLE EVENTS DETECTION IN VIDEO SURVEILLANCE

Video surveillance has gained immense popularity across the globe due to the rising concerns for public safety and security. This leads to demand for technically advanced surveillance systems, thereby, creating huge growth opportunities for video surveillance market. Generally, video surveillance market can be segmented into cameras, servers and encoders, storage, monitors, and analytics or software. According to the global research conducted by Frost and Sullivan in (Frost & Sullivan, 2012), the market is expected to have generated revenues of about US$ 10.3 billion in 2010, with a growth rate of 9.8 percent over the previous year. In Malaysia for example, the last decade has witnessed significant advances in the field of video surveillance. Malaysia video surveillance market was estimated at over US$ 65 million in 2008 with compound annual growth rate of 27% by 2013 (Frost & Sullivan, 2009) and is as shown in Fig. 5.1. Most recently, the analysis of video surveillance market in Malaysia has been discussed further in (J. Lin, 2014), where the actual market revenues for Malaysian market from year 2011 to 2014 demonstrated a steady growth rate of 17%. Furthermore, the annual budget by the Government of Malaysia has shown tremendous interest in curbing crime and providing 'safer city'. This can be seen from the increase in budget allocation over the years. In the recent budget 2014, a total of RM3.9 billion fund was allocated to strengthen public safety (Tun Haji Abdul Razak, 25 October 2013).

Along with the enormous growth in the number of CCTVs deployed in public spaces, rises the need for intelligent analytical solutions (Velastin & Remagnino, 2006). Due to the large number of monitors to be observed closely, as well as other security tasks in hand, it is extremely challenging for human operators to perceive and interpret activities

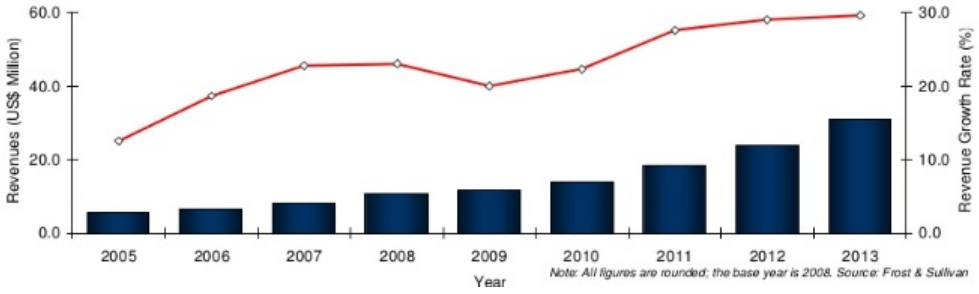


Figure 5.1: IPV6 market size and forecast in Malaysia.

taking place in the scene. Moreover, other factors such as fatigue, lack of knowledge, state of learning, confidence and integrity has been considered to influence the reliance of human monitoring in video surveillance (Dadashi, 2008; Keval & Sasse, 2008; Fookes et al., 2010; Bruckner, Picus, Velik, Herzner, & Zucker, 2012). Hence, there is a dire need to automate some aspects of the real-time surveillance systems. Automatic surveillance systems or analytics are used to detect, track and in higher levels solutions, to analyse the behaviour of objects in the scene (Valera & Velastin, 2005).

There have been considerable efforts in the industry as well as academia, which are focused on developing various algorithms and models for surveillance systems (Draganjac, Kovacic, Ujlaki, & Mikulic, 2008; Chan & Liu, 2009; C. Wang, Lu, Yang, & Liu, 2010; Chen et al., 2011; Albusac, Castro-Schez, Vallejo, Jiménez, & Glez-Morcillo, 2011; Albusac et al., 2014). These systems are commonly designed for specific video surveillance applications, which arise in favour of social welfare and public safety. Amongst the applications include traffic monitoring, loitering and intrusion detection. Favourably, analytics solutions that infer abnormal events or meaningful patterns that suggest complicated circumstances should be flexible enough to deal with multiple events. Current analytics solutions, however, often act separately to detect multiple events in different scenarios. For example, systems that perform loitering detection or/and abnormal trajectory in a given scene is based on two separate modules that work independently. Thus, they are usually not flexible or general enough to allow detections of different events at one time,

or to be generalised to other environments (Dick & Brooks, 2004).

5.1 Compositional Based Multiple Events Detection

This section aims to propose a novel framework, based on the principle of compositionality for multiple events detection in video surveillance. The framework deals with multiple events, on different regions-of-interest (ROI), at a particular time while utilising the low-level features of a given scene. The advancement from detecting singular event to multiple events provides a broader degree of scene understanding in automated video surveillance. This is very critical in the real-world scenarios where different (multiple) events may take place in a scene at the same time. For example, it is very likely that a loitering event happens at the same time as an abandoned luggage in a given scene.

By adopting the principle of compositionality from the Artificial Intelligence (AI) domain into video surveillance, the detection of multiple events in multiple regions is simplified and optimised. This principle proposes that the meaning of a complex expression is determined by the meanings of its constituent expressions, and the rules used to combine them (Pelletier, 1994). It is also referred to as the Frege's Principle, because Gottlob Frege is widely credited for the first modern formulation of it (Janssen, 2001).

The gist of this chapter is to conceptually decompose information obtained from a given scene into several intermediate degrees of abstractions. These low-level descriptions are then integrated and combined using a basic set of rule-packages, which discriminate between different abnormal events to build a complete knowledge of the given scene. In order to represent the contextual information of the scene using the proposed framework, this work investigates two main research questions: i) how to decompose and represent the modularised entities of the knowledge-based system in the video surveillance domain, and ii) how to apply the basic set of rule-packages to perform different abnormally events detection in a given scene.

The rest of this section is organised as the following. Section 5.2 provides the conceptual understanding to construct the proposed compositional-based framework, while Section 5.3 formulates the problem of multiple events detection in the context of compositional model in detail. This is followed by Section 5.4 which discusses a model application which is aimed at detecting multiple abnormal events using the proposed framework. Finally, the experimental results are presented and discussed in Section 5.5. The final section, in Section 5.6, concludes this study and provides insights as well as the future work.

5.2 Conceptual Understanding: Proposed Compositional-based Framework

Thus far, constructing the knowledge-based for video understanding is still an open issue due to the large variety and complexity of real-world scenarios that have to be dealt with, as well as the lack of formalised expertise on programs (Georis, Bremond, & Thonnat, 2007). This chapter explores the potential of exploiting the principle of compositionality for optimised abstractions of the complex surveillance problems.

Using the principle of compositionality, the surveillance problem is modularised into a set of variables comprising of low-level descriptions of the scene. This is made possible as surveillance systems generally involve monitoring different ROI in a given scene, classes (i.e. human, animal, vehicle, etc.) with various quantity dimensions (i.e. single or group) and attributes of the identified object-of-interest (i.e. direction, speed, locality, etc.). The underlying idea is that most of the events in surveillance can be hierarchically decomposed into low-level descriptions. By leveraging on this notion, the proposed method is able to optimise the reasoning of such complex events. This is in contrast to conventional methods, where redundant low-level processing is required to detect multiple events; since each event detection modules operate individually.

The principle of compositionality in logical semantics indicates that the meaning of

a complex expression is a function of the meaning of its constituent expressions and the set of rules used to accomplish them. Thus if an expression E is constituted by $E1$ and $E2$ under the constraint of some syntactic rule, then the semantic meaning of E, i.e. $M(E)$, is acquired by the combination of $E1$'s meaning $M(E1)$ and $E2's$ meaning $M(E2)$ abiding by some semantic rules (Szabó, 2007).

In the domain of video surveillance, the complex understanding of an abnormal event, E can be modularised into its constituents region-of-interest, elements, classes, attributes = $\{E1, E2, E3, E4\}$. Each constituent expression can then be constraint by a set of rules, or notion to create meaning, $\{M(E1), M(E2), M(E3), M(E4)\}$. These components are then combined to infer the complex expression, E. In short, the proposed compositional-based framework is constructed by means of addressing the following questions:

- Which **ROI** in a given scene will be analysed?
- What **class** of object will be monitored (object-of-interest)?
- What **attribute** is associated to each object-of-interest that will be analysed?
- What **notion** (a set of rules) to be applied to the respective attribute or a set of attributes to infer an event?
- What is the **event** that may have taken place? Where and when does the event happen?

5.2.1 General Overview

A scene, C, captured by CCTV(s) can be deemed as comprising multiple sub-environments defined as the ROI, $C = \{R_1, R_2, \cdots, R_N\}$; different activities or events may be taking place in each ROI. Generally, CCTV is fixed at optimum angle to monitor each of the ROI, R_n closely. These cameras are often activated when motion is detected in the field

of view. Otherwise, recording of the scene continues upon activation. In most of the real-time environments, it is practical to decompose a given monitored scene into different ROI as each region might have different activities taken place, and hence different set of rules should be applied for the event detection.

5.2.2 Pre-processing Stage

Pre-processing or also known as the low-level image processing is one of the most important step in any video surveillance applications (Bozdogan & Efe, 2011; Karasulu & Korukoglu, 2012; Cancela, Ortega, Fernández, & Penedo, 2013). The outputs from this step serve as the basic building blocks for higher level understanding and reasoning of the activities or events happening in the scene. In this chapter, several low-level image processing modules are adapted for evaluation of this framework, including background subtraction, object tracking and classification. Note that since the concern of this chapter is on the capability of the compositional-based framework, and not on introducing the best low-level processing techniques, these off-the-shelf methods are selected and revised accordingly, due to their reasonable trade-off between system precision and computational cost.

5.2.2 (a) Background Subtraction

The objective of the background subtraction module is to delineate the foreground from the background. In this study, the background subtraction model proposed in (Jacques, Jung, & Raupp Musse, 2005) is adopted due to its capability in handling illumination changes and shadow with minimal processing time.

In this study, we adapted the work as in (Jacques et al., 2005) due to its capability in handling illumination changes and shadow with minimal processing time. Using this method, each foreground pixels are given a unique label that separates them from the background pixels. The discrimination is done by analysing the intensity change of pixels

across frames and is as described further in the following:

$$V(x,y,\delta t) > (B(x,y) - T\sigma) \tag{5.1}$$

where, $V(x,y)$ is the input image of the video sequence within δt frames , and $B(x,y)$ is the initially learned background model. Each pixel (x,y) is classified as a foreground pixel if the difference is greater than $T\sigma$, where T is a fixed parameter (the empirical settings of in this study is, T=2), and σ is the median of the largest inter frames absolute difference (Jacques et al., 2005).

This is then followed by standard morphological filters to remove the noise and smooth the motion blob's boundaries. Here, the motion blob refers to the group of pixels identified as motion. A size filter threshold is then applied to remove groups of pixels that represent noise, where the noise is usually smaller in size (total number of pixels). The final output of this module is denoted as the motion map, in which connected pixels are grouped into clusters and given unique identifier. Example of the outputs of the background subtraction pre-processing step is shown in Fig. 5.2.

Figure 5.2: Left: Original frame. Right: Example output from background subtraction.

5.2.2 (b) Object Tracking and Classification

In order to estimate the locality of each object, the resultant motion map is fed into a object tracking module. In this framework, the blob-tracking approach as implemented in (S. L. Tang, Kadim, Liang, & Lim, 2010) is adapted. Using this previous-current relationships approach, it is critical to classify objects into four broad categories comprising of: i) new, ii) existing, iii) splitting and iv) merging as illustrated in Fig. 5.3. Each of the category is defined with a set of rules between the corresponding blobs, O in the previous and current frames as follows (j denotes the previous blob number while $j\prime$ represents the current blob number and i as the frame number):

Definition 5.2.1. *(Same Label) If the previous blob, $O_{j,i-1}$ overlaps with one current blob, $O_{j\prime,i}$, then the $O_{j,i}$ is corresponding to an existing object tracker. It will be continued to be tracked and labeled as the same object label assigned to $O_{j,i-1}$.*

$$O_{j,i-1} \cap O_{j\prime,i} \rightarrow \textit{same label} \tag{5.2}$$

Definition 5.2.2. *(New Object) If $O_{j\prime,i}$ does not overlap with any $O_{j,i-1}$, the blob will be considered as a new object to be tracked and this will be invoked to assign an object tracker for the newly identified object.*

$$O_{j,i-1} \cap O_{j\prime,i} = \emptyset \rightarrow \textit{new object} \tag{5.3}$$

Definition 5.2.3. *(Splitting) If $O_{j,i-1}$ overlaps with more than one $O_{j\prime,i}$, then splitting is detected.*

$$\left\{O_{0,i}, O_{1,i}, O_{2,i}, \cdots, O_{j\prime+1,i}\right\} \cap O_{j,i-1} \rightarrow \textit{splitting} \tag{5.4}$$

Definition 5.2.4. *(Merging) If more than one $O_{j,i-1}$ relates to a blob, merging scenario is denoted and a top-down tracking approach will be invoked to estimate the location of each of the object in the merged motion blobs. In this top-down tracking, prior information from the low-level processing tasks are incorporated for tracking.*

$$O_{j\prime,i} \cap \left\{O_{0,i-1}, O_{1,i-1}, O_{2,i-1}, \cdots, O_{j+1,i-1}\right\} \rightarrow merging \tag{5.5}$$

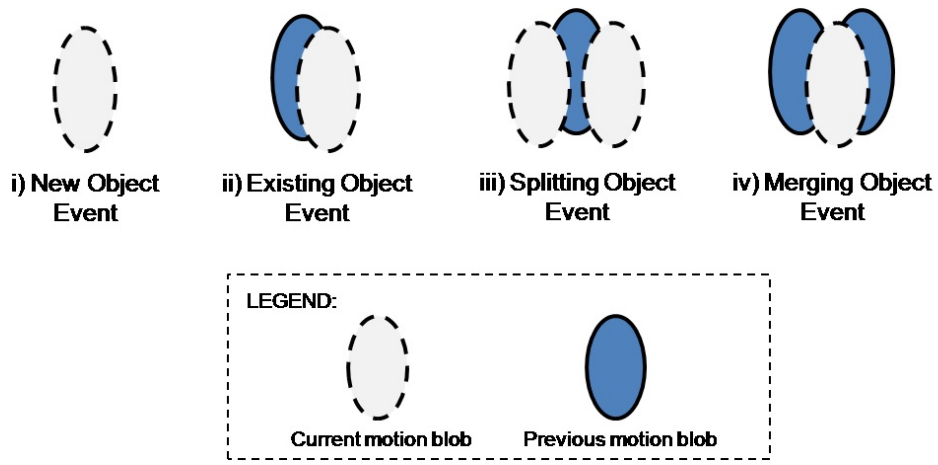


Figure 5.3: Graphical illustration of the previous-current motion blobs relationships.

