Tracking Multiple Mobile Devices in CCTV-enabled Areas

Savvas Papaioannou

Kellogg College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2016

In loving memory of my dear friend Artemis Papamichael
(5 November 1986 - 25 July 2016)

Acknowledgements

First of all I would like to express my sincerest gratitude to my supervisors, Dr. Niki Trigoni and Dr. Andrew Markham, for their continuous support, guidance, and encouragement during my D.Phil study. I am very fortunate to have been their student. My supervisors are not only brilliant scientists but also incredible people.

Next, I would like to thank my friends in our group who have shared their valuable knowledge and experiences with me, and who have helped me in so many ways. In particular, I would like to thank Dr. Hongkai Wen, Dr. Traian Abrudan, Dr. Zhuoling Xiao, Dr. Orfeas Kypris, Dr. Bo Wei, Dr. Sen Wang, Dr. Andrew Symington, Dr. Muzammil Hussain, Ronnie Clark, Xiaoxuan Lu, Linhai Xie and Zhihua Wang.

I am deeply grateful to Laing O'Rourke for funding my studies through a D.Phil scholarship.

Moreover, I would like to thank my former supervisor from Yale University, Dr. Andreas Savvides, for his interesting discussions and generous help.

I would like to thank my examiners for their excellent comments and suggestions on my thesis, Prof. Kwiatkowska, Prof. Gibbons, Prof. McCann and Prof. Rogers.

A very big thanks goes to the one person very special to me. I would also like to thank all my friends who have encouraged me throughout my studies. Finally, I owe my deepest gratitude to my parents Andreas and Chryso Papaioannou, and to my sister Maria for their love and support over the last years. Without them this thesis would not be possible.

Abstract

Over the last decade we have witnessed an unprecedented interest in indoor positioning technologies, with a variety of solutions developed in academic and industrial research labs. Although the field has reached a significant level of maturity, there is still no dominant solution and, as a consequence, positioning services are still lacking in many buildings. In order for a solution to be widely implemented and adopted, two key requirements must be satisfied: low cost and high accuracy. The dichotomy between cost and accuracy has fragmented the technology landscape, leading to a plethora of competing solutions that cannot satisfy both requirements simultaneously.

The key objective of this thesis is to investigate how to unify the two disparate camps, providing high positioning accuracy with very low cost. Many approaches have tried to achieve this goal by fusing different sensor modalities. However, the majority of existing work has only investigated how to fuse sequences of measurements for which the associations with the targets are known (i.e. device personal data). Sensor fusion techniques that combine device personal data and anonymous sensor streams (where the associations between the measurements and the targets are not known) remain under-explored as of today. In this thesis we investigate how to efficiently combine device sensor data and anonymous sensor streams from various sensor modalities in order to build low cost and high accuracy positioning systems. By combining these two types of sensor modalities in one system we see a great potential in designing cost-effective and accurate positioning systems for challenging environments such as for tracking people in highly dynamic industrial settings. Our goal is to design a multi-target multi-sensor tracking framework which will utilise existing sensor infrastructure found in industrial environments and large public buildings (e.g museums) in order to provide reliable positioning services.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Nowadays positioning is becoming the backbone of many services and applications that aim to transform our everyday lives. A positioning system allows a mobile device to determine its location by enabling location-based services. Indoor/outdoor navigation, location-based advertising and assisted living are only some of the applications that could benefit from the development of accurate and robust positioning systems. For example, museums can use positioning services to provide the visitor with information on the exhibits, collect useful statistics about the exhibits (e.g. popularity) and analyse visitor behaviour. In a construction site, workers can get information on the specification and rules pertaining to components they use for building, etc. An accurate positioning system can also provide additional safety to the construction site personnel by issuing different types of warnings (e.g. notifications on entering danger zones). Furthermore, an accurate and cost-effective tracking system can be used to detect how active each individual is, to estimate energy expenditure and/or provide key insights on their well-being. Other applications include smart supermarkets that provide location-based navigation and smart hospitals which are able to track and locate medical personnel and equipment in case of emergency situations.

The Global Positioning System (GPS) [1] is the most widely used satellite-based positioning system providing accurate geographic positions in outdoor environments. However, GPS cannot work indoors or in the presence of obstacles that block Line-of-Sight (LoS) propagation from the satellites. Indeed, the performance of outdoor positioning has become excellent due to GPS but many applications require accurate and sub-meter positioning in both indoor and outdoor environments. Compared to the outdoor setting, indoor environments are more challenging. This is due to a) Non-Line-of-Sight (NLoS) conditions which result in high signal attenuation, scattering and multi-path due to walls, people and other obstacles, b) fast temporal changes and c) demand of high accuracy and low cost.

Furthermore in large buildings (e.g. museums and airports) and industrial settings (e.g. construction sites) positioning is a very challenging problem and quite different from the traditional indoor positioning (i.e. self-positioning using Wi-Fi fingerprinting). Construction sites, for instance, as with many other industrial settings, consist of a combination of complex indoor and outdoor areas. As a result, tracking the workers in a construction site introduces an additional layer of challenges compared to indoor tracking. This is mainly due to the many moving parts and the fast large-scale changes that occur in these complex environments. For instance, in an indoor environment, the positions of walls and floors remain constant over time and the positions of furniture vary little from day to day. Existing indoor positioning systems leverage this environmental stability to provide accurate location services with the use of stable maps. In contrast, the construction site evolves rapidly from day to day, precluding the use of systems which rely on stable, long-term maps for positioning. Currently, there is no system that allows for workers to be tracked reliably and robustly during all phases of construction. As a case in point, consider the challenges in a unified positioning system that works equally as well during deep foundation excavation through to an almost complete multi-storey building. At different points in time, the performance of different techniques alters, with some improving and some degrading.

Existing positioning systems do not provide a cost-effective solution and in addition they lack the necessary accuracy requirements. The goal is to have ubiquitous positioning services and for this to happen we need to have low-cost positioning systems. This is important since for a large number of organisations, positioning is not important enough to their core business to justify investment into bespoke infrastructure. Still, positioning would add key benefits enabling them to advance their businesses.

In this thesis we are interested in designing low-cost positioning systems. With the term low-cost positioning systems we refer to positioning systems which rely on sensor modalities found in todays buildings (e.g. airports, shopping malls, etc) such as WiFi and CCTV cameras. Systems that make use of dedicated positioning equipment have been well studied (i.e. [2, 3, 4]) in the past. These systems could be very accurate (i.e. centimetre accuracy) however they come with additional costs (i.e. the cost dedicated infrastructure, the cost of deployment and the cost of maintenance). For instance, the cost of just 25 UWB beacons [5] can be as high as $500, excluding the cost of UWB tags that people need to carry on them in order to be localized. These kind of systems are not cost-effective and not ubiquitous which renders them unappealing for many businesses. On the other hand many business are interested in positioning systems that can leverage the existing sensor infrastructure (e.g. WiFi, smartphones, cameras) thus providing a low-cost alternative.

Given all the above, we claim that a practical positioning system that can be widely adopted and used in variety of scenarios should satisfy the following requirements:

- **Low-cost**: A positioning system that requires expensive dedicated infrastructure is impractical for many applications and environments. For many environments the deployment of a dedicated positioning infrastructure such as Ultra Wideband and Ultra sound transceivers is not feasible due to the environmental dynamics and sensor limitations. Therefore, instead of requiring special-purpose equipment a practical solution should exploit the existing systems on site, which are being used for security and communication such as CCTV cameras and WiFi access points. Additional costs are due to the maintenance of the system. For instance maintaining a WiFi fingerprinting map can become very expensive, labor intensive and error prone. A practical system should thus be able to automatically keep itself up-to-date and adapt to the changing conditions.

- **Accurate:** Accuracy is a key requirement for the majority of positioning applications. For instance, in a construction site environment a positioning system that monitors the location of workers to indicate working hazards (e.g. red and green zones) should be highly accurate (i.e. with sub-meter accuracy) in order to provide the workers with the necessary safety. The term "accuracy" here indicates the distance between the estimated and true locations/trajectory. Many applications (e.g. location-based advertising) require meter level accuracy and even sub-meter accuracy (e.g. for indicating working hazards in construction sites). In this thesis we investigate whether it is possible to design systems with sub-meter accuracy that make use of multiple sensor modalities which are already deployed in many environments (e.g. museums, construction sites, etc)

- **Robust:** To date, the majority of positioning systems have been designed to operate within environments that have long-term stable macro-structure with potential small-scale dynamics. These assumptions allow for stable maps to be produced and gradually aged with the incorporation of minor variations. However, in many environments the aforementioned assumption is invalid. In highly dynamic industrial settings the environment evolves over time, and thus a good positioning system must be able to adapt to these environmental changes. The term "robustness" indicates the ability of a positioning system to operate equally well under different conditions and this is a fundamental requirement for a variety of applications.

- **Scalable:** A practical positioning system should be able to handle a variable number of tracking objects without significant performance drop. The system should be able to track and identify multiple people inside the camera field of view by utilising multiple sensor modalities. The requirement here is to be able to track and identify a variable number of objects (usually 1-10) inside areas which are covered by CCTV cameras (i.e. areas with sizes approximately 10m $\times$ 10m). In addition, it should be easy to extend the system to larger areas and to incorporate additional infrastructure without significant effort and cost. Positioning systems that require dedicated infrastructure are usually less scalable than systems that re-use existing infrastructure.

1.2 Illustrative Examples

The problem of designing a low-cost, accurate, robust and scalable positioning system is very important because it is the core component in many applications including location-based triggered information in museums, location-based advertising in shopping centres, coordination in hospitals and asset management in industrial settings. In what follows we present illustrative examples that outline the challenges above and motivate further this work.

1.2.1 Construction Sites

The construction industry is one of the most complex businesses. A report published in 2009 by the US National Research Council [6] found that construction lags behind other industries such as manufacturing in terms of productivity, and blamed the situation on problems with planning, coordination, and communication. The main reason for this is due to the fact that monitoring activity and tracking assets across a large, complex construction site is particularly difficult because there are so many moving parts, and because the jobs that are being performed change frequently. In addition, construction sites exhibit rapid large-scale changes in structure, the environment evolves over time, and the workers usually wear similar uniforms which makes them indistinguishable and thus hard to identify.

In order to understand better the challenges and why positioning is important in construction sites we have partnered with the Laing O' Rourke construction company in order to identify the problems the industry faces, its key requirements and how we can design a practical system that can help to improve their businesses. More specifically, the benefits of a positioning system in construction sites are as follows:

Figure 1.1: The figure shows the Blavatnik School of Government, University of Oxford construction site on two different days (left image taken on day 1 and right image taken on day 36). The site changes rapidly from day to day, precluding the use of positioning systems which rely on stable, long-term maps.

- **Productivity:** An accurate positioning system can be used to improve productivity by identifying bottlenecks and by locating personnel and equipment. In addition, it can be used to provide key insights on the daily operations and infer important statistics about the workers such as activity, frequent visited locations, energy expenditure, etc. This technology will allow manual work to be monitored and evaluated automatically.

- **Safety:** By having a system which is able not only to track but also identify individual workers (i.e. by their device/tag ID) we can improve the safety on site. For instance, we can deploy virtual geofencing and we can monitor the location of workers to indicate working hazards which can be individually tailored.

Currently, there is no system that allows for workers to be tracked reliably and robustly in construction sites and in similar environments. There is no system that can fully utilise the existing infrastructure and at the same time provide the required accuracy.

1.2.2 Museums

Positioning systems can also be used in museums with the aim of improving both the museum's management and the experience of the visitors. For instance, an accurate positioning system can be used to provide location triggered-based information to the visitors in order to offer additional history and facts about the exhibits. In addition, a system that can monitor the position of individuals can be used as a virtual guide and also provide useful

statistics (i.e. most popular artefact, etc) to the museum management. The latter can be used to resolve bottlenecks, plan-ahead, and improve scheduling as well as the visitor's experience. Furthermore, a positioning system can provide coordination and communication between the visitors and the museum guides. It can be used to help visitors find certain locations and navigate easily throughout the building.

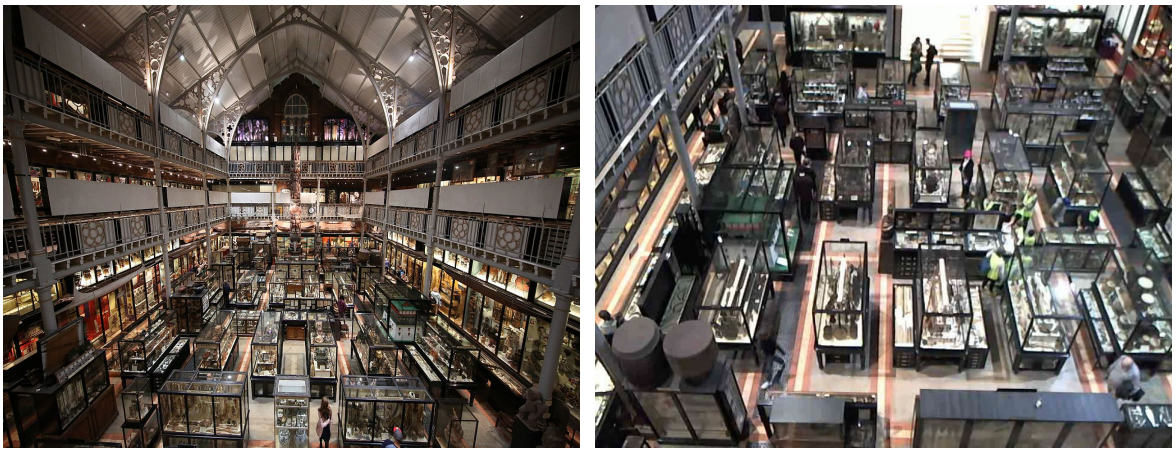


Figure 1.2: The Pitt Rivers Museum in Oxford

However, despite all the above-mentioned benefits, there is currently no positioning system that can cope with the challenges in museums namely:

- Cost requirements and limited infrastructure: Investing in bespoke positioning infrastructure is not effective. Moreover, making use of the existing infrastructure is tricky as it was not designed for positioning purposes.
- Environmental challenges: The number of people varies significantly over time, the room is usually not well lit, and the museum exhibits create occlusions, thus making maintaining tracking very challenging.
- Accuracy requirements: In order to be able to make use of the services discussed above, room-level accuracy is not adequate and sub-meter accuracy is required.

In this thesis we have partnered with the Pitt Rivers Museum (Fig. (1.2)) in Oxford in order to see all the above challenges first hand and be able to propose solutions and evaluate the proposed systems accordingly.

1.3 Overview of Existing Techniques

As we have explained above, in today's large buildings (e.g. museums and airports) and complex industrial environments (e.g. construction sites) the need for advanced planning and scheduling, careful coordination, efficient communication and reliable activity monitoring is essential for productivity and safety purposes. Accurate and cost effective positioning are the two main key requirements in order to meet all the above goals. Unfortunately, no adequate solution exists for providing accurate positioning services across these environments. Motivated by these challenges and by the importance of positioning for some applications we see a great need for positioning systems that, while using existing infrastructure, are able to obtain high accuracy. We first outline existing solutions and their relative merits, before motivating our research questions and direction.

Existing high-cost and high-accuracy solutions: Currently the state-of-the-art positioning systems that attain high accuracy require the use of expensive specialised equipment and substantial deployment effort. Systems based on Ultra-Wideband (UWB) [2, 7, 8, 9] for example can usually achieve centimeter-level positioning accuracy. However, due to the requirement for expensive dedicated transmitter-receiver infrastructure, this technology has never reached the mass market except for only few industrial implementations. Ultrasound (US) based techniques [10, 11, 3, 12, 13] can also achieve centimeter-level accuracy. However, these systems usually require the deployment of large fixed-sensor infrastructure and very careful sensor placement. In addition US-based systems are sensitive to temperature variations and they usually have an operating range of about 10m due to the strong decay profile of sound waves. Other positioning techniques that require the deployment of specialized infrastructure are based on active infrared (IR) and Radio-Frequency IDentification (RFID). Active IR positioning systems [4] usually require infrared receivers to be placed at fixed known locations (e.g. on the ceiling of a room) and mobile nodes transmitting IR beacons. IR-based systems can provide high precision and accuracy but require several receivers to be deployed in each room with additional cost. Such systems are also limited by the fact that IR signals cannot penetrate opaque materials (e.g. walls), thus LoS is required between transmitters and receivers. RFID-based positioning systems [14, 15, 16, 17] do not require LoS conditions and usually consist of several RFID active tags placed at fixed known locations and mobile nodes equipped with RFID readers which scan for nearby tags. Thus the accuracy and the cost of these systems highly depends on the density of the deployed infrastructure.

Existing low-cost and low-accuracy solutions: On the other hand, in order to provide cost-effective location services many indoor positioning systems and techniques exploit or reuse existing infrastructure (e.g. WiFi) or even operate without infrastructure (e.g. inertial tracking). These inexpensive, ubiquitous positioning approaches however cannot provide the sub-meter accuracy that high-end, expensive solutions like UWB or ultrasound can. For example, WiFi-based methods [18, 19, 20, 21, 22, 23, 24, 25] usually provide only several meter-level or room-level accuracy, use a significant number of access points and often require a very stable radio-map or a very accurate radio propagation model. Moreover, methods based on Inertial Measurement Units (IMUs) [26, 27, 28, 29, 30] have become quite popular given the wide spread of devices that include accelerometers, magnetometers and gyroscopes. However, these approaches usually require the IMU to be mounted on feet and/or a floor-plan in order to avoid large positioning errors.

Finally, as cameras are already deployed and available in a wide range of environments, vision-based techniques [31, 32, 33, 34, 35, 36, 37] can be used to detect and track people within their field-of-view (FOV) with high accuracy. However, in many scenarios visual tracking becomes very challenging due to path discontinuities, poor visual features, occlusions, etc. For example, although cameras (i.e. CCTV) provide relatively accurate position information, they are poor at distinguishing among people, especially when the individuals wear uniforms or helmets (e.g. in an industrial plant), when the room is not well lit (e.g. in a museum), or when the camera is pointing down at them making it difficult to distinguish facial features (e.g. bird's-eye view cameras in large stations). In addition, environments such as museums and airports are typically crowded with people crossing paths, and hiding behind dynamically changing obstacles. These conditions create ambiguities and make robust visual tracking a particularly challenging task.

We are now in a position to classify the aforementioned existing low cost solutions into two types namely *id-linked* and *anonymous*. Sensor data originating from a personal device (e.g. WiFi data from a smartphone) are called id-linked. Id-linked sensor modalities such as WiFi and IMUs incorporate a form of identification that can be used to identify the target that produced that data (e.g. WiFi MAC address), yet id-linked data typically provide poor position accuracy. The converse is true for anonymous sensor modalities (e.g. cameras, motion detectors) where the sensor data does not contain any form of explicit identification that can be used directly to link measurements to targets (i.e. anonymous sensor streams). This data, however, typically provides high positioning accuracy.

Our key hypothesis in this thesis is that we can use two types of sensor modalities (i.e. id-linked and anonymous) in order to create a low-cost high-accuracy positioning system. Our aim is to provide a practical system which would be able to operate robustly in a number of different environments from large public buildings such as museums and supermarkets to industrial settings such as construction sites. We believe that the fusion of id-linked and anonymous sensor modalities can help us resolve numerous positioning challenges and provide accurate positioning services.

The vast majority of existing work has only explored fusing id-linked sensor data (e.g. WiFi+FM [38], WiFi+Bluetooth [39, 40], WiFi+Inertial [41, 42], etc.). Our hypothesis is that in view of our requirements of high accuracy and low cost, fusing id-linked and anonymous data is a very promising research direction because it would allow us to combine the unique strengths of each approach and mitigate their limitations (Fig. (1.3)). However this area remains largely unexplored. Previous work (e.g. [43, 44, 45]) has recently unlocked the great potential of combining id-linked and anonymous data, but it is still faced with fundamental research challenges. It cannot be applied in a practical and robust way because it typically 1) requires significant training effort (for the id-linked data models), 2) is not robust to noisy anonymous data, 3) is not robust to changing environmental conditions and 4) none of the solutions has been tested and shown to be effective in any realistic setting.

1.4 Research Problem

As we have already discussed positioning systems that take advantage of id-linked and anonymous sensor modalities are under-explored. In addition, existing positioning systems that utilise id-linked and anonymous sensor data are tailored for specific applications and they do not provide a flexible framework which can be used in a variety of application scenarios. Therefore the main motivating question of the thesis is:

"How can we harness the power of id-linked and anonymous sensor data in order to build a positioning system that only relies on existing infrastructure, and yet is practical, cost-effective, robust and highly accurate ?"

In this thesis the term "positioning" indicates mainly continuous tracking i.e. we would like to design a system that provides the continuous trajectories of all objects being tracked. Before going further we outline our assumptions and key objectives in designing a practical positioning system.

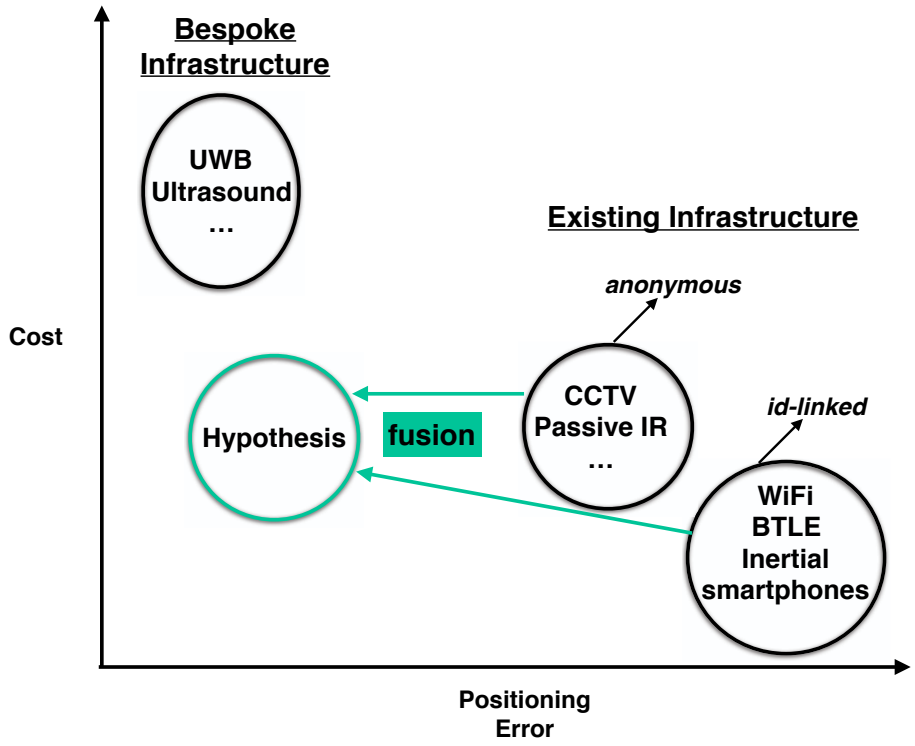


Figure 1.3: Our aim is to design positioning systems that combine the positioning accuracy of anonymous sensor modalities (e.g. cameras) with the identification accuracy of id-linked sensor modalities (e.g. WiFi). By doing so, our hypothesis is that we can design low-cost positioning systems with high-accuracy.

The main assumption we make in this thesis is the use of the existing infrastructure, which can be found in large buildings like museums, airports and industrial settings. Positioning systems that rely on the deployment of expensive, dedicated infrastructure have never reached the mass market. This is mainly because of the high deployment and maintenance costs. In addition, a dedicated fixed positioning infrastructure is not a viable solution for certain settings such as construction sites, where the environment changes significantly over time. From our interactions with our industrial partners we have realised that despite the fact that positioning systems could benefit their operations and organisation, investing heavily in stand-alone positioning infrastructure is not a viable option. Moreover, nowadays the majority of buildings and industrial settings are equipped with CCTV cameras and WiFi networks. In addition, people carry the ultimate positioning device, i.e. the mobile phone, which is equipped with cameras, WiFi/BTLE radios and inertial measurement units (IMUs).

In summary, in this thesis we will show how to utilise id-linked and anonymous sensor modalities (Fig. (1.3)) to design a practical positioning system that is low-cost, accurate,

robust and scalable. Below, we discuss the main research challenges that we have identified in addressing our research problem:

1.5 Research Challenges

Over the past few years we have understood very well that designing a positioning system which is low-cost, accurate, robust and scalable is a very challenging problem. Below are the main challenges we have faced during this time:

- **Limitations of existing infrastructure:**

  For many businesses investing in bespoke positioning infrastructure is not a viable option. Positioning systems that require the deployment of expensive specialised equipment never became widely adopted because of the scalability factor. These systems require substantial deployment effort and since the cost of deployment and maintenance is very high, it renders then impractical and difficult to scale.

  In addition, we should note here that in many scenarios we cannot install additional positioning equipment to aid the positioning task nor do we have the freedom to tune and alter the existing infrastructure for the same purpose (e.g. like in museums, construction sites). For instance, we might have a WiFi network which can provide reliable internet access but not enough WiFi access points that can be used for the task of positioning. Secondly, we have to make use of equipment that has been designed for purposes other than positioning. For example, we can use WiFi signals for positioning but unfortunately this method alone is not reliable enough since WiFi was designed for communication and data transfer purposes. The bottom line is that in large buildings and industrial settings the sensing infrastructure is limited and the sensors that can be used for positioning purposes are not ideal for this task. We cannot equip the construction sites with a dense WiFi infrastructure. WiFi access points can in fact be found only at the perimeter of the site and their distribution is not uniform providing only limited coverage. Additionally, most security cameras are installed to provide a large field of view, typically resulting in a birds eye view of the scene. This top-down perspective makes it difficult to distinguish facial features which can be easily used for identification and positioning (Fig. (1.4)).

  A practical positioning system should be ubiquitous and blend seamlessly in our everyday life. Thus, the positioning systems should be able to rely opportunistically on the available existing infrastructure but at the same time provide accurate positioning, be scalable and cost effective. We see a big need for positioning systems that

easily scale according to the available infrastructure without requiring additional cost and significant maintenance effort.

- **Environmental variability:** A big problem with the existing techniques is the assumption of environmental stability i.e. they have been designed for specific environments and they assume stable macro-structure with potential small-scale dynamics which allows them to use stable maps and use dedicated sensor infrastructure to provide positioning services. Unfortunately, these assumptions are not valid in many situations. Take for instance construction sites which are characterised by rapid large-scale changes in structure which greatly affect the sensor measurements (e.g. WiFi RSS). Furthermore, within periods of a few weeks the environment can change significantly as staircases and entire floors are added making positioning a very challenging task. So we need to design a system that can cope with changing conditions and adapt accordingly by making use of the different sensor modalities available. For instance, when there is an occlusion due to the addition of a new wall and visual tracking cannot be carried out, a good system can potentially switch to inertial tracking using sensor tags or mobile phones carried by the workers to provide a location estimate. In a museum the distribution of people changes significantly over time making visual tracking challenging. In addition the existing WiFi infrastructure is not adequate for positioning and furthermore the exhibit stands create occlusions and multi-path which makes positioning very challenging in these environments. Additional challenges come from noisy sensor measurements, changing lighting conditions, occlusions, path discontinuities due to people crossing paths, etc (Fig. (1.1)).

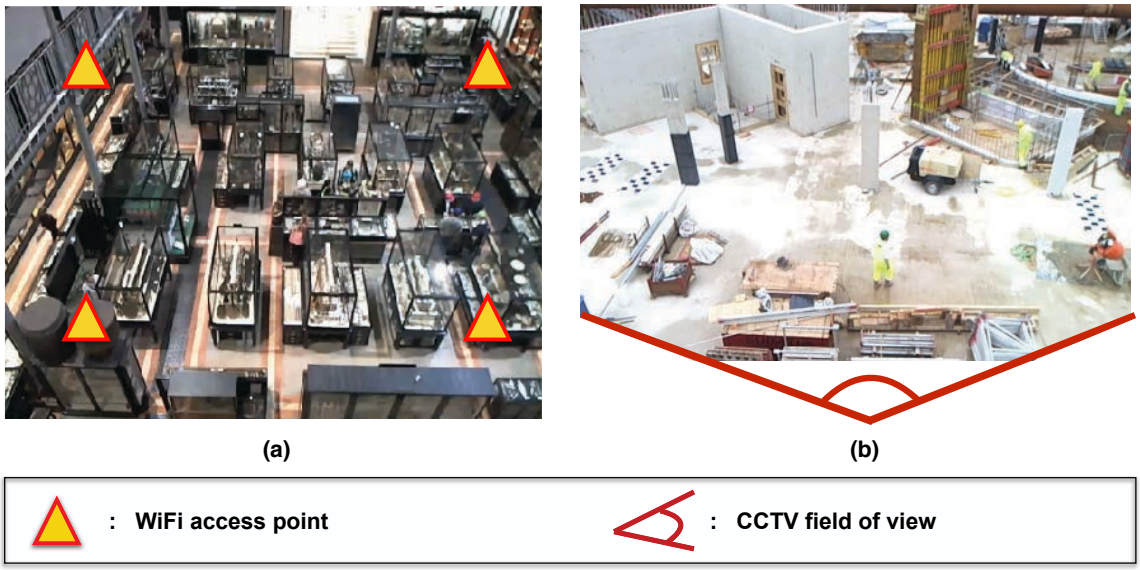


Figure 1.4: (a) The image of a museum and the available WiFi access points (yellow triangles). Due to the building topology, the WiFi access points are installed at fixed locations on the perimeter of the room and provide limited coverage. There are not enough access points to provide accurate positioning using fingerprinting techniques. (b) The image shows the CCTV view from a construction site. Security cameras are sparse and are installed to provide a large field of view. This top-down view makes it difficult to distinguish facial features which can be used for identification and positioning.