Following this, the Modular Adaptive Resonance Theory Map (MARTMAP) as proposed in (Tan, Loy, Lai, & Lim, 2008) is adopted to classify each of the object into predefined classes. The predefined classes in this study include human, vehicle and luggage. The categories of classes are limited to the ones appearing in the set of test sequences used in this study. The prediction of class membership by the MARTMAP classifier is made by collectively combining the outputs from the multiple detectors. In contrast to other neural network methods for multi-class pattern recognition, the MARTMAP network has incremental learning capability and fast convergence (Tan et al., 2008). In the context of this study, the fast convergence criterion of the MARTMAP network allows discrimination between the different classes efficiently. Furthermore, new classes can be added or removed from the existing model as the need arises, without affecting the other trained

detectors (Tan et al., 2008). Thus, the MARTMAP is selected to evaluate this framework.

5.2.2 (c) Discussion

Although at present there are numerous low-level image processing methods that are more sophisticated or can produce better accuracy, there is always a trade-off between their precision and computational requirement. This finding has been supported by many studies including, *"Motion models that only work sometimes"*in (Cifuentes et al., 2012), *"Fuzzy qualitative human motion analysis"* in (Chan & Liu, 2009), *"Do we need more training data or better models for object detection?"* in (Zhu et al., 2012). Therefore, the proposed compositional-based framework adopts and revises off-the-shelf methods that are simple yet effective instead of complex ones. This is because; developing the most accurate analytics or event detectors that run independently of each other is not the focus of this study. Instead, the priority and the underlying notion that motivates this study is that most video understanding problems can be decomposed into similar set of low-level descriptions and thus, event detection can be optimised further using the principle of compositionality. In short, modularising the complex event detection problems into sets of basic descriptions of the scene, allows flexibility to detect multiple events under a single integrated framework.

5.3 Theoretical Understanding and Research Formulation

The general architecture of the proposed framework for multiple events detection is illustrated in Fig. 5.4. In general, the proposed framework categorises the process of detecting multiple events into 3 broad levels comprising the i) Sensory Level (SL) which refers to the data acquisition process, ii) Analysis and Reasoning Level (ARL) in which the knowledge of the environment is firstly constructed, followed by an analysis and reasoning using the principle of compositionality, and the the final level, iii) User Level (UL) where an alarm is triggered upon the detection of any abnormal event or to

alert the authority, so that the appropriate action can be taken immediately.

5.3.1 Analysis and Reasoning Level

As shown in Fig. 5.4, the ARL comprises a knowledge database to keep track of each event detection module, E(s). Note that E is scalable, depending on the given ROI, R, $E_N \leftarrow R_N$ and that the *Pre-processing* step is performed globally. This is to alleviate the need to perform redundant low-level processing tasks across different events. The ARL framework comprises three analysis stages that exploits the low-level cues obtained from the aforementioned pre-processing step. The three stages are defined as:

$$E_n = \left\{ e_n^1; \left\{ e_{1,n}^2, e_{2,n}^2, \cdots, e_{M,n}^2 \right\}; \mu_n \right\} \tag{5.6}$$

where, e^1 denotes the ***Primary Analysis***, e^2 is the ***Secondary Analysis*** and μ refers to the ***Reasoning*** stage. M is the total number of a series of e^2.

Definition 5.3.1. *(Primary Analysis, e^1) The main goal of the compositional-based framework is to decompose the complex expressions of events into the most basic constituents.*

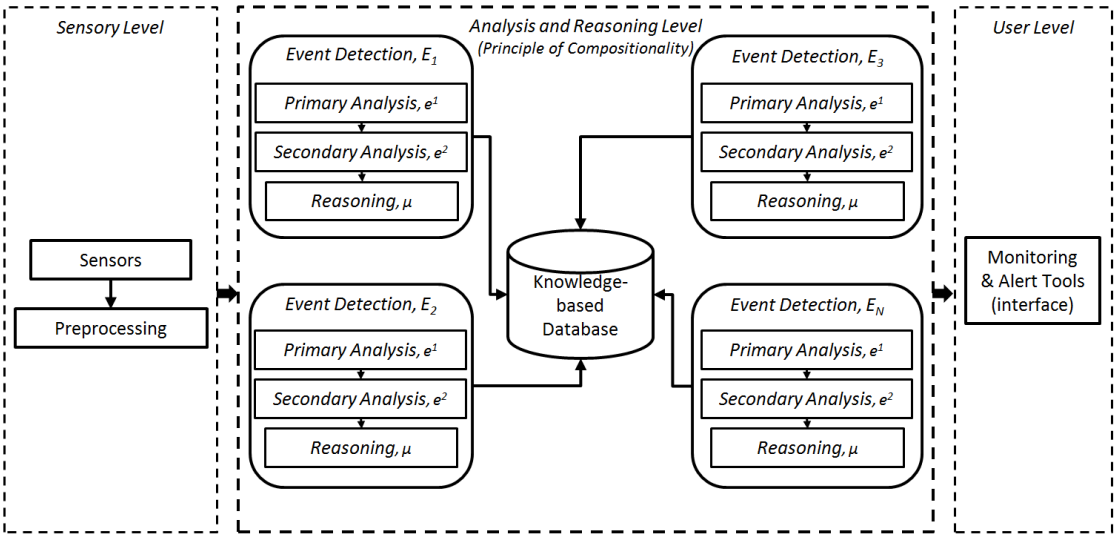


Figure 5.4: The general architecture of the proposed framework for multiple events detection.

The primary analysis is defined as:

$$e^1 = \{c, dim\} \tag{5.7}$$

where,

- *C is a set of classes , $C = \{c_1, c_2, \cdots, c_K\}$, K is the maximum number of classes.*
- *dim is an indication of the dimensions of the object-of-interest. An object can be associated to a single entity or group entity.*

Definition 5.3.2. *(Secondary Analysis, e^2) The secondary analysis is an extension of the outputs from the primary analysis, where each element in the primary analysis is associated to a set of attributes. The attributes include the temporal information (time span of an object appears in the scene), the geometric properties (profile of the motion map in the horizontal and vertical directions), and polarization (locality of an object) which can be represented as:*

$$e^2 = \{v, met\} \tag{5.8}$$

where,

- *V is a set of monitored attributes to infer an event happening in R_n, $V = \{v_1, v_2, \cdots, v_J\}$. J is the maximum number of attributes and comprises time span, speed, profile, locality, etc. Note that the set of attributes are not limited to the aforementioned and may be extended further according to the intended applications. V is derived from the extracted low-level descriptions. For example, speed is derived from the time information and displacement in locality within a duration of time.*

- *Met is a set of associated variables that infer the physical state of each attribute, based on a set of rules.* $Met^{v_j} = \left\{met_0^{v_j}, met_1^{v_j}, met_P^{v_j}\right\}$*, where P is the maximum number of state of variables. For example, the speed variable,* v_{speed} *can be associated to the state of* ***slow, moderate, fast****, where* $Met^{v_{speed}} = \{slow, moderate, fast\}$.

Definition 5.3.3. *(Reasoning,* μ *) The decision to determine if an object is exhibiting normal (*$\mu = 0$*) or abnormal (*$\mu = 1$*) behaviour is defined by means of a crisp set (object meets the constraints) with logical conjunction 'AND'.* μ *is a flag to indicate abnormal event and is considered by the following:*

$$\mu = \begin{cases} 1, & if\ (c_i \wedge Dim^{c_i}) \wedge \left(v_j \wedge Met^{v_j}\right) \wedge \cdots \wedge \left(v_j^{'} \wedge Met^{v_j^{'}}\right) \\ 0, & otherwise. \end{cases} \tag{5.9}$$

5.4 Model Application

The proposed decomposition of surveillance problem into compositional model is described by applying a real world, public surveillance scenario for better understanding of the model. Detailed description of each variables and their associated values are described. The goal of the proposed compositional-based model is to detect abnormal events taking place in different regions of a given scene. The abnormal events comprises of multiple events such as loitering, intrusion and slip and fall. Since the proposed model categorises an input environment into multiple ROIs, it is best suited for wide-area surveillance spaces which include the airport and railway stations. While there are many scenarios that cause abnormalities, this model describes the most common and critical scenarios for evaluation. They include the events of a) Loitering, b) Intrusion, c) Slip and Fall, d) Abnormal Crowd Activity, and e) Unattended Object.

Definition 5.4.1. *(Primary Analysis) The set of variables in e^1 is defined as:*

$$C = \{Human, Luggage, Unknown\} \tag{5.10}$$

$$Dim = \{Single, Group\} \tag{5.11}$$

Definition 5.4.2. *(Secondary Analysis) The set of variables and their corresponding values for e^2 are described as below:*

$$V = \{TimeSpan, Region, Profile, Speed, Direction\} \tag{5.12}$$

$$Met = \{\{Long, Short\}, \{In, Out\}, \{Horizontal, Vertical\}, \{Fast, Slow\}, \{Left, Right\}\} \tag{5.13}$$

Definition 5.4.3. *(Reasoning) This section describes the logical rules or notions that are applied to each event detector according to the definition of the intended module, E. The set of rules are specific to each attributes, and is as defined by the respective events. The reasoning stage is performed over an interval of time, Δt, using the moving window concept to infer an abnormal event. Note that, the moving window concept allows flexibility to configure the sensitivity of the detections, and Δt may vary from one event detector to another. In the following, the model application for the five different events is described based on the compositional model in detail.*

5.4.1 Loitering Detection

Loitering is defined as the act of lingering in a restricted area for more than a specified allowed duration (time). Loitering detection is critical as the act of loitering is often related to subsequent conduct of illegal activities. For example, a person who loiters around the at the entrance to a highly secured building are usually planning an illegal intrusion to the building or a person loitering at bus stops are often found to be dealing with drug trafficking (Bird et al., 2005). Intuitively, the act of loitering is often associated to the time span of a subject in the given scene as described in the first condition, **Condition 5.4.1** below. In another scenario of loitering, the condition can be extended to include the regularity of motion or pace in the scene. For example, in the departure walkway of a train station, where passengers who get off the train are moving towards the exit, the act of moving slowly and lingering at the same region for a long time is often deemed suspicious. Thus, the second condition, **Condition 5.4.2** that defines loitering considers the speed of the subject as well. Based on the proposed compositional framework, the conditions can be easily extended to deal with the variations of conditions that define a particular event.

Condition 5.4.1. *Given $R_1 \in S$, an object that belongs to Class = Human is not allowed to loiter in a restricted region for a long duration or time span as defined below.*

$$E_1 = \{\{Human, Single\}; \{\{Region, In\}, \{TimeSpan, Long\}\}; \{Abnormal, Flag\}\} \tag{5.14}$$

$$\mu_1 = \begin{cases} 1, & if\ (Human \wedge Single) \wedge (Region \wedge In) \wedge (TimeSpan \wedge Long) \\ 0, & otherwise. \end{cases} \tag{5.15}$$

Condition 5.4.2. *Given $R_2 \in S$, an object that belongs to Class = Human, is not allowed to remain in the restricted region with slow motion for a specific duration.*

$$E_2 = \{\{Human, Single\}; \{\{Region, In\}, \{Speed, Slow\}, \{TimeSpan, Long\}\}; \{Abnormal, Flag\}\} \tag{5.16}$$

$$\mu_2 = \begin{cases} 1, & if\ (Human \wedge Single) \wedge (Region \wedge In) \wedge (Speed \wedge Slow) \wedge (TimeSpan \wedge Long) \\ 0, & otherwise. \end{cases} \tag{5.17}$$

5.4.2 Intrusion Detection

Intrusion detection aims to detect scenario where a specified ROI is invaded and is extremely important for perimeter security (Norman, 2012). Generally, the act of intrusion is often linked to loitering, where both detectors are intended to keep people out of unauthorised areas. However, in contrast to intrusion that recognises unauthorised entry immediately, the latter allows leniency in terms of time span. Loitering allows entrance to a particular area, provided that the subject does not stay in the region for more than a specified duration. Intrusion is more sensitive towards detecting any motion in the specified ROI as compared to loitering detector and is extremely useful for applications such as perimeter monitoring and border control.