1.6 Contributions

With this thesis we aim to enable accurate and low-cost positioning services in challenging environments. For this purpose we propose novel positioning techniques that can be used in large public areas, buildings and industrial settings with no additional infrastructure. As of today positioning systems that can operate in such environments remain under-explored. Concretely, the technical contributions of this thesis are as follows:

- We have motivated the positioning problem in a wide range of real-world scenarios and applications. We have also identified the main challenges that exist in these scenarios and applications and why existing techniques are not adequate to solve these problems.

- We provide a fresh perspective on the problem of low-cost high-accuracy positioning in large open-plan indoor spaces such as museums and airports. We show that we can leverage anonymous visual data and id-linked radio data to solve positioning challenges in these environments achieving low cost and high accuracy at the same

time. We propose a deferred decision logic (i.e. we generate historic tracks over a period of time) multi-hypothesis probabilistic approach (RAVEL), in Chapter 3, to fuse radio and camera data that is robust to noisy and incomplete measurements. We leverage the discriminative power of WiFi RSS signature over a period of time and we show that although WiFi measurements are not by themselves sufficiently accurate for positioning services in these environments, when they are fused with camera data, they become a catalyst for pulling together ambiguous, fragmented, and anonymous visual trajectories into accurate and continuous paths, yielding typical errors below one meter.

- We have investigated the problem of real-time tracking in highly dynamic industrial settings and we propose a flexible positioning architecture for these rapidly changing environments. In Chapter 4, our Rao-Blackwellized particle-filter tracking framework, as opposed to RAVEL which is an off-line tracking system, utilises three different sensor modalities (i.e. vision, radio and inertial) to allow for accurate real-time tracking in challenging conditions and environments such as construction sites which are characterised by rapid large-scale changes in structure. Moreover, we show that the addition of inertial measurements in combination with visual and radio observations gives us additional benefits, i.e. we completely bypass the problem of occlusions and we show how we can maintain tracking under long term occlusions.

- We demonstrate the impact of applying the social force model to improve real-time tracking in dynamic environments (Chapter 4). Human motion is not always predictable. In fact in certain situations is very challenging to model and predict accurately the human motion. More specifically, human motion is affected by the environment (i.e. walls, obstacles) and from the motion of other people. In this thesis we show how we can integrate and use social forces to improve the prediction of human motion, as well as the overall tracking accuracy. This is very important in a lot of scenarios (e.g. industrial settings) where high accuracy and real-time operation is of essence.

- We propose a technique for cross-modal sensor parameter learning in Chapter 5. The proposed system is able to automatically tune the parameters of its sub-systems (e.g. radio model, visual detector, step-length estimator) by making use of the tracking output and a subset of sensor modalities. We show how to make our system adaptive and suitable for real-time tracking in dynamic environments by learning for instance the radio model and operate robustly in the presence of occlusions and sensor measurement noise. We further observe that in industrial settings (e.g. construction sites)

the environment changes rapidly with the addition of new walls, corridors, etc. and we show how to take advantage of these environmental changes to learn maps of the environment which can then be used in combination with social-forces to improve the prediction of human motion.

- We have conducted extensive experiments in a real construction site and in a museum with the help and guidance of our industrial partners. We have evaluated the proposed systems in real world conditions, we have observed the real world problems and we have refined our techniques in order to meet the real world challenges. In addition, we have compared the proposed systems against the competing techniques that are currently used in similar environments and we show that our proposed methods outperform these competing methods.

1.7 Publications

The main contributions of this thesis have been published in the following international peer-reviewed conferences and journals:

- **Savvas Papaioannou**, Hongkai Wen, Andrew Markham and Niki Trigoni. *"Fusion of Radio and Camera Sensor Data for Accurate Indoor Positioning."*, In Proceedings of the 11th IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS'14), pp. 109-117. IEEE, 2014.

- **Savvas Papaioannou** Hongkai Wen Zhuoling Xiao Andrew Markham and Niki Trigoni. *"Poster: WiFi Sensors Meet Visual Tracking For An Accurate Positioning System."*, In the 11th European Conference on Wireless Sensor Networks (EWSN'14), 2014.

  In these two papers, we propose a novel positioning system which fuses anonymous visual detections captured by widely available camera infrastructure, with radio readings. We show that although radio measurements are not by themselves sufficiently accurate, when they are fused with camera data, they become a catalyst for resolving visual ambiguities. This method is discussed in Chapter 3.

- **Savvas Papaioannou**, Andrew Markham and Niki Trigoni. *"Tracking People in Highly Dynamic Industrial Environments"*, in IEEE Transactions on Mobile Computing, 2016.

  In this paper we present a novel multiple-target tracking framework suitable for highly dynamic industrial environments. More specifically, we show that the fusion of radio, inertial and WiFi observations and the use of social forces increase the tracking accuracy in challenging environments which are characterised by rapid large-scale changes in structure. This work is discussed in Chapter 4.

- **Savvas Papaioannou**, Hongkai Wen, Zhuoling Xiao, Andrew Markham and Niki Trigoni. *"Accurate Positioning via Cross-Modality Training.*", In Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys'15), pp. 239-251. ACM, 2015.

  In this work we are exploiting the fact that different sensing technologies have uncorrelated failure modes which we use to provide a robust and adaptive positioning framework. We propose a cross-modality learning technique to deal with the environmental dynamics and we use it to learn the internal parameters of our system. This work is discussed in Chapter 5.

I would like to thank Dr Hongkai Wen and Dr Zhuoling Xiao for their help during the experimental setup in a museum and a construction site and my supervisors for their valuable feedback and help.

1.8 Thesis Structure

The rest of this thesis is organised as follows. Chapter 2 provides an overview of related work and outlines limitations of existing systems. The following three chapters present our proposed approaches. Chapter 3 presents RAVEL (Radio And Vision Enhanced Localisation), which fuses anonymous visual observations with radio readings in order to increase the tracking accuracy. Chapter 4 focuses on the problem of on-line multi-target tracking using inertial, WiFi and visual observations. Chapter 5 presents a cross-modality learning approach, which we use to learn the internal parameters of the positioning system. Finally, Chapter 6 concludes this thesis and outlines areas for future work.

Chapter 2

Background

Accurate, practical and cost-effective positioning systems are very important for a variety of applications and services. Hospitals can use location information to track expensive equipment and provide navigation services to patients. Construction sites can benefit from cost-effective positioning systems by providing safety to their personnel whereas museums and shopping malls can use positioning systems to build rich location services that would provide the visitors with location-based triggered context-aware information.

In this chapter we will give a high level overview of existing positioning systems and techniques. Then we will focus on state-of-the-art positioning systems which are more relevant to this thesis outlining the advantages and limitations of each method. More extensive surveys covering a variety of positioning systems and technologies can be found in [46, 47, 48, 49].

First, we give a brief introduction on tracking techniques and discuss fundamental positioning that relies on a single id-linked sensor modality. We then proceed to describe more sophisticated positioning systems that use more than one sensor modality to aid and improve the positioning results. These systems use sensor fusion in order to combine one or more sensor modalities into the positioning task and can be divided into two categories: a) systems that use only id-linked sensor modalities for sensor fusion and b) systems that fuse id-linked and anonymous sensor modalities. Because in the systems of the first category the association of measurements to the targets is known, these systems do not require any special procedure for knowing which target emits each measurement. On the other hand, systems that use anonymous sensor data require to solve the data association problem, otherwise known as motion correspondence in visual tracking in order to figure out which measurement originates from which target.

Table (2.1) gives an overview of the advantages and limitations of positioning systems that are based on particular sensor modalities, including typical accuracies and costs.

Overview of Existing Positioning Systems				
Technology	**Cost**	**Accuracy**	**Advantages**	**Limitations**
id-linked sensors				
WiFi ([18, 50])	Low	m	Ubiquitous inexpensive positioning	Requires manual training to build the radio map.
Inertial ([28, 27])	Depends on the quality of the IMU	m	No infrastructure needed	Usually requires a floor-plan and the IMU to be body mounted. Over time the location estimates diverge from the true location i.e. drift.
RFID ([17, 16])	Varies (low to high)	cm - m	For specific applications it can achieve good price/performance ratio.	High accuracy comes at high cost. Range is low.
Ultrasound ([13, 12])	High	cm	High accuracy	Requires dedicated hardware and careful deployment.
UWB ([9, 2])	High	cm	High accuracy	Expensive transmitter-receiver infrastructure.
Fusion of id-linked sensors				
WiFi + Inertial ([41, 42])	Usually low (i.e. smartphone IMU)	m	It can avoid manual fingerprinting. It can resolve fingerprinting ambiguities.	Accuracy depends on various factors (i.e. number of APs, availability of floor-plan, magnetic disturbances, etc.)
WiFi + FM ([38])	Low	m	Does not require extra infrastructure, more robust than WiFi.	Requires extensive site surveying.
anonymous sensors				
Cameras ([51])	Varies (low to high)	cm	Passive multi-target tracking.	Often targets cannot be identified. Poor visual features, occlusions and people crossing paths cause problems.
Fusion of id-linked and anonymous sensors				
Camera+Radio ([43, 52])	Varies (low to high)	cm - m	Accurate tracking and identification can be achieved.	Requires radio calibration or uses expensive UWB equipment. Radio is not used to resolve visual ambiguities.
Camera+IMU ([44, 45])	Varies (low to high)	cm - m	Accurate tracking and identification can be achieved.	Typically the IMU must be body mounted.

Table 2.1: Overview of the advantages and limitations of positioning systems that are based on particular sensor modalities.

2.1 Tracking

In this section we provide a brief overview of target tracking techniques. The main purpose of a target tracking algorithm is to estimate the trajectory of one or more targets by correctly assigning them a specific label which must be propagated over time. Targets are usually modelled as time-varying systems using state-space models and the purpose of tracking is to estimate at each time step the state (e.g. position, velocity, etc) of every target through noisy measurements. In the probabilistic approach the evolution of each target can be described using a sequence of conditional probability distributions:

$$x_k \sim p(x_k|x_{k-1}) \tag{2.1}$$

$$y_k \sim p(y_k|x_x) \tag{2.2}$$

where $x_k \in \mathbb{R}^n$ is the state of the system at time step k, $y_k \in \mathbb{R}^m$ is the measurement of the target at time k and

- $p(x_k|x_{k-1})$ is the dynamic model of the target which describes the stochastic dynamics and uncertainties (i.e. process noise) of the target states over time.
- $p(y_k|x_x)$ is the measurement model, which describes the distribution of measurements given the state of the target accounting also for the measurement noise.

The above model is assumed to be Markovian which means that x_k given x_{k-1} is independent of anything that has happened before time step $k-1$ i.e. $p(x_k|x_{1:k-1}, y_{1:k-1}) = p(x_k|x_{k-1})$ and also we assume the conditional independence of measurements, that is $p(y_k|x_{1:k}, y_{1:k-1}) = p(y_k|x_k)$.

2.1.1 Optimal Filtering

The purpose of optimal filtering [53] is to recursively compute the marginal posterior distribution of the state x_k at each time step given the history of measurements up to time k i.e. $p(x_k|y_{1:k})$. The Bayesian recursive solution to the filtering problem is computed through a sequence of prediction and update steps as follows:

- Initialisation: The recursion starts from the prior distribution of target states $p(x_0)$.
- Prediction: In the prediction step, the Chapman-Kolmogorov equation is used to calculate the predictive distribution $p(x_k|y_{1:k-1})$ i.e. the predicted pdf of x_k based

on the measurements up to time $k-1$. Given the dynamic model of the system $p(x_k|x_{k-1})$ the predictive distribution is given by:

$$p(x_k|y_{1:k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|y_{1:k-1})\ \mathrm{d}x_{k-1} \tag{2.3}$$

- Update: At time step k, a new measurement y_k becomes available and the Bayes rule can be used to obtain the posterior distribution of the state x_k using the measurement model $p(y_k|x_x)$ as:

$$p(x_k|y_{1:k}) = \frac{1}{Z_k}p(y_k|x_x)p(x_k|y_{1:k-1}) \tag{2.4}$$

where the normalisation constant Z_k is given by $Z_k = \int p(y_k|x_x)p(x_k|y_{1:k-1})\mathrm{d}x_k$.

If this problem of recursively calculating the posterior distribution is solved exactly, the optimal Bayesian solution is obtained. However, this optimal solution only exists in certain cases since it involves the evaluation of complex high-dimensional integrals. In such cases approximate suboptimal solutions can be obtained.

2.1.2 Kalman Filter

The Kalman filter [54] is the closed form solution to the optimal filtering problem discussed above. The exact solution that the Kalman filter provides is due to the following assumptions: First the Kalman filter assumes that the dynamic and measurement models are linear and second that the process and measurement noises are drawn from zero mean Gaussian distributions. This means that the resulting posterior distributions calculated by the Kalman filter are Gaussian. The main advantages of the Kalman filter are the optimal Bayesian solution and the low computational complexity and memory requirements. However, when the target dynamics and measurement model are not linear Gaussian, Kalman filter cannot be used, which is the main drawback of this method.

2.1.3 Extended Kalman Filter (EKF)

As we have already mentioned often when the dynamic and measurement processes in practical applications are not linear and the Kalman filter cannot be applied. In such cases alternative sub-optimal methods can be applied. A popular technique is the extended Kalman filer (EKF) [55, 56] which first approximates the non-linearities with a first order Taylor expansion and then uses the standard Kalman filter equations. In the EKF the process and measurement noises are also assumed to be Gaussian, thus the EKF approximates the posterior distribution of the target states with a Gaussian density. If the system under

consideration has weak non-linearities only, this suboptimal algorithm can be very effective. However, if the non-linearities become severe, the performance of the filter decreases rapidly.

2.1.4 Unscented Kalman filter (UKF)

For highly non-linear systems, the extended Kalman filter can give very poor performance due to the linearised covariance propagation of the underlying non-linear system. An alternative option in this case is the unscented Kalman filter [57, 58]. As opposed to the EKF which tries to approximate the non-linear functions of the system, the UKF approximates the probability distribution of the non-linear system instead with a Gaussian density. The unscented Kalman filter uses a deterministic sampling technique (i.e. the unscented transformation) to pick a minimal set of points that best represent the state distribution. This set of points is then propagated through the non-linear functions from which both the mean and covariance of the state estimate are recovered. Compared to the EKF which provides a first order Taylor series approximation of the non-linear system the UKF captures the posterior mean and covariance accurately to the 3rd order thus is more suitable for severe non-linearities. The disadvantage of UKF is that the posterior distribution is approximated by a Gaussian density as well, which means if the true density is non-Gaussian, the UKF fails to converge.