Condition 5.4.3. *Given $R_3 \in S$, object that belongs to Class = Human is not allowed to appear in an restricted region.*

$$E_3 = \{\{Human, Single\}; \{Region, In\}; \{Abnormal, Flag\}\} \tag{5.18}$$

$$\mu_3 = \begin{cases} 1, & \textit{if } (Human \wedge Single) \wedge (Region \wedge In) \\ 0, & \textit{otherwise.} \end{cases} \tag{5.19}$$

Condition 5.4.4. *Given $R_4 \in S$, object that belongs to Class = Human, which is moving towards a specific direction (e.g. reversed traffic flow) is not allowed in a particular region. This extended condition for intrusion detector is well suited in identifying opposing traffic flow. Intrusion detector is made complex with the addition of motion direction attribute, to detect unauthorised motion. One example application is to detect the act of 'u-turn' or turning back at security check points. In the example below, motion towards the left direction with respect to the image space is not allowed.*

$$E_4 = \{\{Human, Single\}; \{\{Region, In\}, \{Direction, Left\}\}; \{Abnormal, Flag\}\} \tag{5.20}$$

$$\mu_4 = \begin{cases} 1, & \textit{if } (Human \wedge Single) \wedge (Region \wedge In) \wedge (Direction \wedge Left) \\ 0, & \textit{otherwise.} \end{cases} \tag{5.21}$$

5.4.3 Slip and Fall Detection

Slip and fall detection is important to allow immediate assistance to the victims, and is extremely critical in public areas, nursery or elderly homes. Intuitively, the most significant and direct hint to infer a fall event is the change of profile of the motion shape, from a vertical distribution to a horizontal distribution. In addition to the profile attribute, the speed of the fall motion is considered to distinguish similar activity such as lying down from the event of slip and fall. This alludes to the fact that the speed of movement during a slip and fall event is faster as compared to the act of lying down. This framework

analyses the speed and profile of the object-of-interest for a more robust detection.

Condition 5.4.5. *Given $R_5 \in S$, object that belongs to Class = Human is deemed as exhibiting slip and fall iff the Speed = Fast and Profile = Horizontal.*

$$E_5 = \{\{Human, Single\}; \{\{Speed, Fast\}, \{Profile, Horizontal\}\}; \{Abnormal, Flag\}\} \tag{5.22}$$

$$\mu_5 = \begin{cases} 1, & if\ (Human \wedge Single) \wedge (Speed \wedge Fast) \wedge (Profile \wedge Horizontal) \\ 0, & otherwise. \end{cases} \tag{5.23}$$

5.4.4 Abnormal Crowd Activity Detection

Abnormal crowd activity is associated to a group of people, acting in a more aggressive manner collectively as compared to the norm (i.e. running away when there is threat or harmful incidents such as bombing) (N. Li & Zhang, 2011). In the context of this study, abnormal crowd activity is defined as sudden dispersal event, where the crowd activity exhibit irregular motion patterns (i.e. running towards different directions or dispersing from a central point) due to panic escape and evacuation. Abnormal crowd activity is critical as it is often an indication of threat or incident is taking place.

Condition 5.4.6. *Given $R_6 \in S$, a group of object that belongs to Class = Human is deemed as exhibiting abnormal crowd activity when its Speed = Fast.*

$$E_6 = \{\{Human, Group\}\}; \{\{Speed, Fast\}, \{Direction, Irregular\}\}; \{Abnormal, Flag\}\} \tag{5.24}$$

$$\mu_6 = \begin{cases} 1, & \textit{if } (\textit{Human} \wedge \textit{Group}) \wedge (\textit{Speed} \wedge \textit{Fast}) \wedge (\textit{Direction} \wedge \textit{Irregular}) \\ 0, & \textit{otherwise.} \end{cases} \tag{5.25}$$

5.4.5 Unattended Object Detection

Unattended object detection provides an alert when an object such as luggage is abandoned or left unattended in the public space for a specified duration. For this evaluation, the event of unattended object is defined as i) a static object dwelling in a given region over a specified duration and ii) a static object which belongs to non-human class. This event detector is extremely important as the act of abandoning an object can be considered as potential security breach in public safety from terrorism. This is especially true considering all of the terrorist attacks that have happened over the past decade. The source of most major terrorist attacks has been unattended objects. For example, in the recent *Boston Marathon* bombing incident, where a bag containing explosives was left unattended in the incident area.

Condition 5.4.7. *Given $R_7 \in S$, an object that belongs to Class = Luggage is deemed as an unattended luggage iff its TimeSpan = Long.*

$$E_7 = \{\{\textit{Luggage}, \textit{Single}\}; \{\textit{TimeSpan}, \textit{Long}\}; 1\} \tag{5.26}$$

$$\mu_7 = \begin{cases} 1, & \textit{if } (\textit{Luggage} \wedge \textit{Single}) \wedge (\textit{TimeSpan} \wedge \textit{Long}) \\ 0, & \textit{otherwise.} \end{cases} \tag{5.27}$$

Fig. 5.5 demonstrates a sample model application represented in a tree structure. It

is highlighted that based on the proposed compositional model, the conditions for each respective events can be easily extended and added without the need for a learning stage. Moreover, the decomposition of the conditions into primary, secondary and reasoning layers allows optimised detection of multiple events; since most of the events are leveraged from similar low-level features.

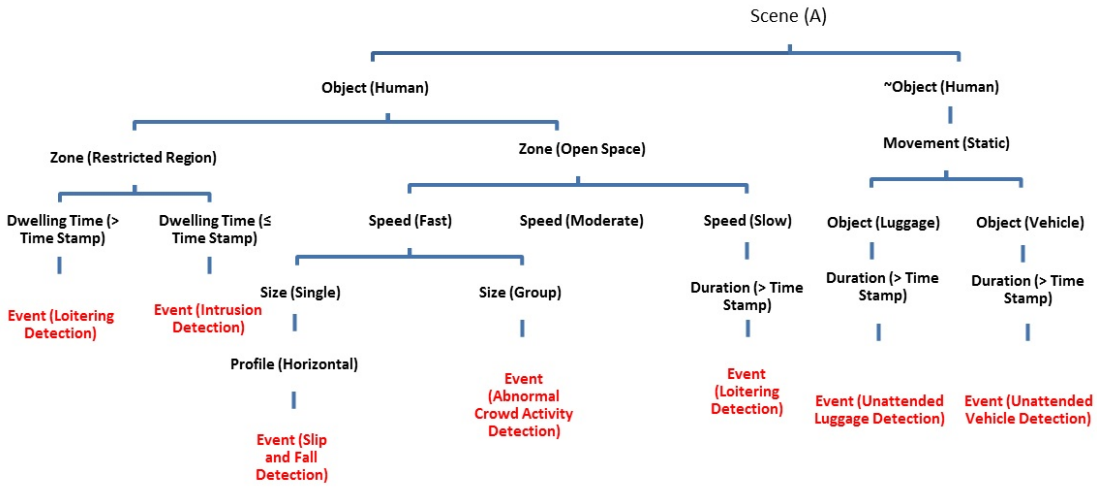


Figure 5.5: Sample tree representation of the model application used for evaluation.

5.5 Experimental Resuts and Discussion

The main goal of this experiment is to evaluate the efficiency of the proposed compositional-based framework in detecting different scenarios of abnormal events. The proposed system is implemented in C++, using the OpenCV image processing library. All experiments were performed on an Intel (R) Core(TM) 2 with CPU frequency of 1.8 Ghz and 2G RAM. Experimental results have demonstrated the performance and robustness of the proposed framework in providing flexibility and efficiency for multiple events detection.

5.5.1 Experiment Setup and Dataset

Each of the five abnormal events discussed in Section 5.4 are tested on 20 datasets (with one or two events in each dataset), respectively. Each dataset comprises a combination of video sequences obtained from standard dataset such as the PETS 2006 and

2007, dataset[1], CANTATA dataset[2], UMN dataset[3] and from the Youtube.com[4] [5] [6]. The sequences comprises mixture of staged and real activities (e.g. unattended luggage, slip and fall, loitering, intrusion, crowd dispersal). Most of the videos were captured at 25 frames per second and stored using the MPEG4 compression format.

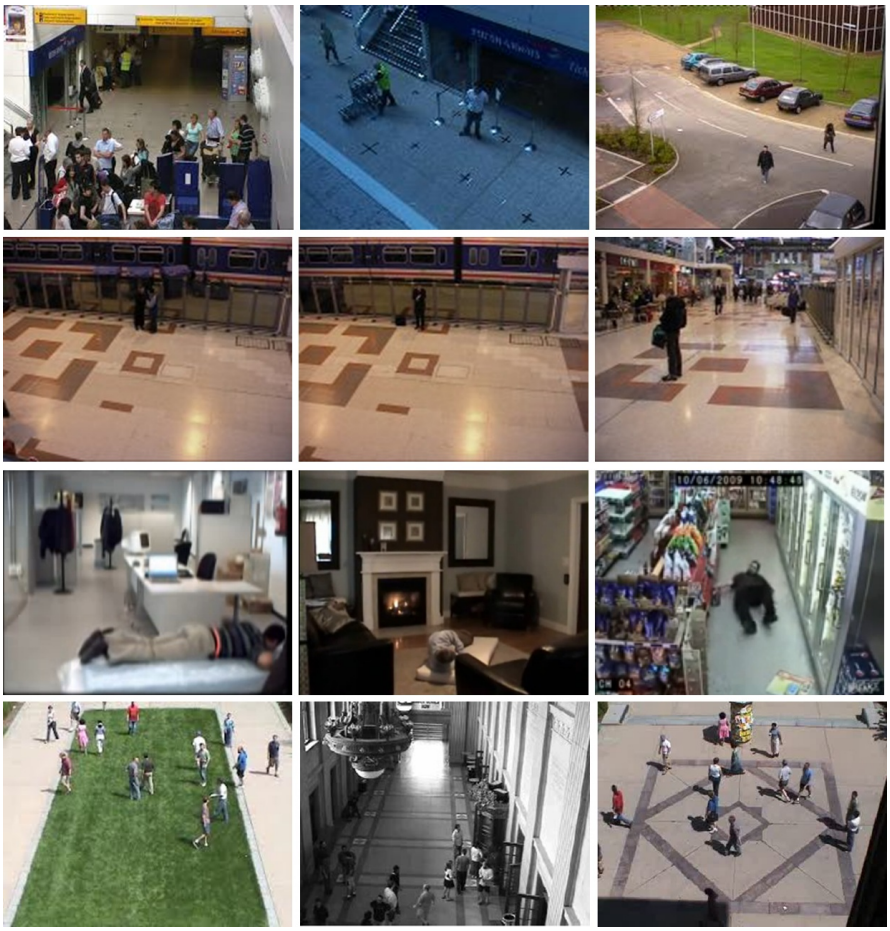


Figure 5.6: Sample benchmarked and public dataset used for evaluation.

5.5.2 Quantitative Result

The proposed framework is evaluated using three common metrics of measurements, including accuracy, detection rate and Positive Predictive Value (PPV).

[1]http://www.cvg.rdg.ac.uk/slides/pets.html

[2]http://www.multibel.be/cantata/

[3]http://mha.cs.umn.edu/movies/crowd-activity-all.avi

[4]http://www.youtube.com/watch?v=V8BLV4Wt3gA

[5]http://www.youtube.com/watch?v=09YrJcMhy9w

[6]http://www.youtube.com/watch?v=W8blP3DRyuM

Definition 5.5.1. ***A****ccuracy: Reflects the ability of the system to correctly identify both the actual abnormal event, and normal events from the population.*

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.28}$$

Definition 5.5.2. ***D****etection Rate: Reflects the ability of the framework to correctly identify actual abnormal event against all ground truth positive events.*

$$Detection\ Rate = \frac{TP}{TP + FN} \tag{5.29}$$

Definition 5.5.3. ***P****ositive Predictive Value (PPV): Reflects the ability of the system to correctly identify actual abnormal event against all positive detections.*

$$PPV = \frac{TP}{TP + FP} \tag{5.30}$$

where TP indicates that the framework detects an event correctly; TN indicates that the framework performs correct rejection, FP indicates the false alarm from the framework and FN indicates the miss detection from the framework. Fig. 5.7 illustrates the representation of TP, TN, FP and FN with respect to the ground truth detections. When available, the ground truths of the benchmarked dataset are used, otherwise they are manually annotated for evaluation; abnormal events are given a unique label from the normal events. The alarm to indicate abnormal event is triggered when any of the object conforms to the rules defined beforehand, for each of the ROI, while a non-event refers to the scenarios where the rules are not met. The collection of the test sequences used in these experiments comprise a fair distribution of both event and non-event scenarios.

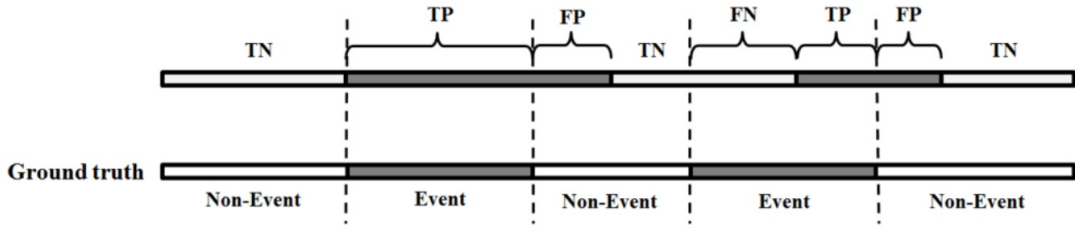


Figure 5.7: TP, TN, FP and FN are labelled with respect to the ground truth.

As shown in Table 5.1, the average accuracy for the five event detectors is 83.02%; average of detection rate is 87.82% and average PPV is 92.79%. Meanwhile the average time performance is 88.4 ms per frame. Meanwhile, Fig. 5.8 illustrates the graphical presentation of the results. The shaded region represents the acceptable detection accuracy, as adopted in most systems. It is observed that while the performance of the proposed framework is not the best in the market, it tends to fall either on the region deemed acceptable or surpasses the satisfactory accuracy as practised and adopted by state-of-the-arts (huperLab, 2014; Velastin et al., 2005; Khoudour et al., 1997; Schwerdt et al., 2005; Fernández-Caballero et al., 2012; VideoIQ & company, 2014).