2.1.5 Particle filter (PF) methods

For many filtering problems the Gaussian approximations of the UKF work well. However in certain cases the filtering distributions can be non-Gaussian, multi-modal or some of the state components might be discrete, in which cases the Gaussian approximations of the UKF are not appropriate. A more general way to approximate the filtering equations is to use particle filter (PF) methods [59]. These methods are also known as sequential Monte Carlo, sequential importance (re)sampling, bootstrap filter. Their main advantage is their ability to approximate complex high dimensional densities, i.e. they can be used to approximate states of non-linear dynamical systems and non-Gaussian noise.

In this paragraph we will give a brief overview of the sequential importance resampling technique (SIR) [60] which is the most widely used particle filter method. The SIR method is used to approximate the posterior distribution $p(x_k|y_{1:k})$ by a set of N weighted particles (finite set of samples), i.e. $\{x_k^n, w_k^n\}_{n=1}^N$ where the large weights indicate the states with high posterior probability. The weights are normalised such that $\sum_n w_k^n = 1$. Therefore,

the posterior distribution at time k can be approximated with a sum of N weighted Dirac deltas as:

$$p(x_k|y_{1:k}) \approx \sum_{n=1}^{N} w_k^{(n)} \delta(x_k - x_k^{(n)}) \tag{2.5}$$

The weights are chosen using the method of importance sampling. According to this, instead of sampling directly from the posterior distribution which can often be difficult, we use an approximate distribution called importance distribution $\pi(x_k|x_{1:k-1}, y_{1:k})$ from which we can easily draw samples. In many target tracking problems the transition prior distribution which is used to model the target dynamics is often used as the importance distribution i.e. $\pi(x_k|x_{1:k-1}, y_{1:k}) = p(x_k|x_{k-1})$. At each time step the weights are obtained from the following recursion:

$$w_k^{(n)} = w_{k-1}^{(n)} \frac{p(y_k|x_k^{(n)}) p(x_k^{(n)}|x_{k-1}^{(n)})}{\pi(x_k^{(n)}|x_{k-1}^{(n)}, y_k)} \tag{2.6}$$

where the dynamic and measurement models have been used to compute the quality of each particle, which is similar to the prediction and update steps of the Kalman filter.
Finally, particles with very low weights are eliminated and new ones are generated based on the particle distribution, a process called re-sampling which greatly improves the stability of the algorithm.

2.1.6 Rao-Blackwellized Particle Filtering

Particle filtering i.e. sequential Monte Carlo (SMC) [59] has become a very popular and practical numerical technique for approximating the Bayesian recursion in many tracking problems. This is because particle filter techniques make no assumptions about the system dynamics or the distributions that govern the states and the measurements. They can be used to approximate high dimensional densities and the states of non-linear dynamical models and non-Gaussian noise. Their adoption in a wide range of problems is also due to their simplicity, efficiency and ease of implementation.

Many tasks that require real-world data analysis involve estimating unknown quantities from given observations. In certain scenarios these observations arrive sequentially in time therefore one is interested in performing inference in real-time. Particle filtering methods however can be inefficient and computationally expensive since in high dimensional state-spaces a large number of samples is needed in order to represent the joint posterior distribution.

In certain cases however there are more efficient ways to use particle filters and this is by marginalising out some of the variables analytically which reduces the size of the state-space that needs to be sampled. In other words, it is possible to calculate some of the variables analytically and approximate others by sampling. This is called Rao-Blackwellisation and the derived filter that combines these two techniques is called Rao-Blackwellized particle filter (RBPF) [61, 59].

More specifically in Bayesian inference, the objective is to compute the posterior distribution $P(z_{1:t}|y_{1:t})$ where $z_{1:t}$ are the hidden states at times 1 to t and $y_{1:t}$ are the received measurements up to time t. Suppose we can decompose the hidden state z_t into two parts: a root variable λ_t and a leaf variable x_t as shown below:

$$P(x_{1:t}, \lambda_{1:t}|y_{1:t}) = P(x_{1:t}|\lambda_{1:t}, y_{1:t})P(\lambda_{1:t}|y_{1:t}) \tag{2.7}$$

If we can compute the conditional posterior distribution $P(x_{1:t}|\lambda_{1:t}, y_{1:t})$ analytically, then we only need to sample from $P(\lambda_{1:t}|y_{1:t})$ using the particle filter. Thus, the main idea of Rao-Blackwellized particle filtering (RBPF) [61, 59] is to reduce the number of variables that are sampled by evaluating some parts of the filtering equations analytically. This reduction makes RBPF computationally more efficient than the standard particle filter, especially in high dimensional state-spaces.

2.2 Positioning with one id-linked sensor modality

2.2.1 WiFi

WiFi localisation techniques are becoming quite popular nowadays due to the wide availability of WiFi access points (APs) in indoor environments. In addition, for many applications these methods are considered a very good choice due to the price-performance ratio. The most popular WiFi based localisation techniques can be divided into two categories: a) attenuation-based and b) fingerprinting-based.

2.2.1.1 Attenuation-based

Attenuation-based techniques use the received signal strength (RSS) at the user's position to estimate his/her distance from various APs of known locations using a radio propagation model. Given the distance estimates from various APs the user's position can be calculated using techniques like lateration. The authors in [62] use the Friis transmission equation to convert the RSSs of a mobile node to distances from three APs of known locations and then use lateration to find the location of the mobile node. More sophisticated radio

propagation models have been proposed [63, 64, 65, 66] which take into account the type of the environment and the structure of the building (i.e. walls, floors) in order to provide more accurate distance estimates. The authors in [50] use a radio propagation model that takes into account the walls in the building. Before using the radio model for distance estimation they use multiple regression analysis to learn the wall attenuation coefficient from training data. The technique in [67] follows a different approach by learning the radio model parameters on the fly using a maximum likelihood estimation technique, avoiding thus the off-line calibration.

The main drawback of these methods is the inaccurate distance estimation often caused by multi-path effects and non-line-of-sight (NLOS) conditions which cannot be modelled precisely by the radio propagation models.

2.2.1.2 Fingerprinting-based

The main positioning technique that is used by the majority of WiFi-based systems is RSS fingerprinting.

Fingerprinting approaches can be divided in two categories namely deterministic and probabilistic. Deterministic techniques build the radio map by representing the signal strength of an access point at a specific location with a scalar value (i.e. mean RSS value). On the other hand, probabilistic approaches represent the signal strengths at the various locations as probability distributions. A detailed survey describing the different fingerprinting methodologies can be found in [68]. Radar [18, 19] is probably the most well known positioning system which demonstrates the idea of fingerprinting. The system uses WiFi APs from the infrastructure and operates in two phases. In phase-one (training phase) a radio map is built by collecting RSS data with a mobile node at known locations with respect to three base-stations. In phase-two, nearest neighbours techniques are being used to infer the node's location from the radio map given its signal strength readings from the unknown location. The initial RADAR system was able to locate a user within a few meters of his/her actual location. More recent probabilistic approaches [21, 22] report improved accuracy by statistically modelling the signal strength as Gaussian distribution.

Although fingerprinting-based localisation methods have become quite popular they usually require extensive and labour intensive training and a large number of APs in order to provide high positioning accuracy. In addition, the fingerprint database usually requires re-training when spatial environmental variations occur. Depending on the density of calibration points, the number of APs and the type of the environment WiFi fingerprinting techniques usually reach accuracies of 2 to 10 meters.

2.2.2 Inertial

Inertial tracking techniques use inertial measurement units (IMUs) containing accelerometers, gyroscopes and magnetometers to estimate the user's trajectory relative to his/her initial position (i.e. pedestrian dead reckoning). Such techniques assume that the user's initial position at a specific time t_0 is known or can be accurately estimated. Then the user's position at time $t_0 + \delta t$ can be calculated by integrating their velocity, or twice-integrating their acceleration, during the time interval δt. Popular techniques in this category [29, 26, 28] use foot mounted IMUs. This is convenient since it allows the use of zero velocity updates (ZUPTs); the velocity inference is set to 0 m/s every time the IMU detects the foot on the ground and thus the accumulating integration errors can be significantly reduced. Other authors including [27, 30] explore more realistic scenarios where the IMU is assumed to be embedded in a smart-phone device. The work in [27] investigates the case where the IMU is in hand whereas the authors in [30] develop an adaptive motion sensing algorithm to cope with the IMU's position and the different walking profiles. Various sources of error (e.g. calibration inaccuracies, external magnetic disturbances, etc.) can quickly cause the location estimate of these techniques to diverge from the true location.

2.2.3 Radio Frequency Identification (RFID) and Infrared (IR)

RFID is an automatic wireless identification technology in which RFID tags can communicate with RFID readers and transmit their serial numbers (IDs). The RFID reader is then capable of knowing the identity and the approximate location of a tag (via proximity or fingerprinting). RFID tags can be either passive (i.e. they operate without batteries) or active (i.e. they operate with batteries). Passive tags are less expensive than active tags but they have a limited range of about 1-2 meters as opposed to active tags which typically exhibit a much longer range (i.e. 10-30 meters). The early LANDMARC system [14] uses the RFID technology to locate mobile users carrying active RFID tags in a 4m by 9m room. The RFID readers are used to scan periodically for nearby tags. Based on the received signal strength information k nearest-neighbors techniques are used to find the location of each tag. More recently Seco et al. [17] conducted a more extensive experiment where 71 active RFID tags were distributed in a building covering a total area of 1600 square meters (55 different rooms). A system calibration phase was performed by collecting RFID RSS measurements from different locations in the building in order to build a radio-map. Finally, a user equipped with a RFID reader could localise him/herself using the radio-map and RSS measurements received from the infrastructure. The median positioning accuracy of this system was 1.5 meters.

Infrared based positioning works in a similar way. More specifically, the Active Badge system [4] illustrates how the IR technology can be used for room level localisation. In Active Badge each person to be localised is wearing a small IR badge which emits a unique code every 15 seconds. These signals are picked up by an IR sensor network around the building. A central server collects this data and provides the location of people in semantic form (i.e. Kitchen).

As we have already mentioned the main disadvantage of these systems is that they require extensive infrastructure to be deployed in the building. Thus their accuracy and cost highly depends on the density of the deployed hardware.

2.2.4 Ultrasound (US) and Ultra Wideband (UWB)

Ultra-Wideband [2, 7, 8, 9] and Ultrasound [10, 11, 3, 12, 13] methods were briefly discussed in the introduction and we are not going to go into more details in this chapter. These methods typically utilise time-based measurements to accurately estimate the distances between sensor nodes. They require expensive dedicated transmitter-receiver infrastructure which makes them non ubiquitous and impractical for a lot of applications.

2.2.5 Discussion

In this section we have outlined systems that rely on one id-linked sensor modality to provide positioning services. These systems have changed the way we see the world and helped us understand the unlimited potential of positioning technologies. In certain scenarios some of the above systems can still provide an acceptable positioning solution. However, as we move forward to a future where positioning is ubiquitous and completely hidden from our lives relying on just one sensor modality is not sufficient. In addition, augmenting our environment with specialised equipment is not very practical nor cost-effective. The next generation of positioning systems must be able to utilise a variety of sensor modalities and utilise the existing infrastructure in order to become truly ubiquitous. The systems in this category provide the back-bone of the more advanced and accurate positioning systems that we are going to discuss next.

2.3 Fusion of id-linked Sensor Streams

Using ubiquitous technologies like WiFi for localisation has received a lot of attention from the research community over the past years. In particular, WiFi fingerprinting provides a reasonable solution for many applications that require low-cost and room-level accuracy.

However, the main disadvantage of fingerprinting is the labor-intensive survey required for building the fingerprint database. Because of this problem many techniques have been proposed [69, 41, 42] that can automatically build the fingerprint map using sensor fusion. More specifically, the recent LiFS system [42] uses the accelerometer on a user's mobile phone and a floor-plan to automatically build the fingerprint database. In the training phase LiFS collects WiFi RSSI measurements from various locations as users walk inside a building carrying smart-phones where their step count is also recorded using the accelerometer. LiFS then exploits the number of steps between two endpoints along a user's trajectory in combination with the floor-plan to establish the geographical relationship among the radio fingerprints in order to build the radio map. During the online phase LiFS uses the radio map generated in phase one to perform fingerprinting localisation. LiFS reported an average localisation error of 5.8 meters and a room-level localisation error of 11%.

Seitz et al. [70] use a Hidden Markov Model (HMM) to combine WiFi fingerprinting with dead reckoning in order to provide more precise and robust localisation. Given the fingerprinting map, this algorithm models the locations of the WiFi fingerprints as the hidden states of the HMM and the RSSs as the observations. Given an initial probability distribution of the user's location the algorithm uses step length and heading information obtained from the IMU to hop among the HMM states. The RSS measurements are then compared to the ones stored in the fingerprinting map to recover the hidden state and thus find the location of the user. The authors reported a localisation accuracy of 5 meters during their experiments in indoor environments.

Other techniques [71, 38] augment the WiFi fingerprint database with information from FM signals to improve the localisation accuracy. Most recently Chen et al. [38] combined WiFi and FM signals to built robust and discriminative wireless signatures for indoor localisation. These authors observed that the FM signals compared to the WiFi ones are less susceptible to human presence, multi-path and fading while they exhibit exceptional indoor penetration. The internal structure of the building greatly affects the propagation of FM radio signals and by exploiting this fact more accurate localisation can be achieved. The authors designed multi-dimensional fingerprints based on FM-RSS, FM-SNR and FM-multipath and WiFi RSSI and reported a room-level accuracy of 98%.

2.3.1 Discussion

In this section we have presented some of the most popular fusion-based positioning systems that utilise id-linked sensor modalities. It is obvious that by using more that one sensor

modality we can build more accurate positioning systems and reduce the tuning and maintenance effort. However, these systems cannot be applied in all the application scenarios and environments. For instance, not all buildings provide adequate WiFi infrastructure thus WiFi based approaches cannot be used reliably. On the other hand the techniques discussed above are designed to use id-linked data only and cannot be applied in cases where anonymous data (i.e. camera feed) are available. In fact many environments i.e. public spaces, large buildings and industrial settings are equipped with CCTV infrastructure for security purposes which can also be used for positioning tasks. Systems that leverage id-linked and anonymous sensor streams are under-explored as of today.

2.4 Data Association Methods for Multi-Target Tracking

The objective of multi-target tracking (MTT) is to collect sensor data from a field of view (FOV) containing multiple targets of interest and potential background clutter and then find the trajectory of each target and filter out the clutter.

If the sequence of measurements associated with each target is known (i.e. id-linked sensor modalities are used) then the MTT reduces to a state estimation problem (e.g. Kalman/particle filters can be used to follow each target). However, when the target-to-measurements association is unknown (i.e. anonymous sensor modalities such as cameras, radars and sonars are used) the data association problem must be solved in addition to state estimation. Essentially the data association problem seeks to find which measurements correspond to each target. In this section five data association methods are discussed which are commonly used in MTT systems.

2.4.1 Nearest-Neighbour Data Association Filter (NNDAF)

NNDAF [72, 73] is probably the simplest solution to the problem. Using this filter each measurement is associated with the target whose predicted position is closest. After that the filter uses the assigned measurement to update the target's state. Since a single measurement may be the nearest neighbour to more than one target certain NNDAF implementations allow this measurement to be used to update more than one track. Other implementations follow a greedy approach where assignments are ranked and processed sequentially. These approaches do not allow a single measurement to be used more than once. The advantage of this filter is obviously its simplicity both conceptually and computationally. However, this filter is only applicable to simple scenarios and it can break down easily in cases with ambiguities and clutter.

2.4.2 Global Nearest-Neighbor Data Association Filter (GNNDAF)

A method closely related to the previous one is the GNNDAF [56, 55]. This method determines a unique assignment so that at most one observation can be used to update a given track by considering all possible measurements-to-track pairings. A cost or likelihood function is used to evaluate all possible measurement-to-track combinations which are then used as input to the well-known assignment problem. The best assignments can be then calculated efficiently with the Hungarian algorithm [74, 75, 76].

2.4.3 Probabilistic Data Association Filter (PDAF)

The two previous methods allow at most one measurement to be used to update a given track and they usually do not permit a measurement to be used more than once. Such methods are termed unique-neighbour methods. In addition, they use cost or likelihood functions to evaluate the measurements-to-track pairings. However, in cases of densely moving targets and severe clutter many measurements-to-track combinations may have similar likelihoods which leads to many association errors from which these methods cannot recover.

In order to compensate for this problem the all-neighbours data association methods update the track's position by using all the measurements within a specific region (called gate) about the predicted track's location.

The PDAF [77], is an all-neighbours method which assigns probabilities to the measurements found within a track's gate and then performs averaging to update the track's position based on all measurements. Given N measurements within the gate of a track, PDAF forms $N + 1$ hypotheses (i.e. H_0: none of the measurements is valid, H_j: the j_{th} measurement is valid, etc.). For each hypothesis (i.e. association) a probability is being assigned governed by the hypothesis quality with respect to the track's predicted location. The filtering algorithm uses these probabilities in a weighted sum to update the position of the target. The PDAF has shown to be very effective in situations with severe clutter for one target outperforming both NNDAF and GNNDAF.

2.4.4 Joint Probabilistic Data Association Filter (JPDAF)

Despite the great success of PDAF in cluttered environments, later results showed that this filter breaks in dense target environments and so it was extended to JPDAF [78]. JPDAF has the same principle of operation as PDAF however the association probabilities are calculated using all observations and all tracks. Essentially, JPDAF assumes that each target generates only one measurement and thus enumerates all possible measurements-to-track associations and computes association probabilities. Similarly to PDAF, given an

association the filtering algorithm estimates the state of the target and then this estimate is weighted by the corresponding association probability. The final target position is then calculated as the weighted sum of all estimates. Even though JPDAF performs significantly better that PDAF both filters have a tendency towards track coalescence for closely spaced targets that share measurements in their gates.

2.4.5 Multiple Hypothesis Tracking (MHT)

The multiple hypothesis tracking (MHT) [79, 80] is another popular data association method. MHT is a deferred decision logic method which maintains multiple track hypotheses associating past measurements with targets. In MHT alternative data association hypotheses are formed whenever there are measurement-to-track ambiguities. Similarly to NNDAF and GNNDAF, MHT is also a unique-neighbour method. However, the measurement-to-track association decision is postponed until enough measurements are collected that can be used to resolve the association ambiguities. The performance of MHT is superior from all the techniques discussed previously. However, the main disadvantage of MHT is its computational complexity since the number of hypotheses grows exponentially over time. Various techniques including the formation of only the $k-$best hypotheses, hypothesis pruning and hypothesis merging have been developed in order to avoid the combinatoric explosion.

2.5 Positioning with Anonymous Sensor Streams

As we have explained in the previous section when the measurements-to-targets association is not known (i.e. the system uses anonymous measurements for tracking) the data association problem must be solved prior to the state-estimation (i.e. tracking). Systems that use anonymous sensor streams from cameras, sonars, passive IR use data association techniques to find the association of measurements to targets. In this section we will give a brief overview on visual tracking (i.e. the anonymous observations are camera measurements such as visual coordinates) which is more relevant to this thesis.

The basic task of visual tracking is to estimate the trajectories of one or more interacting targets from a sequence of images. Visual tracking algorithms [81, 82] typically consist of two components: a) an object detector and b) a tracker. The purpose of the detector is to detect the objects of interest within a given frame. The tracker on the other hand is responsible for establishing the correspondences of the detected objects across all frames. Kalman filters, in particular, have been widely applied in visual tracking scenarios where the process and measurement models are linear and noise sequences are Gaussian [83]. When the

process and measurement models are nonlinear, but the posterior state density is still modelled as Gaussian, extended / unscented Kalman filters are used instead. Hidden Markov models (HMMs) [84] are typically applied when the state space is discrete and we have known transition probabilities between states. Particle filters are also very popular for target tracking [85], and represent a more general class of filters, in which the current density of the state is represented by a set of random samples with associated weights. In multiple target scenarios the data association problem must be solved prior to the position estimation thus the Joint Probabilistic Data Association Filter (JPDAF) [78] estimates measurement to target association probabilities across all targets. The Multi-Hypothesis tracking approach (MHT) [79] estimates the conditional probability of a measurement given a target. Further work aims at tackling the multi-target tracking problem in the presence of occlusions and split merge conditions using track linking approaches [86]. To solve this problem efficiently, in [36], the authors map the data association problem into a cost-flow network, and solve it with a min-cost flow algorithm. The authors in [37] formulate the data association as a generalised minimum clique problem (GMCP). More specifically, under the GMCP framework the detections across all frames are first grouped into disjoint clusters and then a complete weighted undirected graph between detections belonging to different clusters is formed. The weights on the edges of this graph define the similarity between a pair of detections according to appearance and motion features. To find the trajectory of one person an optimisation problem must be solved to find the subset of detections with the minimum cost (sum of weights) by selecting exactly one detection from each cluster. The found detections are then removed from the graph and the optimisation problem is repeated until all people's trajectories are found. Also, multi-target data association with higher-order motion models has been proposed in [87]. Poor visual features, crowded environments and occlusions greatly affect the accuracy of the above methods.

2.6 Positioning using Id-linked and Anonymous Sensors

These positioning systems combine id-linked and anonymous sensor modalities. In this section we will present the state-of-the-art positioning systems that use id-linked and anonymous sensor modalities.

2.6.1 Fusion of Camera and Inertial Data

A first technique that fuses accelerometer (entity-linked) and camera (entity-unknown) traces for people identification and tracking is presented in [44]. This system fuses motion

traces obtained from one stationary camera with motion information from wearable accelerometer nodes to uniquely identify people in the FOV using their accelerometer node IDs. Background subtraction is used to detect people from the video footage and then their floor-plane acceleration is extracted by double differentiation. The camera acceleration traces are then compared against the overall body acceleration obtained from the accelerometer nodes using the Pearson's correlation coefficient. The acceleration correlation scores between all possible combinations of camera-accelerometer pairs are then used to form an assignment matrix. Finally, the assignment problem is solved using the Hungarian algorithm in order to find the association between camera and accelerometer measurements and thus identify the people in the video with their accelerometer node ID.

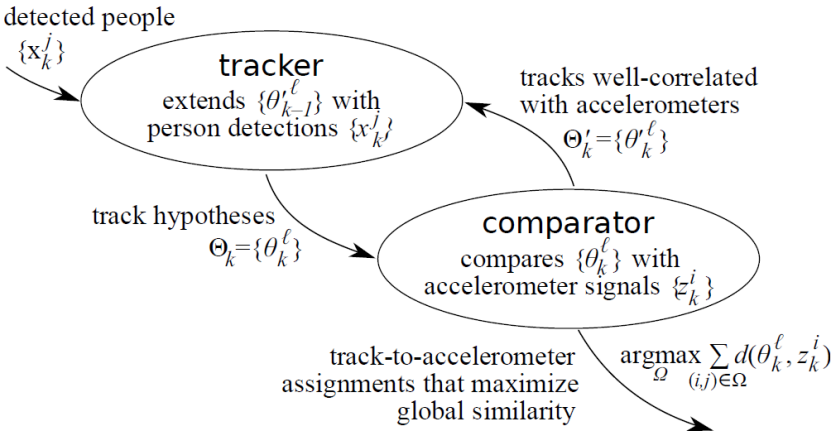


Figure 2.1: Overview of the algorithm in [44] that is used to identify and track people using accelerometer and camera traces.

In order to handle motion ambiguities the authors use a data association filter containing two modules as depicted in Figure (2.1). In the first step they use a tracker which is responsible of assigning visual detections to tracks. At each time-step k whenever there are multiple detections within a track's gate the tracker spawns multiple hypotheses. Then the comparator module takes all hypotheses Θ_k and the accelerometer measurements and solves the assignment problem using the Hungarian algorithm. The output of the Hungarian algorithm (i.e. assigned hypotheses to accelerometer data) is then passed back to the tracker which uses this information to filter out the wrong hypotheses. To evaluate their method the authors captured five 1-minute videos and the corresponding accelerometer traces of a single person walking in a room. Then they used these data to emulate multi-target scenarios. The disadvantage of this method is that it cannot distinguish among people with similar acceleration patterns. Finally, the authors have not tested their method in realistic

and dense target environments and there is no indication of how this method will perform in such cases.

Subsequent work [45] extends the previous method to a multi-camera system. The authors fuse information from wearable inertial sensors (i.e. accelerometer and magnetometer) with camera traces to identify multiple people in the FOV. For each person in the FOV the authors use a hidden Markov model (HMM) with two observations. The first observation (i.e. motion measurements) is obtained from the IMU and includes the person's ID but not his/her location. The second observation (i.e. user's motion) comes from the camera network and contains the person's location but not his/her ID. Then an optimisation problem is solved in order to find the best matching pairs between the two observations among all possible combinations of IMU and camera measurements. Once the matching pairs are identified (i.e. association between the IMU and camera observation) the HMM state (i.e. location) can be recovered and thus the location and identity of each person are obtained. The overall identification accuracy of this method was 90%, however the evaluation was done in rather ideal conditions. The authors captured 15 1-minute camera and IMU traces from a single person walking in a room, and they used these data to create a multi-target scenario consisting of 1 to 4 targets. The advantage of this method compared to their later work ([44]) is that instead of using only acceleration data they also incorporated information from the magnetometer (i.e. yaw) which made it possible to distinguish between similar acceleration patterns.

More recently the OPTIMUS system [88] uses a similar approach, where inertial measurements from smartphones are used to identify visual trajectories. The algorithm uses histograms of oriented gradients (HOG) descriptors to detect people in a scene covered by one stationary camera and then optical flow is used to group consecutive detections together when there are no ambiguities forming tracklets. A track-level association procedure is then performed to merge the found tracklets and create full trajectories. At the identification stage accelerometer readings from the smartphones are converted into binary vectors that indicate the user's movement and are matched with movement vectors extracted from the visual trajectories using the Hungarian algorithm.

Unlike our work, the methods described above use inertial data mostly for identification purposes; they are not used for identification and positioning. In addition, these methods are specifically designed having only one sensor modality in mind and thus they do not provide a general multi-sensor multi-target tracking framework.

2.6.2 Fusion of Camera and Radio Data

Given a number of people walking inside a scene with camera coverage the recent EV-Loc [43] system showed how to localise and identify each individual by fusing WiFi signals from their mobile devices with visual traces from the video footage. The architecture of EV-Loc (Figure (2.2)) consists of four major modules namely data collection, signal processing, electronic and visual signal matching and signal fusion.

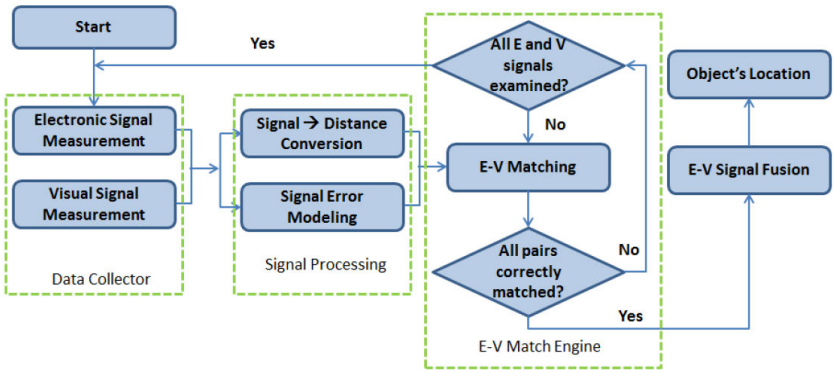


Figure 2.2: Overview of the EV-Loc architecture [43]

At each time step the EV-Loc system first collects synchronised electronic signals (i.e. RSS WiFi measurements) from people's mobile phones and visual signals (i.e. images) from a stationary camera. In the next step the distance of each person from the WiFi access points is estimated using both the electronic and visual signals. The signal processing module uses object detection techniques to detect the people in the image and calculate their distance from all APs. Also, a pre-trained radio propagation model is used in this step to convert the RSS measurements to distances. After all electronic and visual signals have been converted into distance measurements the algorithm calculates the Euclidean norm between all pairs of visual and electronic distance estimates. Then the E-V match engine uses this information to find the best matching between electronic and visual signals by the Hungarian algorithm. Finally, when the associations between electronic and visual signals are established the location of each person is calculated from the weighted average of the corresponding electronic and visual distance estimates. The above procedure is used when the targets are stationary. However, when the targets are moving inside the camera FOV, EV-Loc first tries to establish the correspondences of the detected targets across all frames (i.e. data association) using only visual signals and when this happens the visual trajectories are converted into distance measurements and the previous procedure is applied.

The authors performed 10 indoor and 10 outdoor experiments in order to evaluate their system. In each case 15 visual and electronic frames were captured and five stationary targets were present in the FOV. The system reported a 90 percentile distance error of 0.47m and 0.91m for the indoor and outdoor experiments respectively.

Compared to our proposed approach EV-Loc concentrates on the problem of matching N wireless devices with N visual detections, rather than inferring the trajectory of a device, using both the device's radio data and anonymous camera detections from the scene. Unlike our work, they assume a known, calibrated radio model, and have performed tests in rather ideal conditions with static targets.

Also Mandeljc et al. [52] recently proposed a fusion algorithm that incorporates radio data into the camera-based probabilistic occupancy map (POM) framework. The POM [89] is a multi-camera framework for people detection and localisation in which the area of interest (i.e. the ground plane as viewed by the cameras) is divided into a number of cells forming a grid. Under this framework humans are represented as simple rectangles, and detections are generated using background subtraction techniques. The algorithm then models each cell of the grid as random variables representing the probability of a cell being occupied by a person. Finally, the goal of the algorithm is to estimate the probabilities of occupancy for each cell given binary images obtained from a background subtraction process from multiple overlapping cameras.