While it is acknowledged that there are existing analytics solutions that are able to deliver better performance in terms of accuracy ($\geq 90\%$) and processing time ($\leq 60ms$ per frame), they are either concentrating on detecting one event at a particular time or do not deal with events which are similar to the ones recommended in this evaluation (Staff, 2012). Furthermore, the accuracy is often highly dependent on camera position, lighting, quality, and application scenario (Technology, 2012; Cisco Systems, 2014). Thus a fair comparison in terms of quantitative measurements against the existing systems is not applicable. Furthermore, the datasets used for most of the systems differ from one another, depending on their event detectors, leading to difficulty to perform a fair comparison. The acceptable detection rate and performance of analytics solutions have always been a constant debate between the systems' providers and consumers. Ideally, the consumers would like to have 100% accuracy of correct detection and 0% false alarms per camera

per day. However, no present technology makes this possible (Rozmus, 2012). This issue is made even more challenging with the fact that precise and unambiguous definition of the alarm condition is sometimes very difficult and thus also the classification of detections as false (positives or negatives) alarms is often challenging. Due to the ambiguity between a human perspective and automated systems, this often leads to debates whether an alarm triggered is correct or false. For example, according to i-LIDS sterile zone test the systems have 10 seconds to report an alarm state after an alarm event begins in the evaluation footage. During this time multiple alarm reports will be disregarded; an alarm event is either detected or not. After this 10 second window, any further alarms reported will be deemed 'false positives' (Scholz, Kawan, & Schindelhauer, 2012). Nevertheless experience in the industry shows that customers do not count alert as false alert when they occurred after 10 seconds window. An in depth discussion regarding this can be found by i-LIDS where a platform for evaluation is provided for analytics suppliers in (Branch, 2013; Crouwel, 2013). It is important to note that the exact values for detection accuracy and performance are confidential and not made public.

Table 5.1: Accuracy and performance measures for the five scenarios of abnormal events.

Event	Accuracy	Detection Rate	PPV	Time Performance
Loitering	83.33%	88.24%	93.75%	93.5ms
Intrusion	83.33%	90.00%	95.20%	84.5ms
Slip and Fall	88.46%	86.96%	95.00%	89.5ms
Abnormal Crowd Activity	80.00%	86.96%	90.00%	85.5ms
Unattended Object	80.00%	86.96%	90.00%	88.9ms

5.5.3 Comparison Result

Although the quantitative comparison between the proposed framework and existing systems are not available due to the reasons as discussed in the preceding section, this section provides a comprehensive comparison in terms of functionality. Table 5.2 summarises the functionality provided by the proposed compositional-based framework and state-of-the-art surveillance frameworks; CROMATICA in (Khoudour et al., 1997),

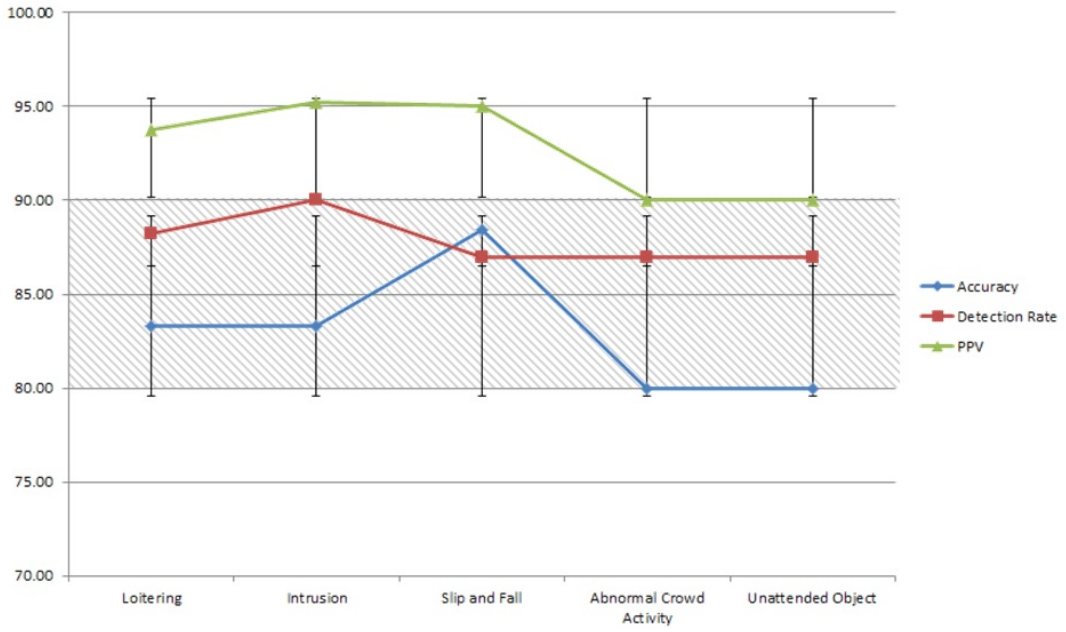


Figure 5.8: Accuracy measurement for the five scenarios of abnormal events.

PRISMATICA in (Velastin et al., 2005), Fuentes *et al.* in (Fuentes & Velastin, 2004), EAGLE in (Schwerdt et al., 2005), Black *et al.* in (Black et al., 2005), Fernández-Caballero *et al.* in (Fernández-Caballero et al., 2012), Axis in (Scholz et al., 2012), Bosch in (Bosch Security Systems, 2008), iOmniscient in (iOmniscient, n.d.) and finally VideoIQ in (VideoIQ & company, 2014).

The functionality is gauged in terms of the events that can be detected by each framework (e.g. loitering and intrusion), and the deployment scenarios (e.g. indoor and outdoor). The categorisation of the deployment scenarios is based on the sequences used for evaluation and is classified into 3 broad categories; indoor, outdoor, crowd (i.e. more than 10 persons). From Table 5.2, it is noticed that most of the solutions are able to detect the event of unattended object. This is probably due to rising concern of threat from terrorism that has been shown to be linked to the act of leaving objects comprising explosives unattended. Another common application is on detecting intrusion which triggers alert when a person enters a prohibited area or moves towards unauthorised direction. There are only a few suppliers that deal with fall detection which is extremely useful for smart

home systems to monitor the elderly or the disabled. The closest work which provides the most similar functions is iOmniscient in (iOmniscient, n.d.). Although iOmniscient provides crowd counting functionality, they do not deal with detection of crowd dispersal due to panic or evacuation. In summary, the proposed framework demonstrates sufficient detections of multiple events and is comparable to most of the state-of-the-art solutions, in terms of functionality. Although there are other functionalities such as crowd counting and face recognition which are provided by the existing solutions, they are not highlighted in the comparison table. The additional or sophisticated functionality is not in the scope of this study, and can be extended further for future investigations. It is emphasised that at this point of study, the main goal is to develop an integrated and optimised framework that allows detections of multiple events in different regions of the same scene. This is in contrast to other solutions that are tailored made to fit into detecting a particular event at one time. Nonetheless, the promising results of the proposed framework demonstrate the potentials of the proposed compositionality model in the domain of video surveillance.

Table 5.2: A comparison between our proposed compositional-based framework with state-of-the-art systems. The symbol '•' denotes available functions whereas blank columns indicate non availability.

Framework	**Event**					**Dataset**		
	Loitering	Intrusion	Slip and Fall	Abnormal Crowd Activity	Unattended Object	Indoor	Outdoor	Crowd
CROMATICA		•		•	•	•		•
PRISMATICA		•		•	•	•		•
Fuentes *et al.*		•	•		•	•		
EAGLE	•	•			•	•		
Black *et al.*		•	•		•	•		•
Axis		•				•	•	
Bosch	•	•			•	•	•	
iOmniscient	•	•	•		•	•	•	•
VideoIQ	•	•		•	•	•	•	•
Proposed Framework	•	•	•	•	•	•	•	•

5.5.4 Qualitative Results

This section presents the qualitative results for all five scenarios of abnormal events and discusses the evaluation settings comprehensively.

5.5.4 (a) Loitering Detection

The capability of the proposed system to detect loitering event is evaluated using the video sequences obtained from a combination of lab simulated dataset as well as the PETS2006 and PETS2007, respectively. In each video sequence, the ROI is defined as a restricted region, and the time span threshold to differentiate between long and short duration in which an object appears in the scene is defined in terms of number of frames. In this experiment, the threshold is set according to the ground truth and it varies from one sequence to another, depending on the staged loitering activity. For example, in the PETS2007 benchmarked sequence, the ground truth time span is fixed to 1500 frames (60sec x 25fps). In practice, the time span to determine loitering differs from one application to another, as well as on the deployment environment. In a critical area such as the entrance to an authorised building (i.e. bank), for instance, the time span to trigger loitering event is usually much shorter (i.e. 100 frames (4sec x 25fps)) as compared to the detector at the corridor of a shopping outlet, where the act of lingering is common.

Fig. 5.9 illustrates sample detections of loitering event. The loitering threshold is set to 200 frames (8sec x 25fps). The left image of Fig. 5.9 shows a loitering event occurring at frame 1097 from camera view 4, whereas on the right side shows a loitering event at frame 1219 from camera view 3.

Figure 5.9: The highlighted region denotes the ROI, and the red bounding box encloses the subject.

5.5.4 (b) Intrusion Detection

The capability of the proposed system to detect intrusion is evaluated using sequences obtained from a combination of lab simulated dataset as well as the PETS2006 dataset. Unlike the loitering event detector which detects human subject that lingers around the restricted region for a duration of time, intrusion detection module detects an event immediately (almost) upon the entry of a human subject into the restricted region. From Table 5.1, it is observed that the detection rate for intrusion event is better than loitering with an increase of 2%. This is due to the inconsistency of the tracking module in providing the same label for the same object over a prolonged period of time, which eventually leads to miss or late detection of loitering event.

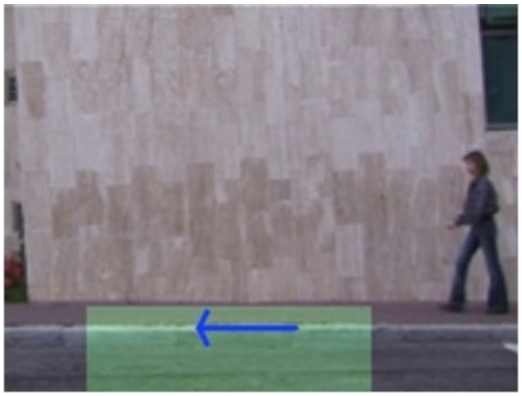
Fig. 5.10 illustrates sample outputs of the intrusion detection event. Accordingly, the restricted region is highlighted in green, and a person will be deemed as intruding a defined region if it appears in the region. Fig. 5.10a shows no detection as there is no object appearing in the restricted region. Fig. 5.10b triggers an alert when there is a person appearing in the ROI. The second sequence is used to evaluate the second scenario of intrusion, where the motion direction feature is incorporated for reasoning. The blue arrow in Fig. 5.10c denotes the prohibited motion direction, whereby any person in the ROI moving towards the left direction from the image space will trigger an alarm. Fig. 5.10d shows example detection of intrusion with direction.

5.5.4 (c) Slip and Fall Detection

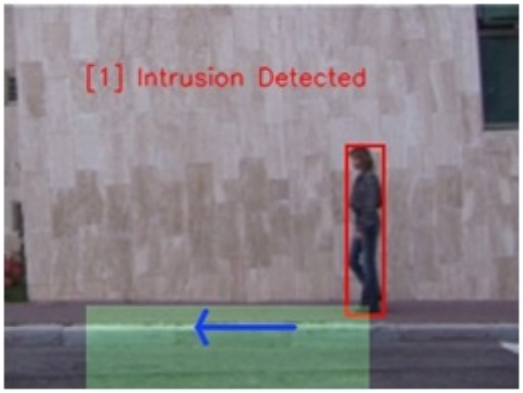
The capability of the proposed system to detect slip and fall events is evaluated using the video sequences obtained from a combination of lab simulated dataset as well as real-time dataset from the Youtube video. In this experiment, the histogram projection algorithm is applied, where for each direction x (rows) and y (columns), a 2D histograms are computed for each motion blob; H_x and H_y, respectively. In order to differentiate

(a) Original image. Highlighted region denotes the prohibited region, where entry is not allowed.

(b) Intrusion detected (regardless of the motion direction).



(c) Original image.



(d) Intrusion detected (unauthorised motion direction is incorporated).

Figure 5.10: Sample detections of intrusion event. Best viewed in color.

between an object who is standing and an object who has fallen down, the profiles are analysed. A standing posture has a dominant $H_y \geq H_x$, while the fall posture has a more dominant $H_x \geq H_y$ such as illustrated in Fig. 5.11. In addition to the profile change from the vertical distribution towards a horizontal distribution, the speed information is used to discriminate between the act of lying down and actual falling. Empirically, the act of lying down tend to have a slower motion speed as compared to fall. In this experiment, the sensitivity threshold is set to 0.7, empirically, and the window size is set to 10 frames. A sensitivity threshold of '1' indicates that the system is set at non-sensitive mode, where the human subject must be completely lying down over an interval of time in order to trigger the event; whereas a sensitivity threshold of '0' indicates greater sensitivity, where the slightest motion activity will trigger an alarm. The settings of the threshold and window

size are subjected to the test sequence, and the optimal values may vary from one sequence to another. The values recommended here are based on the defined dataset. Since the priority of this evaluation is on the capability of the compositional model in detecting multiple events under a unified framework, simple yet accurate techniques are used for reasoning. While it is acknowledged that there are existing works that have demonstrated sophisticated and reliable methods for robust slip and fall detection, they are beyond the scope of this study. Fig. 5.12 demonstrates sample detections of slip and fall events on public scenes.

(a) Sample stand profile, where, $H_y \geq H_x$. (b) Sample fall profile, where, $H_x \geq H_y$.

Figure 5.11: Illustrations on the profile feature between standing and fall posture.

5.5.4 (d) Abnormal Crowd Activity Detection

It is important to emphasise that the term abnormal crowd activity is broad and covers a wide applications of crowd anomalies including, crowd formation and crowd counting. In this context, it is defined as the act of sudden crowd dispersal. The sequences from PETS 2009 and UMN are used for evaluation. This module decomposes the scenario of crowd dispersal into layers of analysis using the compositional model. The spatio-temporal features which include the moving pixels obtained from the background subtraction stage, motion direction and magnitude from optical flow estimation are combined for reasoning. These features are aggregated into histograms and a window-based analysis is performed to monitor the motion changes within the window of frames. Basically, the proposed algorithm is influenced by two parameters; which are the time interval and

Figure 5.12: Example of detected fall events on four different scenes.

the normalised magnitude change. For evaluation, the empirical values are - the threshold of normalised magnitude differences is set as exceeding 200 when there is a sudden flow change and the time interval is set to 50.

It is important to point out that the evaluation at this stage does not consider the exact frame between the detection and occurrence of dispersal. This is in contrast to existing solutions which are focused at detecting crowd dispersal events in particular (Mehran et al., 2009; Thida, Eng, Dorothy, & Remagnino, 2011). Instead, the rule of thumb as recommended by i-LIDS is adopted, where an alert is classified as false if it happens 10 seconds after an event has occurred (Scholz et al., 2012). Fig. 5.13a - 5.13d present sample detections of sudden crowd dispersal event from camera view 1, 2, 3 and 4 using the UMN sequences respectively. Fig. 5.13e - 5.13g show detections of sudden crowd dispersal event for 3 different indoor and outdoor scenes from the UMN dataset.

5.5.4 (e) Unattended Object Detection

Fig. 5.14a - 5.14b illustrate two different scenarios of staged unattended object events. The first scenario simulates an unattended object scene, where the car is parked and in a way left unattended in the scene. This function is useful in various traffic related applications such as unauthorised parking detection, illegal stop or vehicle breakdown detections. The second scenario depicts an actual unattended luggage event, where a person enters the scene before leaving behind a bag. Based on the proposed framework, and the set of notions for reasoning, an object is deemed unattended if it is left static in the scene for more than 1 minute and belongs to the 'vehicle' or 'luggage' object class.

5.5.4 (f) Intrusion and Loitering Detection

Fig. 5.15a - 5.15b show sample outputs of multiple event detection. In this sequence, which is obtained from the PETS 2007, sequence 0, view 3, the individual appears to be in the region-of-interest at frame 152. At frame 160, the proposed framework detected an intrusion event. At frame 610, a loitering event is detected. This is due to the time span of the individual, lingering within the defined region-of-interest for a time span of 450 frames (equivalent to 18 seconds, assuming that the frame rate is 25fps). It is highlighted that the time threshold to determine if an individual is exhibiting intrusion or loitering behaviour is subjected to the dataset used.

(a) PETS 2009 camera 1.

(b) PETS 2009 camera 2.

(c) PETS 2009 camera 3.

(d) PETS 2009 camera 4.

(e) UMN sequence 1.

(f) UMN sequence 2.

(g) UMN sequence 3.

Figure 5.13: Example of the sudden crowd dispersal events.

(a) CANTATA sequence, where a car is parked (left unattended).

(b) PETS 2006 sequence, where a luggage is left unattended in the ROI.

Figure 5.14: Example of the detected unattended object.

(a) Intrusion event is triggered.

(b) Loitering event is detected.

Figure 5.15: Example outputs of multiple events, where the scenario depicts an individual intruding and lingering within the region-of-interest (highlighted in green). An intrusion event is deteced at frame 160, of the sequence, dataset, and at frame 630, a loitering event is triggered.

5.6 Summary

This chapter presents a framework for multiple event detections in surveillance videos. Based on the principle of compositionality, the surveillance problem is modularised into a set of sub-problems to allow flexibility and ease of fine-tuning for scalability, to include other real-time events. In order to demonstrate the functionality of the knowledge constructed based on the proposed concept of compositionality, comprehensive experiments using 100 videos obtained from the selected benchmark dataset (PETS, UMN and CANTATA), as well as real-time public videos obtained from the Youtube are used. Experiment results and a comparison, in terms of functionality, with the state-of-the-art solutions have shown the efficiency of the proposed framework in detecting multiple events efficiently. One of the drawback of this work, at present, is the lack of testing data to further evaluate the robustness of the proposed system. Therefore, amongst the future work of this study is to collect and built a wider selection of dataset for benchmarking purpose within the research community. Although the necessity for an independent evaluation of such capabilities becomes more and more prominent as the capabilities advertised by commercial analytics providers increase, the implementation of a unified performance framework for benchmarking is not as straightforward. Currently, there are no published efforts in the literature or independent data that can sustain the claims of existing analytics providers (Goldgof, Sapper, Candamo, & Shreve, 2009). Nonetheless, other work includes an extension of the proposed framework to include more complex representations of the variables (i.e. classes, attributes and notions) to deal more challenging real-time scenarios, as well as the development of new domain knowledge (e.g. airport).

CHAPTER 6

CONCLUSION

This thesis has set off to explore the prospect of devising computer vision algorithms for activity understanding and abnormal event detection in video surveillance. Specifically, the thesis is driven towards solving the three main issues in conjunction with the three main trends, progressing towards proactive video surveillance as discussed in Chapter 2.1. The first issue is on providing a robust visual tracking algorithm that deals with abrupt motion. The second is to identify salient regions, which could ultimately lead to unfavourable events in dense crowd scenes. Finally, the third aims to provide an integrated framework to detect multiple events in different regions-of-interest of a given scene. These problems are non trivial towards effective and proactive implementation of video surveillance. Activity understanding and abnormal event detection becomes even more challenging with the enormous growth in the number of CCTVs deployed nowadays. Visual ambiguity, clutter and occlusion, rarity and unpredictability of abnormal events, owing to the diversity of human behaviour, complexity of the environment and massive number of CCTVs remain an issue. The overview of the current state and the gap towards solving the three main issues in this thesis are as discussed in Chapter 2.

6.1 Tracking Abrupt Motion

Chapter 3 has presented a novel swarm intelligence-based tracker for visual tracking that deals with abrupt motion efficiently. Specifically, the proposed method explores a new direction by considering motion estimation or tracking as an optimisation problem. The proposed SwATrack has the advantage of optimising the search for the optimal distribution without making assumptions or learning the motion model before-hand. In

addition, the experience-sharing between particles is exploited and further refined with an adaptive mechanism to allow self-tuning of the parameters. This allows accurate tracking while keeping the number of samples at its minimal, thus resulting in lower processing cost with comparable tracking accuracy. Experimental results have shown that the proposed SwATrack improves the accuracy of tracking abrupt motion, with an average accuracy of **91.39%**, while significantly reduces the computational overheads, with an average processing time of **63** milliseconds per frame. On top of that, findings from this study has defied the common understanding in most sampling-based tracking approach such as the PF, where an increase in the number of particles in PF has been shown to not directly increase the tracking efficiency. The preliminary results at this stage, create prospects for a new paradigm in object tracking which is very highly motivated by the notion proposed in (Zhu et al., 2012; Cifuentes et al., 2012); i) *Will continued progress in visual tracking be driven by the increased complexity of tracking algorithms?* ii) *How far should the increased in complexity be, since motion models only work sometimes?*

There are several aspects of this work for possible extensions. Firstly, the proposed SwATrack has not been evaluated to *track multiple objects* in crowded scenarios. A broad understanding of visual tracking leads to intuitive suggestion that this extension will require data association techniques such as the known Joint Probability Data Association Filtering (JPDAF) and Multiple Hypothesis Tracking (MHT) to address the issue of correspondence. The second aspect is on *resolving occlusion*. Occlusion has been a constant issue in most surveillance scenarios and is amplified when tracking multiple objects. A multi-swarm approach which incorporates data association techniques would be a potential room for future exploration. Also, the experience-sharing advantage of swarm intelligence can be further exploited to model the interactions between multiple objects. Finally, *contextual information* can be included to attune the tracker to particular deployment scenarios. The accumulated trajectories in a given scene can be used to

self-imposed constraints in the motion model of objects and vice versa to allow on-going self-adjustment of tracking mechanism with the environment.

6.2 Behaviour Analysis in Crowd

In order to reduce the cognitive overload in CCTV monitoring, it is critical to have an automated way to direct the attention of operators on potentially precarious regions acquiring attention, especially in crowded scenes. Detection of interesting regions in crowded scene is difficult to be perceived by the naked eyes, due to the large variations of crowd densities and occlusions. Therefore, chapter 4 of this thesis discussed the implementation of two frameworks for salient region detection in the crowded scenes, where one is the extension of the other. The proposed extension is to allow detection of subtle motion change caused by local irregular motion. The proposed framework eliminates the need to track each object individually, prior information or extensive learning to identify anomalies by observing the flow activities in a given scene for inference. In addition, the projection of low-level motion flow into the global similarity structure to characterize stability and phase changes has been shown to be an effective indicator of high motion dynamics and irregularities in the crowded scenes. Experimental results have shown the potential of the proposed framework in detecting obvious and subtle motion change caused by instability, bottleneck, or occlusion, and also local irregular motion, with an average accuracy of **78%** on the defined dataset.

At present, research in crowd behaviour analysis lacks standard dataset and performance framework for benchmarking purpose. At its current state, most solutions are evaluated on partially overlapping sequences due to the differing causes of salient regions to be detected. Furthermore, it is often very difficult to determine the ground truth information, as the salient region such as bottleneck in crowded scenes is often very subjective and not easily perceived by the human eye. Thus, another aspect of the future investiga-

tion is to prepare *comprehensive datasets and performance measurement framework for benchmarking*. The current framework suppresses the dominant motion flow and narrow down the analysis to regions with high motion dynamics to infer salient region. It would be interesting to *include the dominant flow summarised over a time period* for refined salient detection. In practice, a system would be more user friendly and efficient if it able to accept input or feedback from the environment, or user for active learning. Another aspect for future exploration is to *include feedback from user* in the learning mechanism to reduce false alarms and improve true detection. Also, further analysis on the detected salient regions *to infer higher level descriptions* would be practical. Finally, the *feature representation can be enriched by including additional feature* such as projection of the texture feature.

6.3 Multiple Events Detection

Chapter 5 presented a new direction in event detection by modularising analytics into a set of sub-problems. The decomposition method is inspired by the principle of compositionality, where the knowledge of video analytics is decomposed into low-level descriptions which are then integrated and combined using a basic set of rule-packages to infer various events. The compositional representation of the contextual information of a given scene allows simple and optimised detections of multiple events. This is in contrast to conventional analytics frameworks, where redundant low-level processing is required to detect multiple events; since each event detection modules operate individually. Experiment results have demonstrated the capability of the integrated framework to detect five different scenarios of abnormality, with an average accuracy of **83%**. At this stage, the five scenarios include intrusion, loitering, slip and fall, crowd dispersal and abandoned object detection.

Video analytics solutions aim to fully complement and enhance the existing security

infrastructure to provide defence against, as well as proactive understanding of, security vulnerabilities. Such capabilities as advertised by commercial analytics providers have been increasing tremendously to meet the increasing demand. However, there are no published efforts in the literature or independent data that can sustain the claims of existing analytics providers (Goldgof et al., 2009) and most of these information are kept confidential. Thus amongst the future work in this aspect is to *collect and built a better and wider selection of dataset for benchmarking purpose* within the research community, or even industry. Furthermore, there is a need for an independent evaluation of such capabilities or functionalities in order for video analytics to be widely accepted by the public, as well as industry. The current framework is designed to deal with five scenarios of abnormality and can be easily extended *to handle a wider range of events* such as object removal or people counting. Although the presented framework is flexible enough to deal with more scenarios, comprehensive and thorough investigation is required to further advanced the capabilities of current framework to deal with more complex scenarios such as aggression detection. The leap forward for analytics solutions is to deal with early detections of behaviours that could possible prevents criminal incidents. The *fusion of data from multiple sensors* (e.g. multiple cameras, audio and infrared) is another interesting area that would enrich the functionality of video analytics. Other future work includes the extension of the proposed framework to include *complex representations of the variables*, (i.e. classes, attributes and notions) to deal with the complexity of various real-time scenarios, as well as the *development of new domain knowledge* (i.e. underground station and airport).

6.4 Summary

Activity understanding and abnormal event detection have improved dramatically in recent years. A significant change has been observed in high level of accuracy that video analytics systems can perform, and the increasing number of tasks or functionali-

ties they can accomplish. However, given the marketing and advertising gimmicks nowadays, there is misconception on what the current technology can achieved. Amongst the hypes includes, “analytics can detect terrorist walking along the street” or “analytics can identify an offender out from a sea of faces in the mall”. It is important to understand that video analytics are still very much at their infancy, and should not be a technology that is “over-promised” and “under-delivered”. The utmost importance is continuous research efforts in this domain towards achieving the ideal goal of complementing the existing security infrastructure, in the hope of creating a better and safer society. This thesis has set off to contribute to such goal, by introducing growing recognition capabilities and better analytics solutions for the advancement of video surveillance, from the domain of computer vision in particular.