In [52] the authors extend POM to non-visual sensor modalities providing an illustrative example using Ubisense's ultra-wideband radio sensors. More specifically, time-synchronised UWB receivers are deployed in the scene and transmitters (tags) are placed on the objects to be tracked. Their algorithm estimates the probability of each cell being occupied both with the cameras and with UWB using time-of-arrival measurements. They performed an experiment in an office room where 4 cameras and 4 UWB receivers were deployed and they evaluated the localisation performance of their system. A six-minute experiment was conducted involving 3 individuals walking in a 7.1m by 6.9m room. The distances between the estimated and ground truth positions were calculated and a 95 percentile error of 0.73 m, 0.48 m and 0.40 m was reported for UWB positioning, Camera positioning and fusion positioning respectively.

Unlike our work, the radio measurements in [52] are only applied to the sub-problem of estimating a ground plane occupancy model (that gives the probability of a cell being occupied by any of the humans given sensor data) and they are not directly used for identification (i.e. to estimate the probability of a particular person being located in a cell given that the cell is occupied). In addition, the authors claim that this framework can use

arbitrary sensing modalities. However, sensors like WiFi do not possess the ranging capabilities or the precision required to give meaningful information about the probability of occupancy of a particular cell (i.e. 20x20 cm), and thus the sensor fusion with WiFi will probably not improve the performance of the system.

More recent work [90] added a second fusion stage to [52], where anonymous detections are augmented with identity information from radio tags. The cost of mapping an anonymous detection to a radio-based identified detection is evaluated based on the Euclidean distance between the two, and the optimal assignment between two sets of detections in a frame is evaluated using the Hungarian algorithm. The systems presented in [52] and [90] require multiple cameras with overlapping FOVs and expensive state-of-the art ranging equipment which must be carefully deployed.

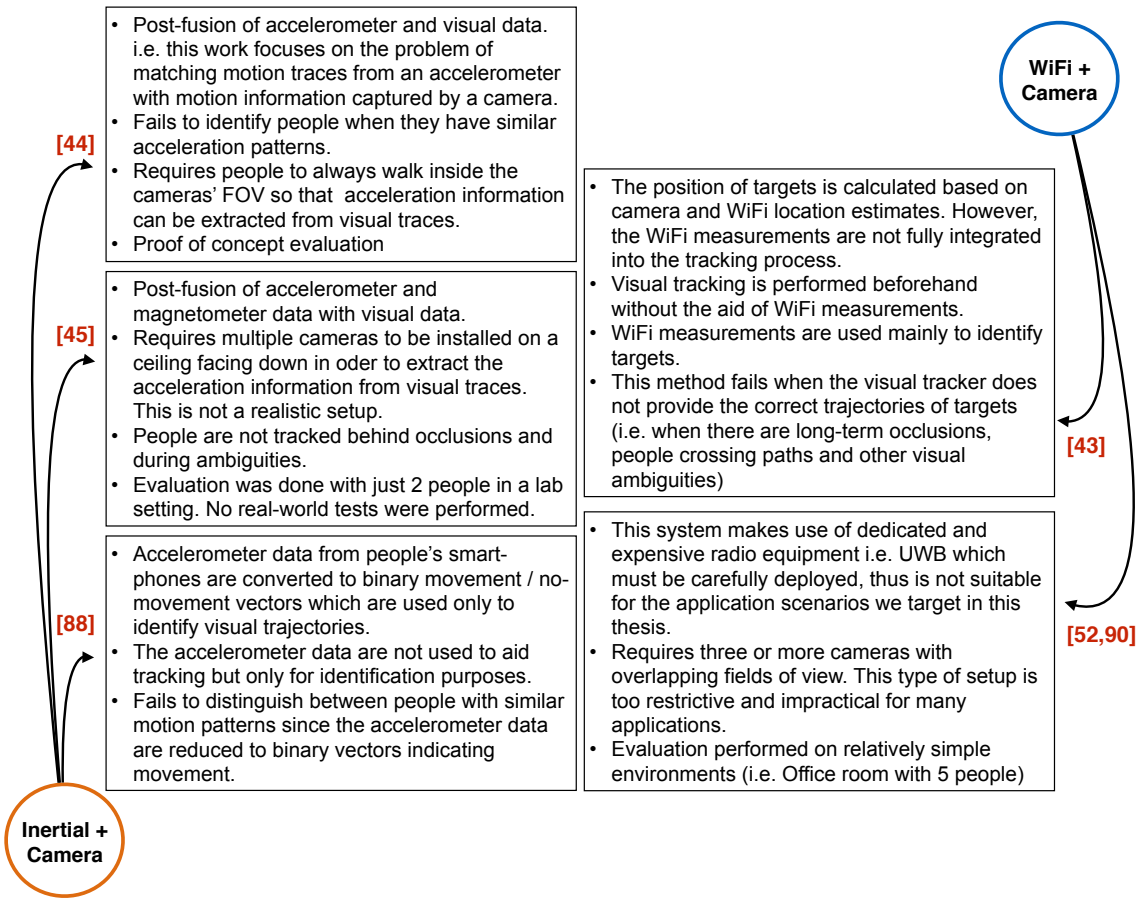


Figure 2.3: Limitations of existing techniques

To conclude this section we summarise in Figure (2.3) the most recent work on positioning which fuses id-linked and anonymous sensor modalities. In this figure we have outlined the limitations of each method which motivates the work of this thesis.

2.6.3 Discussion

In this section we have presented the state-of-the-art positioning systems that fuse id-linked and anonymous sensor modalities. Currently the most promising and cost-effective positioning systems use ubiquitous sensors (i.e. WiFi, IMUs, etc.) that do not require the deployment of additional infrastructure. Unfortunately, none of them can meet the accuracy capabilities of the high-end systems. Research that has tried to merge id-linked and anonymous sensor modalities is still at its infancy and more investigation is required in this direction to build a robust and accurate positioning system. The most relevant positioning systems to this thesis are tailored for specific applications and they do not provide a general multi-sensor multi-target tracking framework.

The recent positioning systems that fuse camera and inertial data (i.e. [44, 88]) concentrate on the problem of matching a set of camera observations with a set of inertial observations in order to identify people. They do not use inertial measurements to improve the overall tracking performance of their systems. For instance, inertial measurements are not used to aid the prediction of human motion in real time or to maintain tracking under long-term occlusions.

Moreover, the state-of-the-art systems than fuse radio and camera measurements lack important features that are required for todays' applications. For instance, step-by-step fusion is lacking in [43]. In other words WiFi is used after the visual trajectories have been created (i.e. post-fusion). As a result the WiFi measurements are not fully integrated into the tracking framework thus they cannot be used to help resolve visual ambiguities and make the system more robust to missing detections and people crossing paths. In addition, the system described in [90] uses expensive dedicated transmitter-receiver equipment and multiple cameras with overlapping fields of view which makes it very difficult and impractical for use outside the lab setting.

In addition, none of the systems mentioned above has been tested in real-world scenarios. The most relevant positioning systems to this thesis have been evaluated in relatively simple environments (i.e. office room) with weak assumptions (e.g. no occlusions, linear human motion, etc).

Motivated by the limitations of the existing techniques we are going to show in this thesis how to use the existing infrastructure (e.g WiFi, cameras, inertial) to build an accurate and cost effective positioning system for multiple-target tracking which is flexible and robust and can be used in a plethora of scenarios. In the next chapter we will show how radio (e.g. WiFi) measurements can be integrated deeply into the multi-target tracking problem and help to resolve tracking challenges such as visual ambiguities, occlusions and missing detections.

Chapter 3

Tracking with Camera and Radio Data

3.1 Introduction

In this chapter, we propose a novel positioning system, RAVEL (Radio And Vision Enhanced Localisation), which fuses anonymous visual detections captured by widely available camera infrastructure, with radio readings (e.g. WiFi radio data). Our fundamental observation is that many applications that require high positioning accuracy are in large public or commercial spaces, such as airports, shopping centres, museums and industrial plants. These large public spaces are typically extensively covered with CCTV cameras for reasons of safety and security. We propose the use of existing camera infrastructure for the originally unintended task of indoor positioning. However, although camera based tracking can provide excellent position information in ideal conditions, the challenges provided by a real deployment are numerous. In particular, most security cameras are installed to provide a large field of view, typically resulting in a bird's eye view of the scene. This top-down perspective makes it difficult to distinguish facial features and accurate identification is made even more challenging when the room is not well lit (e.g. in museums). Furthermore maintaining tracking as people move behind obstacles, exit the field of view, or cross paths is an exceedingly difficult task.

Instead of trying to overcome the limitations of camera-based tracking through increasingly sophisticated vision-based algorithms, in this chapter we will show how we exploit opportunistic and ubiquitous radio signals (e.g. WiFi/Bluetooth Low Energy). Although, radio-based tracking is limited in terms of accuracy, it can be used to add context to trajectories obtained from camera based tracking. At a coarse level, this provides an identifier, which can be used instead of advanced video processing techniques like face recognition. At a finer level, the sequence of radio signal strengths, albeit noisy, can be used to disambiguate between multiple possible trajectories, aiding to merge and split discontinuous traces.

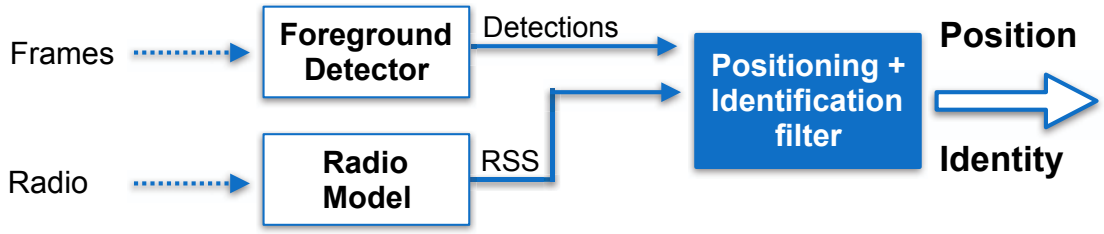


Figure 3.1: The figure shows a high-level overview of the proposed system architecture. Our system uses background subtraction techniques (denoted as foreground detector in this figure) to detect moving people inside the camera's FOV. The positions of detected people are then fused with radio measurements (i.e. WiFi RSS signals) from people's smart-phones to guide the tracking process via a multiple-hypothesis tracking framework. The "Positioning + Identification filter" provides the trajectories and identities of all people carrying smart-phones.

Motivated by the noisy nature of both visual and non-visual sensor modalities, we explore whether we can effectively combine them to overcome each other's weaknesses. We present a generic vision+radio tracking framework, RAVEL that can be used both in receiver-centric (i.e. people carrying smartphones) and transmitter-centric (i.e. low-cost tags in warehouse or industrial sites) applications. More specifically, we focus on the following questions: Can we make use of existing camera infrastructure to provide people with an accurate positioning system? Can personal electronic data (i.e. WiFi data from a person's smartphone) be used to resolve tracking ambiguities in visual data? How should we fuse anonymous visual and personal radio data to make the most of the two modalities? What are the benefits of combining these modalities in practical settings?

3.2 Problem Definition

Let us assume the indoor environment is monitored by a calibrated (known extrinsic and intrinsic parameters) stationary camera. For a given time window of size W, $W \in \mathbb{Z}$, the camera captures a series of frames $[f_1, ..., f_W]$ within its field of view (FOV). We assume that each frame f_i contains a number of camera detections of moving objects within the FOV, denoted as $C_i = \{c_i^1, ..., c_i^j, ...\}$, $1 \leq i \leq W$, $1 \leq j \leq |C_i|$. A camera detection c_i^j is represented as a bounding box of the detected object, or simply by box center's coordinates, i.e. $c_i^j = (x_i^j, y_i^j)$. We also assume that at each time i, the mobile device carried by a particular user *receives* a set of radio measurements $r_i = \{r_i^1, ..., r_i^m, ...\}$, where r_i^m is the Received Signal Strength (RSS) measurement of the m-th radio basestation at time i.

The problem is how to *estimate the trajectory of a user given the sequence of anonymous camera detections* $[c_1, ..., c_W]$ *and personal radio measurements* $[r_1, ..., r_W]$.

Figure (3.1) shows a high-level diagram of the problem we are trying to solve in this chapter. The foreground detector module processes a series of camera frames $[f_1, ..., f_W]$ and extracts the locations of all moving targets within those frames (i.e. camera detections). These detections along with the radio measurements of a particular user $[r_1, ..., r_W]$ are then fused inside the "Positioning + Identification filter". This module is responsible for the main tracking and identification process. In the following sections we highlight the contributions of this chapter, describe the architecture of the proposed system and explain the operation of each module.

3.3 Contributions

To the best of our knowledge, this is the first attempt that proposes a practical solution of radio-aided visual tracking, and tests it in a truly complex and realistic scenario. Specifically, our contributions are:

- We provide a fresh perspective on the problem of low-cost high-accuracy positioning in large open-plan indoor spaces, enabled by the fusion of anonymous visual data and radio data.
- We formulate the problem of tracking people by combining visual- and radio-based positioning modalities. We highlight the key challenges that we have encountered in a complex indoor museum environment, and explain why existing approaches are not designed to cope with reality.
- We design a novel multi-hypothesis probabilistic approach (RAVEL) to fuse radio and camera data that is robust to noisy and incomplete measurements.
- We integrate the radio measurements into the visual tracking process as opposed to the competing techniques which use radio measurements mostly for identification. This allows us to resolve motion ambiguities when people cross paths and maintain tracking as people move behind obstacles.
- We show how the radio propagation model for a particular environment can be learnt online, requiring no site-specific surveying.
- We evaluate the proposed approach in a museum setting, and compare it with the competing state-of-the-art approaches.

3.4 System Architecture

We are now in a position to present in more detail the proposed Radio And Vision Enhanced Localisation (RAVEL) system. The main components of RAVEL are: a) a *visual-based detector* and b) a *radio-aided tracker* as shown in Fig. (3.2). The *visual-based detector* is responsible for detecting all the moving people in the scene and for maintaining hypotheses about their locations. On the other hand the *radio-aided tracker* uses the private WiFi measurements of a person to select and merge the most likely location hypotheses in order to create the complete trajectory of this person.

We should note here that these components can run in the same device (e.g. a server collecting measurements from multiple transmitters) or across different devices (e.g. smart cameras running the visual-based detector and disseminating traces to mobile devices, each running their own radio-aided tracker). The rest of this chapter considers the latter, more challenging case of distributed tracking, but the algorithms presented are equally applicable to centralised tracking.

3.4.1 Visual-based Detector

The visual-based detector processes the captured camera footage as follows. For each frame f_i, a set of anonymous camera detections $C_i = \{c_i^1, ..., c_i^j, ...\}$ of the interesting objects (i.e. moving people) are firstly extracted. In our implementation, we use a lightweight MoG-based background subtraction approach [91] to detect moving people, which does not require heavy training and with some optimization it can run in real-time in embedded camera networks [92]. However, the computed set of detections C_i can be very *noisy* and *unreliable*. Fig. 3.3 shows a frame with three typical cases of noisy detections observed in our experiments, namely multiple detections from a single person (splitting); single detections corresponding to multiple people (merging); and null detections (empty bounding box, caused by a visual artifact such as a moving shadow). In this case, linking the detections into coherent trajectories is not a trivial task. Note that the visual-based detector only links detections into short trajectory segments (referred to as *tracklets* hereafter) when they unambiguously belong to the same target, as discussed in Sec. 3.5.1. The tracklets could be broadcasted to phones in a number of ways, including overloading the WiFi beacon frames (i.e. beacon stuffing) [93] which could preserve the user's privacy.

3.4.2 Radio-aided tracker

The tracklets generated by the visual-based detector tend to be very short, and are still *anonymous*: we have no knowledge of which tracklet should belong to a given user. There-

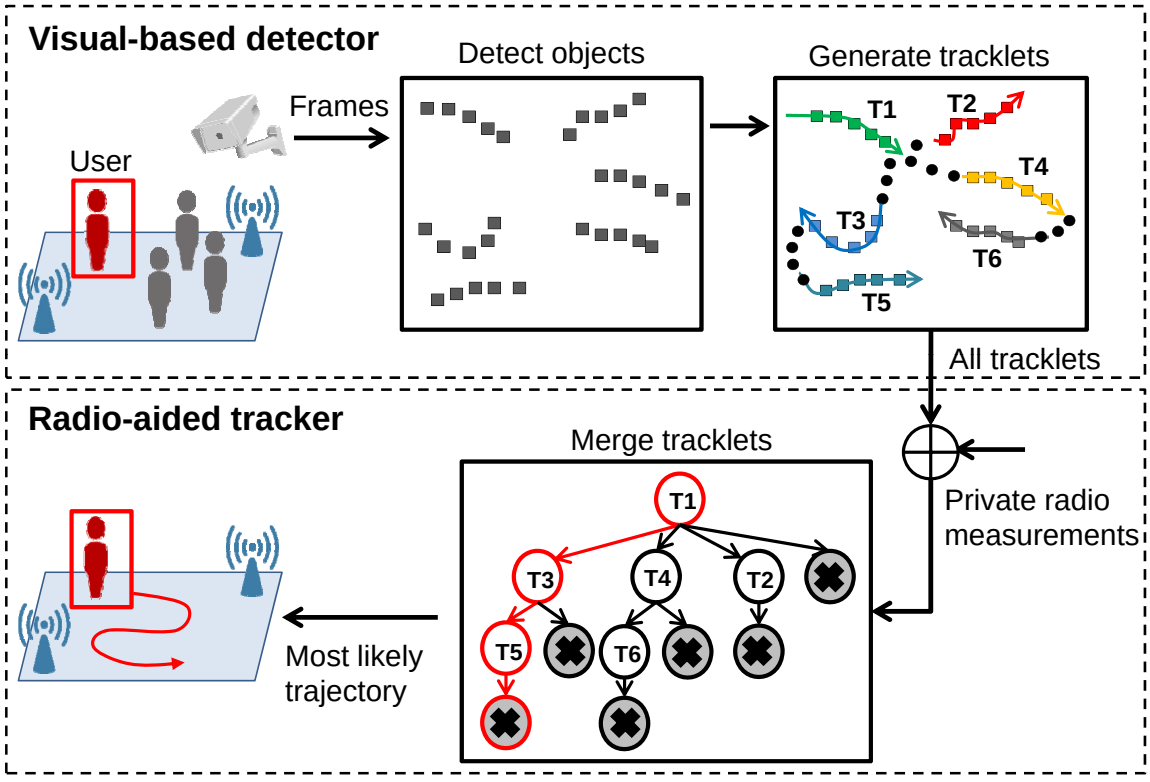


Figure 3.2: RAVEL is composed of two modules: a visual-based detector and a radio-aided tracker. The visual-based detector accomplishes two tasks. First it receives camera frames from the CCTV infrastructure and uses background subtraction techniques to detect the people in the scene. The problem is then to find the correspondence of detections across all frames. To do that it generates hypotheses about the people's locations across frames i.e. when a set of locations has high probability originating from the same user it generates small trajectories (called tracklets). Finally, the radio-aided tracker receives all the generated tracklets within a time window. It then cross references the users' private radio footprint, to determine which tracklets should be selected and merged to produce the full trajectory of this particular user.

fore, the proposed system broadcasts the generated tracklets (which are just streams of two dimensional coordinates) to the radio-aided tracker, which runs on the mobile devices carried by the users. With the received anonymous tracklets, the radio-aided tracker cross references the user's private radio footprint, i.e. radio measurements from the known bases-tations, to determine which tracklets should be selected and merged to produce the complete trajectory of this particular user. To achieve this, our tracker uses a multi-hypothesis tracklet merging approach. In a nutshell, it builds *tracklet trees* that encode all hypotheses of user paths. We are now in a position to present the details of the proposed algorithms.

Figure 3.3: Three cases of noisy detections generated by our visual-based detector: a) multiple detections are generated for one moving target (D1 and D2), b) a detection contains no moving targets at all (D4), and c) one detection contains multiple moving targets (D5).

3.5 Proposed Algorithm

In this section, we discuss the core algorithms of the RAVEL tracking system: the *tracklet generation algorithm* (Sec. 3.5.1) used in the visual-based detector, and the *tracklet merging algorithm* (Sec. 3.5.2), which is the key competency of our radio-aided tracker.

3.5.1 Tracklet Generation Algorithm

Given a fixed window of frames $[f_1, ..., f_W]$, and extracted sets of camera detections i.e. $[C_1, ..., C_W]$, the task of the tracklet generation algorithm is to link the elements of C_i-s into unambiguous trajectory segments, referred to as tracklets. The proposed algorithm leverages a realistic model of human motion to ascertain that a sequence of detections from consecutive frames belong to the same person with high certainty, and can thus be grouped together into a tracklet τ. It starts by considering detections in the first frame and proceeds in chronological order until all detections are grouped into a set of tracklets $\mathcal{T} = \{\tau_1, ..., \tau_N\}$.

More specifically, at the beginning, or when an existing tracklet cannot be further extended with new camera detections, a new tracklet is initiated with the first available (not-visited) camera detection, say c_i. To extend the tracklet with a second detection, we first check if there is a detection c_{i+1} in the next frame such that $\|c_i - c_{i+1}\| < \mathrm{DT}$ where DT is the maximum displacement of a target in the period between two consecutive frames. If there is no such detection, or if there are more than one such detections, c_i becomes a

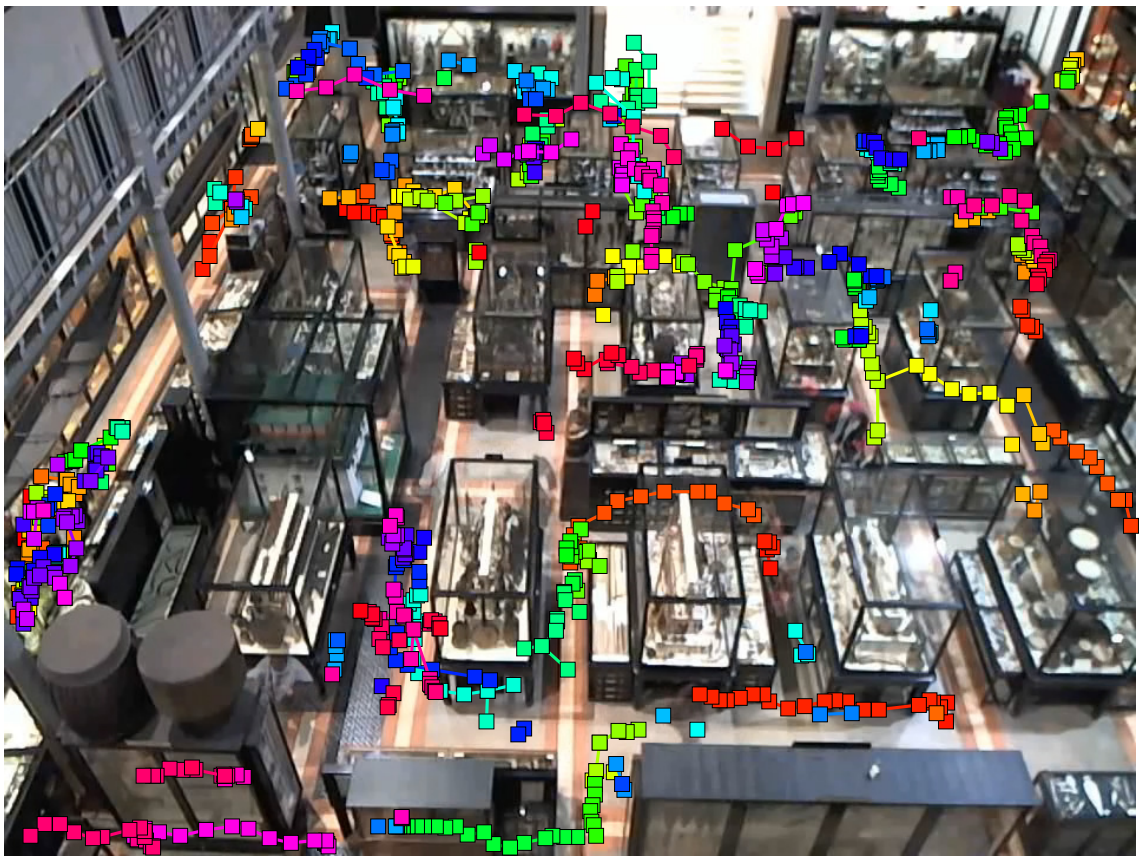


Figure 3.4: Tracklet generation: The figure shows the tracklets that have been generated by the proposed technique over a period of 120 frames. The squares indicate camera detections. Squares of the same colour indicate that these detections originate from one target with high confidence and so are linked together.

singleton tracklet. Otherwise the two detections must belong to the same person, so they are grouped together in the same tracklet.

Once we have at least two detections in the same tracklet, we can extend it with another detection taking into account the fact that humans normally avoid abrupt changes of direction and speed [94]. For example, consider the last two detections in the tracklet (say, c_i and c_{i+1}), and the potential to extend them with one of the available detections in the next frame, say c_{i+2}. In order to assess its suitability, we measure changes in the direction and in the speed of motion, and combine them to obtain a metric of confidence that the three detections concern the same person:

$$Q(c_{i+2}; c_i, c_{i+1}) = w_d Q_d + w_s Q_s \tag{3.1}$$

where Q_d is the cost of direction change while Q_s is the cost of speed change, weighted by

w_d and w_s respectively. In our case, Q_d and Q_s are defined as:

$$\begin{aligned} Q_d &= 1 - \frac{(c_{i+1} - c_i) \cdot (c_{i+2} - c_{i+1})}{\|c_{i+1} - c_i\| \|c_{i+2} - c_{i+1}\|} \\ Q_s &= 1 - 2\frac{\sqrt{\|c_{i+1} - c_i\| \|c_{i+2} - c_{i+1}\|}}{\|c_{i+1} - c_i\| + \|c_{i+2} - c_{i+1}\|} \end{aligned} \tag{3.2}$$

From the above formula, we can see that Q_d is actually the cosine of the angle between the displacement vectors $(c_{i+1} - c_i)$ and $(c_{i+2} - c_{i+1})$, which accounts for changes in the motion direction. On the other hand, Q_s is the ratio between the geometric and arithmetic mean of the magnitude of the displacement vectors. It measures the variation in the speed of motion and penalizes those detections that are very far away from the current tracklet (note that Q_s becomes 0 when the lengths of the two displacement vectors are the same). As shown above, the algorithm uses Eqn. (3.1) to evaluate the smoothness of motion between the last two detections of the current tracklet and a candidate detection in the next frame.

If at most one detection is found with cost Q less than a predefined threshold then we include that detection in the tracklet and continue to the next frame; otherwise we stop extending this particular tracklet. The above procedure is repeated until all detections are grouped into tracklets. This is shown in Fig. (3.4).

3.5.1.1 A note on human motion modelling

In the previous paragraph we have exploited the fact that people often avoid abrupt changes in their direction and speed and thus we have used Eqn. (3.1) to asses the smoothness of their motion for the purpose of human motion prediction. We have observed this type of human behaviour experimentally in a number of different scenarios (e.g. in museums and shopping malls).

This type of human motion modelling however cannot be used in all scenarios. In certain situations (e.g. in construction sites), human motion can be more complex and follows highly non-linear patterns with abrupt changes in the direction and speed. People are usually driven by an inner motivation towards some goal or task and their motion is often influenced by obstacles along their path and by the motion of other people [95]. Thus, motion models which make certain assumptions regarding the human motion [96, 94, 97] (e.g. smoothness of motion, constant velocity, linear-dynamics, etc) cannot be used. In Chapter 4 we investigate further the problem of human motion prediction and we show how to design motion models that can be used in scenarios where the human motion appears to be complex and non-linear.

3.5.2 Tracklet Merging Algorithm

Given the generated tracklets, the tracklet merging algorithm, which is the core of our radio-aided tracker, attempts to decide which tracklets should be merged together to produce the complete trajectory of the particular user. It first builds a set of tracklet trees that encode all possible trajectory hypotheses, and then searches for the trajectory hypothesis that is most consistent with the user's radio data.

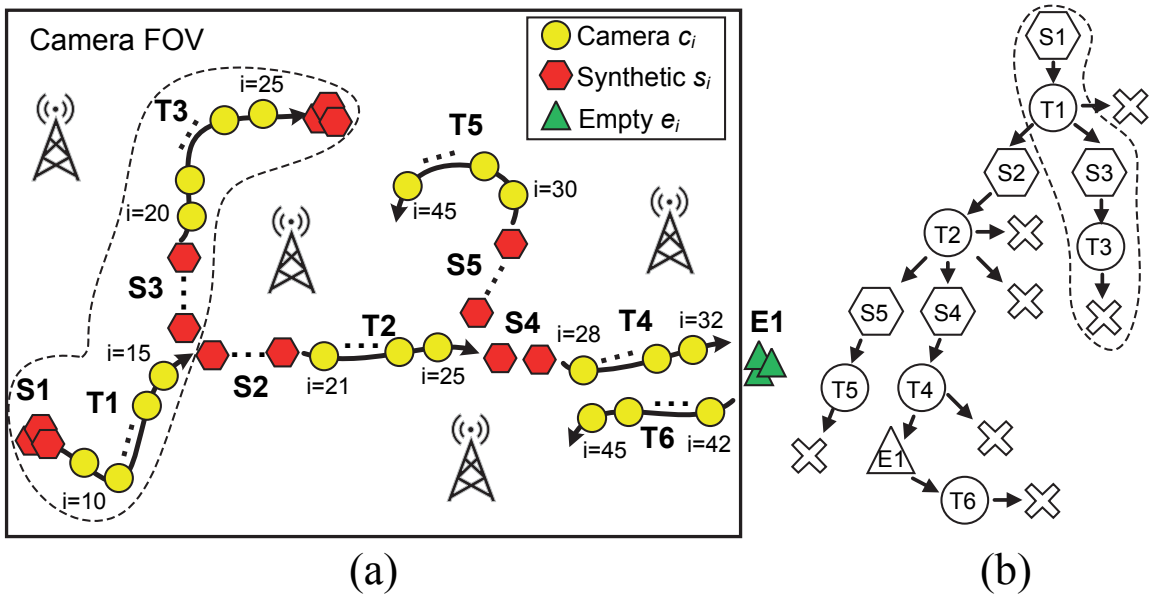


Figure 3.5: (a) The tracklets generated by our algorithm, where the gaps are filled by synthetic and empty detections. (b) The built tracklet tree with nodes containing camera (circle nodes), synthetic (hexagon nodes), and empty (triangle nodes) detections. Each camera node is also attached with a EOT node.