Appendix

APPENDIX A

PUBLICATIONS

Chapter 3: Tracking Abrupt Motion

Conference

M. K. Lim, C. S. Chan, D. Monekosso and P. Remagnino (2013) SwATrack: A Swarm Intelligence-based Tracking of Abrupt Motion, IAPR International Conference on Machine Vision Applications, pp. 37-40.

M. K. Lim, C. S. Chan, D. Monekosso and P. Remagnino (2013) SwATrack: A Swarm Intelligence-based Tracking of Abrupt Motion, IEEE workshop on Visual Object Tracking Challenge, pp. 1-14.

Journal

M. K. Lim, C. S. Chan, D. Monekosso and P. Remagnino (2014) Refined Particle Swarm Intelligence Method for Abrupt Motion Tracking", Information Sciences.

Chapter 4: Crowd Behaviour Analysis

Conference

M.K. Lim, V. J. Kok, C. C. Loy and C. S. Chan (2014) Identifying Anomalies in Crowded Scenes via Global Similarity Structure, IAPR International Conference on Pattern Recognition.

Journal

M. K. Lim, C. S. Chan, D. Monekosso and P. Remagnino (2014) Detection of Salient Regions in Crowded Scenes, IET Electronics Letters, pp. 363-365.

Chapter 5: Multiple Events Detection

Journal

M. K. Lim , S. Tang and C. S. Chan (2014) iSurveillance: Intelligent Framework for Multiple Events Detection in Surveillance Videos, Expert Systems and Applications, 41 (10), pp. 4704-4715.

REFERENCES

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

Ahmed, H., & Glasgow, J. (2012). *Swarm intelligence: Concepts, models and applications* (Tech. Rep.). Queen's University.

Alamri, Y. A. (2014). *Emergency management in saudi arabia: Past, present and future.* Retrieved from http://training.fema.gov/EMIWeb/edu/Comparative%20EM%20Book%20-%20EM%20in%20Saudi%20Arabia.pdf

Albusac, J., Castro-Schez, J. J., Vallejo, D., Jiménez, L., & Glez-Morcillo, C. (2011). A scalable approach based on normality components for intelligent surveillance. In P. Remagnino, D. N. Monekosso, & L. Jain (Eds.), *Innovations in defence support systems* (Vol. 336, pp. 105–145). Springer.

Albusac, J., Vallejo, D., Castro-Schez, J. J., Glez-Morcillo, C., & Jiménez, L. (2014). Dynamic weighted aggregation for normality analysis in intelligent surveillance systems. *Expert Systems with Applications*, *41*(4), 2008–2022.

Ali, S., & Shah, M. (2007). A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Computer vision and pattern recognition* (p. 1-6).

Ali, S., & Shah, M. (2008). Floor fields for tracking in high density crowd scenes. In *European conference on computer vision* (Vol. 5303, pp. 1–14).

Aliakbarpour, H., Khoshhal, K., Quintas, J., Mekhnacha, K., Ros, J., Andersson, M., & Dias, J. (2011). Hmm-based abnormal behaviour detection using heterogeneous sensor network. In L. Camarinha-Matos (Ed.), *Technological innovation for sustainability* (Vol. 349, p. 277-285). Springer.

Anderson, J., & McAtamney, A. (2011). Considering local context when evaluating a closed circuit television system in public spaces. *Trends and Issues in Crime and Criminal Justice*, *430*, 1-10. Retrieved from http://www.aic.gov.au/media_library/publications/tandi_pdf/tandi430.pdf

Andrade, E. L., Blunsden, S., & Fisher, R. B. (2006). Modelling crowd scenes for event detection. In *International conference on pattern recognition* (Vol. 1, p. 175-178).

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *Signal Processing*, *50*(2), 174–188.

Benezeth, Y., Jodoin, P., & Saligrama, V. (2011). Modeling patterns of activity and detecting abnormal events with low-level co-occurrences. In B. Bhanu, C. V. Ravishankar, A. K. Roy-Chowdhury, H. Aghajan, & D. Terzopoulos (Eds.), *Distributed video sensor networks* (p. 137-148). Springer.

Bird, N. D., Masoud, O., Papanikolopoulos, N. P., & Isaacs, A. (2005). Detection of loitering individuals in public transportation areas. *Transactions on Intelligent Transportation Systems*, *6*(2), 167–177.

Black, J., Velastin, S., & Boghossian, B. (2005). A real time surveillance system for metropolitan railways. In *Advanced video and signal based surveillance* (pp. 189–194).

Bosch Security Systems, I. (2008). *Focus your attention: Bosch intelligent video analysis (iva).* Retrieved from http://st-nso-us.resource.bosch.com/media/en/us_product_test/04_customer_service_1/02_contact_8/doc_25/commercial_brochure_enus_1558886539_iva.pdf

Bozdogan, A. O., & Efe, M. (2011). Improved assignment with ant colony optimization for multi-target tracking. *Expert Systems with Applications*, *38*(8), 9172 - 9178.

Branch, H. O. S. D. (2013, March). *Imagery library for intelligent detection systems (i-LIDS).* Retrieved from https://www.gov.uk/imagery-library-for-intelligent-detection-systems

Bruckner, D., Picus, C., Velik, R., Herzner, W., & Zucker, G. (2012). Hierarchical semantic processing architecture for smart sensors in surveillance networks. *Transactions on Industrial Informatics*, *8*(2), 291–301.

Buxton, H. (2003). Learning and understanding dynamic scene activity: a review. *Image and Vision Computing*, *21*(1), 125-136.

Cancela, B., Ortega, M., Fernández, A., & Penedo, M. G. (2013). Hierarchical framework for robust and fast multiple-target tracking in surveillance scenarios. *Expert Systems with Applications*, *40*(4), 1116 - 1131.

Challenger, R., Clegg, C. W., & Robinson, M. A. (2010). *Understanding crowd behaviors: Guidance and lessons identified* (Tech. Rep.). University of Leeds.

Chan, C. S. (2008). *Fuzzy qualitative human motion analysis* (Unpublished doctoral dissertation). University of Portsmouth.

Chan, C. S., & Liu, H. (2009). Fuzzy qualitative human motion analysis. *Transactions on Fuzzy Systems*, *17*(4), 851-862.

Chan, C. S., Liu, H., David, B., & Kubota, N. (2008). A fuzzy qualitative approach to human motion recognition. In *International conference on fuzzy systems* (p. 1242-1249).

Chen, Y.-L., Wu, B.-F., Huang, H.-Y., & Fan, C.-J. (2011). A real-time vision system for nighttime vehicle detection and traffic surveillance. *Transactions on Industrial Electronics*, *58*, 2030–2044.

Cifuentes, C. G., Sturzel, M., Jurie, F., & Brostow, G. J. (2012). Motion models that only work sometimes. In *British machine vision conference* (p. 1-12).

Cisco Systems, I. (2014, May). Cisco video analytics user guide [Computer software manual].

Clerc, M., & Kennedy, J. (2002). The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *Transactions on Evolutionary Computation*, *6*(1), 58–73.

Comaniciu, D., Ramesh, V., & Meer, P. (2003). Kernel-based object tracking. *Transactions on Pattern Analysis and Machine Intelligence*, *25*(5), 564–575.

Crouwel, F. (2013, August). *How useful is i-lids?* Professional Security Magazine. Retrieved from `http://www.professionalsecurity.co.uk/news/interviews/how-useful-is-i-lids/`

Dadashi, N. (2008). *Automatic surveillance and cctv operator workload* (Unpublished master's thesis). University of Nottingham.

Darby, M., P.and Johnes, & Mellor, G. (2005). *Soccer and disaster*. Psychology Press.

Davies, A., & Velastin, S. (2005). A progress review of intelligent cctv surveillance systems. *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, 417–423.

Dee, H. M., & Velastin, S. (2008). How close are we to solving the problem of automated visual surveillance?: A review of real-world surveillance, scientific progress and evaluative mechanisms. *Machine Vision and Applications*, *19*(5-6), 329–343.

Dick, A. R., & Brooks, M. J. (2004). A stochastic approach to tracking objects across multiple cameras. In G. Webb & X. Yu (Eds.), *Advances in artificial intelligence* (Vol. 3339, pp. 160–170).

Dore, A., Soto, M., & Regazzoni, C. S. (2010). Bayesian tracking for video analytics. *Signal Processing Magazine*, *27*(5), 46–55.

Doucet, A., & Johansen, A. M. (2011). A tutorial on particle filtering and smoothing: fifteen years later. In *Oxford handbook of nonlinear filtering* (pp. 656–704).

Draganjac, I., Kovacic, Z., Ujlaki, D., & Mikulic, J. (2008). Dual camera surveillance system for control and alarm generation in security applications. In *International symposium on industrial electronics* (p. 1070-1075).

Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. In *International symposium on micro machine and human science* (p. 39-43).

Ellis, L., Dowson, N., Matas, J., & Bowden, R. (2011). Linear regression and adaptive appearance models for fast simultaneous modelling and tracking. *International Journal of Computer Vision*, *95*(2), 154–179.

Epitropakis, M. G., Plagianakos, V. P., & Vrahatis, M. N. (2012). Evolving cognitive and social experience in particle swarm optimization through differential evolution: A hybrid approach. *Information Sciences*, *216*, 50-92.

Everingham, M., Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fernández-Caballero, A., Castillo, J. C., & Rodríguez-Sánchez, J. M. (2012). Human activity monitoring by local and global finite state machines. *Expert Systems with Applications*, *39*(8), 6982–6993.

Fookes, C., Denman, S., Lakemond, R., Ryan, D., Sridharan, S., & Piccardi, M. (2010). Semi-supervised intelligent surveillance system for secure environments. In *International symposium on industrial electronics* (pp. 2815–2820).

Frost, & Sullivan. (2009, Dec). Malaysia cctv / e-surveillance [Computer software manual]. U.S.A..

Frost, & Sullivan. (2012). *Safer cities - an inevitable trend in urband development* (Tech. Rep.). Author. Retrieved from `http://www.nec.com/en/global/solutions/safety/pdf/Safer_Cities_WP.pdf`

Fuentes, L. M., & Velastin, S. (2004). Tracking-based event detection for cctv systems. *Pattern Analysis and Applications*, *7*(4), 356–364.

Fullerton, E., & Kannov, S. E. (2008). *Keeping watch on the city: Ip video surveillance* (Tech. Rep.). Denmark: Milestone System Incorporation.

Gad-el Hak, M. (2008). *Large-scale disasters: Prediction, control, and mitigation* (M. Gad-el Hak, Ed.). Cambridge University Press.

Georis, B., Bremond, F., & Thonnat, M. (2007). Real-time control of video surveillance systems with program supervision techniques. *Machine Vision and Applications*, *18*(3-4), 189–205.

Goldgof, D. B., Sapper, D., Candamo, J., & Shreve, M. (2009). *Evaluation of smart video for transit event detection* (Tech. Rep.). University of South Florida.

Gong, S., Loy, C. C., & Xiang, T. (2011). Security and surveillance. In T. B. Moeslund, A. Hilton, V. Krüger, & L. Sigal (Eds.), *Visual analysis of humans* (p. 455-472). Springer.

Gong, S., & Xiang, T. (2011). *Visual analysis of behaviour: From pixels to semantics* (S. Gong & T. Xiang, Eds.). Springer.

Gouaillier, V., & Fleurant, A.-E. (2009). *Intelligent video surveillance: Promises and challenges.* (Tech. Rep.). Canada: RIM and Techn^opole Defence and Security.

Gustafsson, F., Gunnarsson, F., Bergman, N., Forssell, U., Jansson, J., Karlsson, R., & Nordlund, P.-J. (2002). Particle filters for positioning, navigation, and tracking. *Transactions on Signal Processing*, *50*(2), 425-437.

Haller, G. (2000). Finding finite-time invariant manifolds in two-dimensional velocity fields. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *10*(99), 99–108.

Helbing, D., & Mukerji, P. (2012). Crowd disasters as systemic failures: Analysis of the love parade disaster. *EPJ Data Science*, *1*(1), 1-40.

Hernon, J. (2003). Cctv: Constant cameras track violators. *National Institute of Justice Journal*, *249*, 16-23.

Höferlin, M., Höferlin, B., Weiskopf, D., & Heidemann, G. (2011). Uncertainty-aware video visual analytics of tracked moving objects. *Journal of Spatial Information Science*, *2*(1), 87-117.

Holmes, S. A., Klein, G., & Murray, D. W. (2009). An o(n) square root unscented kalman filter for visual simultaneous localization and mapping. *Pattern Analysis and Machine Intelligence*, *31*(7), 1251-1263.

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*(1-3), 185–203.

Hospedales, T. M., Li, J., Gong, S., & Xiang, T. (2011). Identifying rare and subtle behaviors: A weakly supervised joint topic model. *Pattern Analysis and Machine Intelligence*, *33*(12), 2451–2464.

Hsu, E. B., & Burkle, F. M. J. (2012). Cambodian bon om touk stampede highlights preventable tragedy. *Prehospital and Disaster Medicine*, *27*, 481–482.

Hu, D. H., Zhang, J., X.-X.and Yin, Zheng, V. W., & Yang, Q. (2009). Abnormal activity recognition based on hdp-hmm models. In C. Boutilier (Ed.), *International joint conference on artifical intelligence* (p. 1715-1720).

Hu, M., Ali, S., & Shah, M. (2008). Learning motion patterns in crowded scenes using motion flow field. In *International conference on pattern recognition* (p. 1-5).

huperLab. (2014). *hupervision 2d & 3d video analytics overview.* Retrieved from `http://www.huperlab.com/english/product/HV/keyFeature_05.htm`

iOmniscient. (n.d.). *iomniscient detection.* Retrieved from `http://iomniscient.com/index.php?option=com_content&view=article&id=56&Itemid=53`

Isard, M., & Blake, A. (1998). Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, *29*(1), 5–28.

Jacques, J., Jung, C., & Raupp Musse, S. (2005). Background subtraction and shadow detection in grayscale video sequences. In *Brazilian symposium on computer graphics and image processing* (p. 189-196).

Jacques Junior, J., Raupp Musse, S., & Jung, C. R. (2010). Crowd analysis using computer vision techniques. *Signal Processing Magazine*, *27*(5), 66-77.

Jäger, M., Knoll, C., & Hamprecht, F. A. (2008). Weakly supervised learning of a classifier for unusual event detection. *Transactions on Image Processing*, *17*(9), 1700-1708.

James, J. J., Benjamin, G. C., Burkle, F. M. J., Gebbie, K. M., Kelen, G., & Subbarao, I. (2010). Disaster medicine and public health preparedness: A discipline for all health professionals. *Disaster Medicine and Public Health Preparedness*, *4*(2), 102–107.

Janssen, T. M. V. (2001). Frege, contextuality and compositionality. *Journal of Logic, Language and Information*, *10*(1), 115-136.

Jiang, F., Wu, Y., & Katsaggelos, A. K. (2009). A dynamic hierarchical clustering method for trajectory-based unusual video event detection. *IEEE*, *18*(4), 907-913.

Jo, H., Chug, K., & Sethi, R. J. (2013). A review of physics-based methods for group and crowd analysis in computer vision. *Journal of Postdoctoral Research*, *1*(1), 56-59.

Jodoin, P.-M., Konrad, J., & Saligrama, V. (2008). Modeling background activity for behavior subtraction.

In *International conference on distributed smart cameras* (p. 1-10).

Julier, S. J., & Uhlmann, J. K. (1997). A new extension of the kalman filter to nonlinear systems. In *Signal processing, sensor fusion, and target recognition vi* (Vol. 182).

Karasulu, B., & Korukoglu, S. (2012). Moving object detection and tracking by using annealed background subtraction method in videos: Performance optimization. *Expert Systems with Applications*, *39*(1), 33 - 43.

Kennedy, C. A., & Carpenter, M. H. (2003). Additive runge-kutta schemes for convection-diffusion-reaction equations. *Applied Numerical Mathematics*, *44*(1-2), 139–181.

Keval, H., & Sasse, M. (2008). "not the usual suspects": A study of factors reducing the effectivenss of cctv. *Security Journal*(2), 1-21.

Keval, H., & Sasse, M. A. (2006). Man or gorilla? performance issues with cctv technology in security control rooms. *International Ergonomics Association*.

Khoudour, L., Deparis, J.-P., Bruyelle, J.-L., Cabestaing, F., Aubert, D., Bouchafa, S., ... Wherett, M. (1997). Project cromatica. In *Image analysis and processing* (pp. 757–764).

Klontz, J. C., & Jain, A. K. (2013). *A case study on unconstrained facial recognition using the boston marathon bombings suspects* (Tech. Rep. No. MSU-CSE-13-4). Michigan: Michigan State University.

Krahnstoever, N. (2011). *Video analysis/analytics: Can we use it to detect criminal behaviors and activities?* (Tech. Rep.). General Electric (GE) Global Research.

Kratz, L., & Nishino, K. (2010). Tracking with local spatio-temporal motion patterns in extremely crowded scenes. *Computer Vision and Pattern Recognition*, 693-700.

Krausz, B. (2012). *Detection and simulation of dangerous human crowd behavior* (Unpublished doctoral dissertation). University of Bonn.

Kuettel, D., Breitenstein, M. D., Van Gool, L., & Ferrari, V. (2010). What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *Computer vision and pattern recognition* (p. 1951-1958).

Kwon, J., & Lee, K. M. (2008). Tracking of abrupt motion using wang-landau monte carlo estimation. In *European conference on computer vision* (Vol. 1, pp. 387–400).

Kwon, J., & Lee, K. M. (2010). Visual tracking decomposition. In *Computer vision and pattern recognition* (pp. 1269–1276).

Kwon, J., & Lee, K. M. (2013). Wang-landau monte carlo-based tracking methods for abrupt motions. *Pattern Analysis and Machine Intelligence*, *35*(4), 1011–1024.

Lavee, G., Khan, L., & Thuraisingham, B. (2007). A framework for a video analysis tool for suspicious event detection. *Multimedia Tools and Applications*, *35*(1), 109–123.

La Vigne, N. G., Lowry, S. S., Markman, J. A., & Dwyer, A. M. (2011). *Evaluating the use of public surveillance cameras for crime control and preventiona* (Tech. Rep.). The Office of Community Oriented Policing Services (US Department of Justice). Retrieved from `http://www.cops.usdoj.gov/Publications/e071112381_EvalPublicSurveillance.pdf`

Lee, C.-K., Ho, M.-F., Wen, W.-S., & Huang, C.-L. (2006). Abnormal event detection in video using n-cut clustering. In *International conference on intelligent information hiding and multimedia signal processing* (p. 407-410).

Lee, M. (2012). *A literature review of emergency and non-emergency events*. Fire Protection Research Foundation. Retrieved from `http://books.google.com.my/books?id=T7t1kgEACAAJ`

Lee, M. L. (2014, January). *A parent's worst fear.* Retrieved from http://www.thestar.com.my/Lifestyle/Family/Features/2014/01/17/A-parents-worst-fear-for-child-safety/

Lerner, A., Chrysanthou, Y., & Lischinski, D. (2007). Crowds by example. *Computer Graphics Forum*, *26*(3), 655-664.

Li, N., & Zhang, Z. (2011). Abnormal crowd behavior detection using topological methods. In *Software engineering, artificial intelligence, networking and parallel/distributed computing* (p. 13-18). doi: 10.1109/SNPD.2011.21

Li, W., Zhang, X., & Hu, W. (2009). Contour tracking with abrupt motion. In *International conference on image processing* (pp. 3593–3596).

Li, X., Wang, K., Wang, W., & Li, Y. (2010). A multiple object tracking method using kalman filter. In *International conference on information and automation* (p. 1862-1866). doi: 10.1109/ICINFA.2010.5512258

Li, Y., Ai, H., Yamashita, T., Lao, S., & Kawade, M. (2008). Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans. *Pattern Analysis and Machine Intelligence*, *30*(10), 1728-1740.

Lin, J. (2014). *Advanced digital security & surveillance solution* (Tech. Rep.). Intel Corporation. Retrieved from http://www.in2ftp.com/intelembedded/Digital%20Security/Malaysia_Embedded_Event.pdf

Lin, W., Sun, M.-T., Poovendran, R., & Zhang, Z. (2008). Human activity recognition for video surveillance. In *International symposium circuits and systems* (p. 2737-2740).

Liu, C., Freeman, W. T., Adelson, E. H., & Weiss, Y. (2008). Human-assisted motion annotation. In *Computer vision and pattern recognition* (pp. 1–8).

Liu, G., Tang, X., Huang, J., Liu, J., & Sun, D. (2007). Hierarchical model-based human motion tracking via unscented kalman filter. In *International conference on computer vision,* (p. 1-8).

Liu, H., & Sun, F. (2012). Efficient visual tracking using particle filter with incremental likelihood calculation. *Information Sciences*, *195*, 141-153.

Liu, N.-H., Chiang, C.-Y., & Chu, H.-C. (2013). Recognizing the degree of human attention using eeg signals from mobile sensors. *Sensors*, *13*(8), 10273–10286.

Ljung, L. (1979). Asymptotic behavior of the extended kalman filter as a parameter estimator for linear systems. *Transactions on Automatic Control*, *24*(1), 36-50.

Loy, C. C., Xiang, T., & Shaogang, G. (2012). Salient motion detection in crowded scenes. In *International symposium on communications control and signal processing* (p. 1-4).

Maggio, E., & Cavallaro, A. (2009). Accurate appearance-based bayesian tracking for maneuvering targets. *Computer Vision and Image Understanding*, *113*(4), 544–555.

Makris, D., & Ellis, T. (2005). Learning semantic scene models from observing activity in visual surveillance. *Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *35*(3), 397–408.

Malone, S., & McCool, G. (2013, April). *Boston officials say 264 injured in marathon bombing.* Retrieved from http://wwww.reuters.com/article/2013/04/16/us-usa-explosions-boston-injuries-idUSBRE93F10X20130416

Mancas, M., Riche, N., Leroy, J., & Gosselin, B. (2011). Abnormal motion selection in crowds using bottom-up saliency. In *International conference on image processing* (p. 229-232).

Maxion, R. A., & Tan, K. M. C. (2000). Benchmarking anomaly-based detection systems. In *International conference on dependable systems and networks* (p. 623-630).

Mehran, R., Oyama, A., & Shah, M. (2009). Abnormal crowd behavior detection using social force model. In *Computer vision and pattern recognition* (p. 935-942).

Moore, B. E., Ali, S., Mehran, R., & Shah, M. (2011). Visual crowd surveillance through a hydrodynamics lens. *Communications of the ACM*, *54*(12), 64-73.

Morris, B. T., & Trivedi, M. M. (2008). A survey of vision-based trajectory learning and analysis for surveillance. *Transactions on Circuits and Systems for Video Technology*, *18*(8), 1114-1127.

Nam, Y. (2013). Loitering detection using an associating pedestrian tracker in crowded scenes. *Multimedia Tools and Applications*, 1-23.

Nedrich, M., & Davis, W., J. (2010). Learning scene entries and exits using coherent motion regions. In *Advances in visual computing* (Vol. 6453, pp. 120–131).

Neri, F., Mininno, E., & Iacca, G. (2013). Compact particle swarm optimization. *Information Sciences*, *239*, 96 - 121.

Ngai, K. M., Lee, W. Y., Madan, A., Sanyal, S., Roy, N., Burkle, F. M. J., & Hsu, E. B. (2013). Comparing two epidemiologic surveillance methods to assess underestimation of human stampedes in india. *PLoS currents*, *5*.

Norman, B. C. (2012). Assessment of video analytics for exterior intrusion detection applications. In *International carnahan conference on security technology* (p. 359-362). doi: 10.1109/CCST.2012.6393585

Oussalah, M., & Schutter, J. D. (2000). Possibilistic kalman filtering for radar 2d tracking. *Information Sciences*, *130*(14), 85 - 107.

Pelletier, F. J. (1994). The principle of semantic compositionality. *Topoi*, *13*(1), 11-24.

Pitt, M. J., & Shephard, N. (2001). Auxiliary variable particle filters. In *Sequential monte carlo methods in practice* ((edited by A. Doucet, J.F.G de Freitas and N.J. Gordon) ed., p. 273-293).

Pitt, M. K., & Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, *94*(446), 590–599.

Poli, R. (2008). Analysis of the publications on the applications of particle swarm optimisation. *Journal of Artificial Evolution and Applications*, *2008*, 4:1–4:10.

PomÃ¡rico-Franquiz, J., Khan, S. H., & Shmaliy, Y. S. (2014). Combined extended fir/kalman filtering for indoor robot localization via triangulation. *Measurement*, *50*(0), 236 - 243.

Popplewell, J. (1986, January). *Committee of inquiry into crowd safety and control at sports grounds - final report.* Retrieved from `http://bradfordcityfire.files.wordpress.com/2013/02/popplewell-final-report-1986.pdf`

Ramendran, C. (2007, October). *Cops make public cctv footage of suspected killers.* Retrieved from `http://powerpresent.blogspot.com/2007/10/more-pics-video-inflated-rm90m-malaysia.html`

Rana, S., Jasola, S., & Kumar, R. (2011). A review on particle swarm optimization algorithms and their applications to data clustering. *Artificial Intelligence Review*, *35*(3), 211–222.

Ratcliffe, J. (2006). *Video surveillance of public places* (Tech. Rep.). United States: U.S. Department of Justice.

Ripley, A. (2008, Dec). *How to prevent a crowd crush.* Retrieved from `http://content.time.com/time/nation/article/0,8599,1864855,00.html`

Rodriguez, M., Sivic, J., Laptev, I., & Audibert, J.-Y. (2011). Data-driven crowd analysis in videos. In

International conference on computer vision (p. 1235-1242).

Rodriguez, S. K. T., M.and Ali. (2009). Tracking in unstructured crowded scenes. In *International conference on computer vision* (p. 1389-1396).

Rozmus, J. M. (2012, August). *Performance and limits of video analytics at the perimeter.* Government Security News Magazine. Retrieved from http://www.gsnmagazine.com/node/27012?c=perimeter_protection

Ruck, D. W., Rogers, S. K., Kabrisky, M., Maybeck, P. S., & Oxley, M. E. (1992). Comparative analysis of backpropagation and the extended kalman filter for training multilayer perceptrons. *Pattern Analysis and Machine Intelligence*, *14*(6), 686-691.

Rui, Y., & Chen, Y. (2001). Better proposal distributions: Object tracking using unscented particle filter. In *Computer vision and pattern recognition* (pp. 786–793).

Saisan, P., Medasani, S., & Owechko, Y. (2005). Multi-view classifier swarms for pedestrian detection and tracking. *Computer Vision and Pattern Recognition*, 18-24.

Sarafraz, A. (2013). *Boss: The crowd-scanning facial recognition system.* Retrieved from http://www.computervisiononline.com/blog/boss-crowd-scanning-facial-recognition-system

Scholz, S., Kawan, N., & Schindelhauer, A. (2012). *Report on video analytics enhanced* (Tech. Rep.). Secured Urban Transportation. Retrieved from http://www.secur-ed.eu/wp-content/uploads/2012/08/SECUR-ED.EU-D33.1-Report-on-Video-Analytics-enhanced.pdf

Schultz, P. D. (2008, June). *The future is here: Technology in police departments.* Retrieved from http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=print_display&article_id=1527&issue_id=62008

Schwerdt, D., K.and Maman, Bernas, P., & Paul, E. (2005). Target segmentation and event detection at video-rate: the EAGLE project. In *Advanced video and signal based surveillance.* (pp. 183–188).

Seekircher, A., Abeyruwan, S., & Visser, U. (2011). Accurate ball tracking with extended kalman filters as a prerequisite for a high-level behaviour with reinforcement learning. *Workshop on Humanoid Soccer Robots Conference*.

Shafie, H. (2008). *Video surveillance in public spaces* (Tech. Rep.). Malaysian communications and Multimedia Commision.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Shi, Y., & Eberhart, R. (1998). A modified particle swarm optimizer. In *Evolutionary computation proceedings* (p. 69-73).

Smith, K., Gatica-Perez, D., Odobez, J., & Ba, S. (2005). Evaluating multi-object tracking. In *Computer vision and pattern recognition* (p. 36).

Smith, K. C. (2007). *Bayesian methods for visual multi-object tracking with applications to human activity recognition* (Unpublished master's thesis). The University of Illinois.

Solmaz, B., Moore, B. E., & Shah, M. (2012). Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *Pattern Analysis and Machine Intelligence*, *34*(10), 2064–2070.

Staff, S. (2012, January). *Tsa report: Video analytics at buffalo's airport 100 percent accurate.* Retrieved from http://www.securitydirectornews.com/public-sector/tsa-report-video-analytics-buffalos-airport-100-percent-accurate

Star, T. (2013, November). *Suspects in baby freddie abduction freed as pair have alibis.* Malaysia. Retrieved from http://www.thestar.com.my/News/Nation/2013/11/20/Suspects-in-baby

-Freddie-abduction-freed-as-pair-have-alibis.aspx/

Star, T. (2014, February). *Public safety of utmost importance to council.* Malaysia. Retrieved from http://www.thestar.com.my/News/Community/2014/02/18/Public-safety-of-utmost-importance-to-council/

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *iConference*, 654-662. doi: doi:10.9776/14308

Stauffer, C. (2003). Estimating tracking sources and sinks. In *Computer vision and pattern recognition workshop* (Vol. 4, pp. 35–42).

Still, G. K. (2000). *Crowd dynamics* (Unpublished doctoral dissertation). University of Warwick.

Sun, C., Zeng, J., Pan, J., Xue, S., & Jin, Y. (2013). A new fitness estimation strategy for particle swarm optimization. *Information Sciences*, *221*, 355-370.

Szabó, Z. G. (2007). *Compositionality.* Retrieved from http://plato.stanford.edu/archives/spr2007/entries/compositionality/

Tan, C. P., Loy, C. C., Lai, W. K., & Lim, C. P. (2008). Robust modular artmap for multi-class shape recognition. In *International joint conference on neural networks* (pp. 2405–2412).

Tang, S. L., Kadim, Z., Liang, K. M., & Lim, M. K. (2010). Hybrid blob and particle filter tracking approach for robust object tracking. *Procedia Computer Science*, *1*(1), 2549–2557.

Tang, X., Wang, X., & Zhou, B. (2012). Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. *Computer Vision and Pattern Recognition.*, 2871-2878.

Taylor, L. J. (1989, April). *The hillsborough stadium disaster - interim report.* Retrieved from http://www.southyorks.police.uk/sites/default/files/Taylor%20Interim%20Report.pdf

Technology, V. (2012, September). *Video analytics faq.* Retrieved from http://www.vcatechnology.com/public/VCA-FAQs-(DRSandDani-6-9-2012).pdf

Thida, M., Eng, H.-L., Dorothy, M., & Remagnino, P. (2011). Learning video manifold for segmenting crowd events and abnormality detection. In *Asian conference on computer vision* (pp. 439–449).

Thida, M., Remagnino, P., & Eng, H.-L. (2009). A particle swarm optimization approach for multi-objects tracking in crowded scene. In *International conference on computer vision workshops* (pp. 1209–1215).

Thida, M., Yong, Y. L., Climent-Pérez, P., Eng, H.-L., & Remagnino, P. (2013). A literature review on video analytics of crowded scenes. In *Intelligent multimedia surveillance* (pp. 17–36). Springer Berlin Heidelberg.

Tian, Y.-L., Feris, R., & Hampapur, A. (2008). Real-time detection of abandoned and removed objects in complex environments. In G. Jones, T. Tan, S. Maybank, & D. Makris (Eds.), *International workshop on visual surveillance.*

Tong, G., Fang, Z., & Xu, X. (2006). A particle swarm optimized particle filter for nonlinear system state estimation. In *Congress on evolutionary computation* (p. 438 -442).

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Computer vision and pattern recognition* (pp. 1521–1528).

Tribunal of Inquiry on the Fire at the Stardust, D., Artane. (1981, February). *Report of the tribunal of inquiry on the fire at the stardust, artane, dublin on the 14th february, 1981.* Published by the Stationary Office (Dublin). Retrieved from http://www.lenus.ie/hse/bitstream/10147/45478/1/7964.pdf

Tun Haji Abdul Razak, M. N. (25 October 2013). *The 2014 budget speech* (Tech. Rep.). Government of Malaysia. Retrieved from http://www.pmo.gov.my/dokumenattached/bajet2013/SPEECH_BUDGET_2013_28092012_E.pdf

Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology*, *18*(11), 1473-1488. doi: 10.1109/TCSVT.2008.2005594

Valera, M., & Velastin, S. A. (2005). Intelligent distributed surveillance systems: A review. In *Vision, image and signal processing* (p. 192 - 204).

Van den Bergh, F., & Engelbrecht, A. P. (2006). A study of particle swarm optimization particle trajectories. *Information Sciences*, *176*(8), 937-971.

Van der Merwe, R., Doucet, A., de Freitas, N., & Wan, E. A. (2000). The unscented particle filter. *Technical Report*(CU ED/ F INF ENG/ R 380).

Van Der Merwe, R., & Wan, E. A. (2001). The square-root unscented kalman filter for state and parameter-estimation. In *International conference on acoustics, speech, and signal processing* (Vol. 6, pp. 3461–3464).

Varadarajan, J., & Odobez, J. (2009). Topic models for scene analysis and abnormality detection. In *International conference on computer vision workshops* (p. 1338-1345).

Velastin, S., Boghossian, B., Lo, B., Sun, J., & Vicencio-Silva, M. (2005). Prismatica: toward ambient intelligence in public transport environments. *Systems, Man and Cybernetics, Part A: Systems and Humans*, *35*(1), 164–182.

Velastin, S., & Remagnino, P. (2006). *Intelligent distributed video surveillance systems* (S. Velastin & P. Remagnino, Eds.). Institution of Engineering and Technology.

VideoIQ, & company, A. (2014). *Videoiq vertical solutions.* Retrieved from http://www.videoiq.com/product/public-and-private-communities/

Vishwakarma, S., & Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, *29*(10), 983-1009.

Wan, E. A., & Van Der Merwe, R. (2000). The unscented kalman filter for nonlinear estimation. In *Adaptive systems for signal processing, communications, and control symposium* (pp. 153–158).

Wang, C., Lu, H., Yang, W., & Liu, N.-H. (2010). Outdoor camera calibration method for a gps & camera based surveillance system. In *International conference on industrial technolog* (p. 263-267). .

Wang, F., & Lu, M. (2012). Hamiltonian monte carlo estimator for abrupt motion tracking. In *International conference on pattern recognition* (p. 3066-3069).

Wang, H., Sun, H., Li, C., & Rahnamayan, J.-S., S.and Pan. (2013). Diversity enhanced particle swarm optimization with neighborhood search. *Information Sciences*, *223*, 119–135.

Wang, L., Cheng, L., Zhao, G., & Pietikäinen, M. (2011). *Machine learning for vision-based motion analysis*.

Wang, X., Tieu, K., & Grimson, E. (2006). Learning semantic scene models by trajectory analysis. In *European conference on computer vision* (Vol. 3953, pp. 110–123).

Weiss, Y., & Adelson, E. (1996). A unified mixture framework for motion segmentation: incorporating spatial coherence and estimating the number of models. In *Computer vision and pattern recognition* (pp. 321–326).

Welch, G., & Bishop, G. (1995). *An introduction to the kalman filter* (Tech. Rep.). University of North Carolina at Chapel Hill.

Welsh, B. P., & Farrington, D. C. (2008). Effects of closed circuit television surveillance on crime. *Campbell Systematic Reviews*, 1-73.

Weng, S.-K., Kuo, C.-M., & Tu, S.-K. (2006). Video object tracking using adaptive kalman filter. *Journal of Visual Communication and Image Representation*, *17*(6), 1190–1208.

Wiliem, A., Madasu, V. K., Boles, W. W., & Yarlagadda, P. K. D. V. (2008). Detecting uncommon trajectories. In *Digital image computing: Techniques and applications* (p. 398-404).

Wong, K. C. P., & Dooley, L. S. (2011). Tracking table tennis balls in real match scenes for umpiring applications. *British Journal of Mathematics & Computer Science*, *1*(4), 228-241.

Wu, K. K., Tang, C. S., & Leung, E. Y. (2011). *Healing trauma: A professional guide*. Hong Kong University Press.

Wu, W. (2008). Tennis touching point detection based on high speed camera and kalman filter. *Technical Report Clemson University*.

Xia, Y., Deng, Z., Li, L., & Geng, X. (2013). A new continuous-discrete particle filter for continuous-discrete nonlinear systems. *Information Sciences*, *242*(0), 64 - 75.

Xiang, T., & Gong, S. (2006). Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, *67*(1), 21–51.

Xiang, T., & Gong, S. (2008). Video behavior profiling for anomaly detection. *Pattern Analysis and Machine Intelligence*, *30*(5), 893-908.

Xu, L.-Q. (2007). Issues in video analytics and surveillance systems: Research / prototyping vs. applications / user requirements. In *Advanced video and signal based surveillance* (p. 10-14).

Yan, F., Christmas, W., & Kittler, J. (2005). A tennis ball tracking algorithm for automatic annotation of tennis match. *British Machine Vision Conference*, 619–628.

Yan, J., & Pollefeys, M. (2006). A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *European conference on computer vision* (Vol. 3954, p. 94-106).

Yang, C., Duraiswami, R., & Davis, L. S. (2005). Fast multiple object tracking via a hierarchical particle filter. In *International conference on computer vision* (Vol. 1, p. 212-219).

Yang, H., Shao, L., Zheng, F., Wang, L., & Song, Z. (2011). Recent advances and trends in visual tracking: A review. *Neurocomputing*, *74*(18), 3823–3831.

Yang, J., Ji, H., & Liu, J. (2011). Particle swarm optimization algorithm for passive multi-target tracking. *Procedia Engineering*, *15*(0), 2398 - 2402.

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *ACM Computing Surveys*, *38*(4), 1-45.

Yokota, T., Ishiyama, S., Yamada, Y., & Yamauchi, H. (2002). Medical triage and legal protection in japan. *The Lancet*, *359*(9321), 1949.

Zhan, B., Monekosso, D. N., Remagnino, P., Velastin, S. A., & Xu, L.-Q. (2008). Crowd analysis: A survey. *Machine Vision and Applications*, *19*(5-6), 345–357.

Zhang, X., Hu, W., & Maybank, S. (2010). A smarter particle filter. In *Asian conference on computer vision* (Vol. 5995, pp. 236–246).

Zhang, X., Hu, W., Maybank, X., S.and Li, & Zhu, M. (2008). Sequential particle swarm optimization for visual tracking. In *Computer vision and pattern recognition* (pp. 1–8).

Zhang, X., Hu, W., Wang, X., Kong, Y., Xie, N., Wang, H., ... Maybank, S. (2010). A swarm intelligence based searching strategy for articulated 3d human body tracking. In *Computer vision and pattern recognition workshops* (pp. 45–50).

Zhang, Y., & Liu, Z.-J. (2007). Irregular behavior recognition based on treading track. In *International conference on wavelet analysis and pattern recognition* (Vol. 3, p. 1322-1326).

Zhen, W., Mao, L., & Yuan, Z. (2008). Analysis of trample disaster and a case study - mihong bridge fatality in china in 2004. *Safety Science*, *46*(8), 1255 - 1270.

Zheng, Y., & Meng, Y. (2008). Swarm intelligence based dynamic object tracking. In *Congress on evolutionary computation* (p. 405-412).

Zhong, H., Shi, J., & Visontai, M. (2004). Detecting unusual activity in video. In *Computer vision and pattern recognition* (Vol. 2, pp. 819–826).

Zhou, J., D.and Weston, Gretton, A., Bousquet, O., & Schölkopf, O. (2004). Ranking on data manifolds. In *Advances in neural information processing systems.*

Zhou, X., Lu, Y., Lu, J., & Zhou, J. (2012). Abrupt motion tracking via intensively adaptive markov-chain monte carlo sampling. *Transactions on Image Processing*, *21*(2), 789 -801.

Zhu, X., Vondrick, C., Ramanan, D., & Fowlkes, C. (2012). Do we need more training data or better models for object detection? In *British machine vision conference* (p. 1-11).

Zuriarrain, I., Lerasle, F., Arana, N., & Devy, M. (2008). An mcmc-based particle filter for multiple person tracking. In *International conference on pattern recognition* (p. 1-4).