Build the tracklet trees: Let us denote the set of tracklets generated within a window of frames $[f_1, ..., f_W]$ as $\mathcal{T} = \{\tau_1, ..., \tau_N\}$. We first define $\mathcal{T}_{start}$ as the subset of tracklets in $\mathcal{T}$ that start early (before a certain time threshold) and thus could be used to start a user's trajectory. For each of those candidate tracklets, the algorithm initiates a tracklet tree. The algorithm then expands these trees as follows: for a given tree node (parent tracklet), its children become all the tracklets that start soon after it finishes and for which their first detection is at close proximity to the parent tracklet's final detection. This spatio-temporal threshold is a tuning parameter which constraints the size of the tracklet trees. Proper tuning would allow this module to run in real time on mobile devices.

Now the tracklets in $\mathcal{T}$ are topologically connected into paths, but these paths are still incomplete. For example, consider tracklets T1 and T3 in Fig. (3.5a), which form a path in the tree shown in Fig. (3.5b) (ignore the nodes other than T1 and T3 for now). For this path, we can see that: a) it does not start from the beginning of time window (T1 starts at time 10); b) it ends early (T3 ends at time 25 while the window size is 45); and finally c) there is a gap of missing detection in between (the final detection in T1 is at time 15, and the first detection in T3 is at 20).

To address this, we introduce two extra types of detections, *synthetic detections* $s_i \in S$ and *empty detections* $e_i \in E$, in addition to the *camera detections* $c_i \in C$. The synthetic detections are used to address the problem of missing camera detections when the user is actually within the camera's FOV, which are caused by occlusions or pauses of the user. On the other hand, the empty detections are introduced to account for situations where the user exits and possibly reenters the camera's FOV, and temporarily has no known coordinates.

Concretely, our algorithm completes paths with synthetic and empty detections according to the following rules:

- *First tracklet rule:* If the first tracklet of a path does not start at time 1 (of the window), then we precede it with empty detections if it starts at the boundary of the scene, or with synthetic detections (positioned at the location of its first detection) if it does not start at the boundary.

- *Gap rule:* If there is a gap between a parent and a child tracklet in the tree, we fill the gap with synthetic detections positioned at interpolated points between the final detection of the parent tracklet and the first detection of the child tracklet. In the unusual case that the parent tracklet ends at the boundary of the FOV and the child tracklet begins at the boundary, we fill the gap with empty (instead of synthetic) detections.

- *Last tracklet rule:* If the last tracklet of a path does not end at the end of the window, we extend it with empty detections if it ends at the boundary of the scene, or with synthetic detections (positioned at the location of its last detection) if it does not end at the boundary. We also add an end of trajectory (EOT) node to every node containing camera detections, indicating that the user may stop there.

After applying these rules, every path of the built tracklet tree represents a possible trajectory that contains W consecutive detections whether camera, synthetic or empty. For example, the highlighted path in Fig. (3.5b) corresponds to the highlighted trajectory in Fig. (3.5a), which indicates that the user moves from the bottom left corner of the FOV to

the top middle. We refer to such a trajectory as a *hypothesis*, denoted as $H = [h_1, ...h_W]$, where $h_i \in C \cup S \cup E$.

Search the tracklets trees: Once we have identified all trajectory hypotheses, we proceed to identify the most likely one. Specifically, the likelihood score of a hypothesis is defined as the sum of two parts:

$$L(H, R, \lambda) = L^v(H) + L^r(H, R, \lambda) \tag{3.3}$$

where $L^v(H)$ is the *visual-based likelihood score* of the hypothesis H, while $L^r(H, R, \lambda)$ is the *radio-based likelihood score* given observed radio measurements R and a radio propagation model λ. Therefore, our algorithm evaluates the likelihood score of a hypothesis in two steps: 1) it firstly estimates the user trajectory $X = [x_1, ..., x_W]$ based on vision data in hypothesis H, and evaluates the likelihood of vision data; 2) it then evaluates the likelihood of the radio data given the estimated user trajectory X. The first step is already applied by existing multiple hypothesis tracking approaches (e.g. [79, 80]), which neglect radio data and select the hypothesis that maximizes $L^v(H)$ only. We could of course derive the overall likelihood score in one step, using a Bayesian filter to jointly fuse vision and radio data. Our design choice to separate these steps stems from our desire to reuse existing implementations of vision-only trackers and extend them flexibly with a new step that additionally exploits radio data. Now we explain how $L^v(H)$ and $L^r(H, R, \lambda)$ are evaluated in detail.

Evaluate $L^v(H)$: For each hypothesis H, our algorithm maintains a filter, to estimate the trajectory X of the user. This can be implemented using a variant of a Bayesian filter, such as a Kalman filter [54]. Let x_i^- be the estimated position (state) of the user given the detections $h_{1:i-1}$, with covariance P_i^-. For a non-empty detection h_i, we define its incremental visual-based score as:

$$\Delta L_i^v = \begin{cases} \log[p_v f(h_i; x_i^-, P_i^-, \theta)] & , h_i \in C \\ \log[(1 - p_v) f(h_i; x_i^-, P_i^-, \theta)] & , h_i \in S \end{cases} \tag{3.4}$$

where p_v is the constant likelihood of having a camera detection when the user is in the FOV. f is a function that evaluates the likelihood of h_i given the estimated x_i^-, P_i^- and the parameters θ of the filter, i.e. it assesses how h_i agrees with the state estimated by the filter. Then $L^v(H)$ is the normalized sum of all ΔL_i^v:

$$L^v = |H|^{-1} \sum_{i=1}^{W} \Delta L_i^v \tag{3.5}$$

where $|H|$ is the number of non-empty detections in the hypothesis H.

Evaluate $L^r(H, R, \lambda)$: Let us assume that radio model λ in the indoor environment can be described with the log-normal shadowing model with parameters $\{P_0, n, \sigma\}$, and the Received Signal Strength (RSS) $r(a)$ at a point which is a meters to the known basestation is given by:

$$r(a) = P_0 - 10n \log_{10}(a) + \chi_{\sigma^2} \tag{3.6}$$

where P_0 is the RSS measurement at a reference distance of 1 meter, n is the path loss factor, and $\chi_{\sigma^2} \sim \mathcal{N}(0, \sigma^2)$ is the random shadowing variation. Let x_i^+ be the state estimated with detections $h_{1:i}$ from the previous step, which is the best guess on the location of a user at time i. For a non-empty detection h_i, we define its incremental radio-based likelihood score as:

$$\Delta L_i^r = \sum_{m=1}^{M} \log f_{\mathcal{N}}[r(dist(x_i^+, B_m)); r_i^m, \sigma^2] \tag{3.7}$$

where $dist(x_i^+, B_m)$ is the distance between x_i^+ and basestation B_m, and $r(dist(x_i^+, B_m))$ is the expected RSS value at x_i^+ (given by Eqn. (3.6)). $r_i^m \in r_i$ is the received RSS measurement of B_m on the user's mobile device. $f_{\mathcal{N}}$ is Gaussian probability density function. Therefore, the score of h_i is determined by how its corresponding state estimate x_i^+ fits the measured r_i. Then the score for the entire hypothesis H and radio model λ is the normalised sum of all ΔL_i^r:

$$L^r = |H|^{-1} \sum_{i=1}^{W} \Delta L_i^r + \log L_\lambda \tag{3.8}$$

where L_λ is the likelihood of radio model $\lambda = \{P_0, n\}$ given prior distributions on these two parameters. Our algorithm does not require perfect knowledge of λ, but attempts to learn the best model by a search through the parameter space. For simplicity, we assume the parameter σ is known, and only P_0 and n vary. In our experiments, we assume P_0 and n are independent, and follow the raised cosine priors governed by known hyper-parameters. Sec. 3.6 will show how in practice our algorithm finds the best radio model that is very close to the actual one.

To sum up, for each hypothesis H and possible radio model λ, our algorithm evaluates the likelihood score $L(H, R, \lambda)$ as shown above, and finds the best solution $\{H^*, \lambda^*\}$ which is given by: $\{H^*, \lambda^*\} = \arg\max_{H,\lambda} L(H, R, \lambda)$.

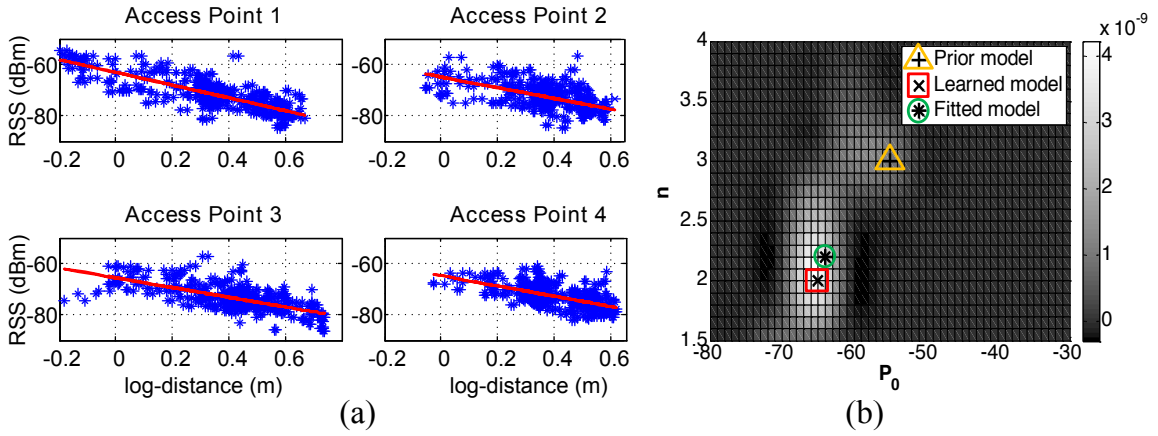


Figure 3.6: (a) Fitted radio model from 4 APs. (b) Learning the radio model parameters by searching the parameter space.

3.6 System Evaluation

3.6.1 Experimental Setup

We have conducted a real world experiment in a three-storey museum building to evaluate the performance of the proposed approach. In our experiment we placed the camera (i.e. off-the-shelf webcam) 10m above the ground in the third floor of the building facing down covering a 11m $\times$ 12m area. All people to be tracked were walking on the first floor while looking at the museum exhibits. The duration of our experiment was 40 minutes with the camera taking video at 2fps with a resolution of 960 $\times$ 720 px. The total number of people in the scene was varying as the museum visitors were entering and leaving the scene. The minimum number of people in one frame was 8 and the maximum 20 with 4 of them having WiFi enabled smartphones. The objective of the experiment is to determine the accuracy with which a person with WiFi data can track themselves in a busy environment with several other people. The WiFi measurements were taken by smartphones receiving beacons from 4 APs at a default rate of 2 samples/sec.

In order to obtain accurate ground truth trajectories from the video footage we supplied all four people carrying WiFi smartphones with hats of different colours. Then we used a mean-shift tracker [98] to track the coloured hats and label the ground truth trajectories of all 4 people for our entire 4800 frame dataset. The colour features were only used to acquire accurate ground truth; and both our approach and the competing approaches use only grayscale images. We noticed that in the absence of these distinctive hats, appearance features are not informative enough to tell apart one person from another, due to the dim lighting and the fact that with the downward-facing camera we did not get a view of

distinctive face/body features.

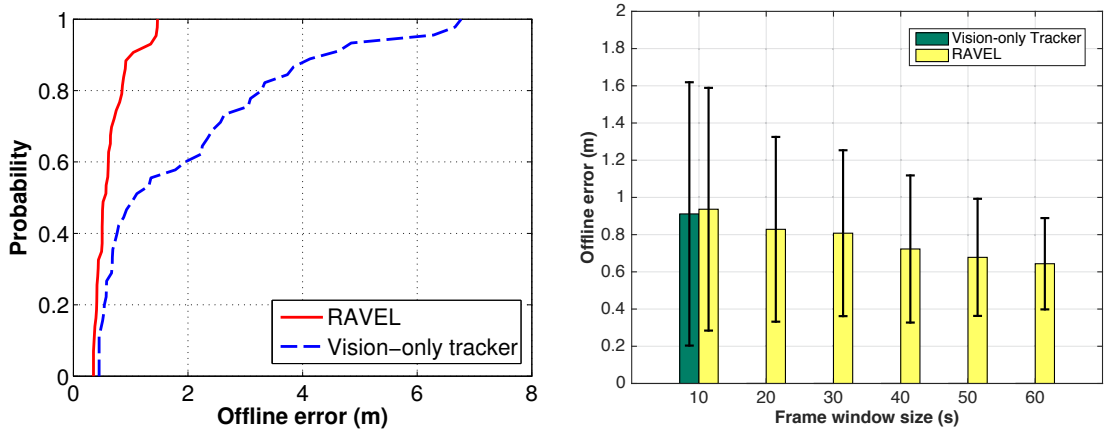
Algorithms: The competing approach, referred to as *Vision-only tracker*, only uses the smoothness of motion to connect visual detections into trajectories; it is a multi-hypothesis tracking approach widely used by the vision community [80, 79]. On the other hand, our proposed approach, *RAVEL* (Radio And Vision Enhanced Localisation) ravels out ambiguous visual threads by exploiting both the smoothness of motion and radio signal strength data. The state-of-the-art EV-Loc system [43] which uses WiFi and visual observations for positioning is also compared to the proposed approach next in the Evaluation section.

Performance metrics: We distinguish between two main operating contexts: 1) offline case, in which the competing/proposed algorithms are given data over a time window of size W, and are tasked to estimate the trajectory of a target during that period; and 2) online case, in which the task is to estimate the current location of the target given a window of historical data of size W.

For the offline case, we use two key metrics to evaluate our approach - offline location error and overlap error. *Offline location error* is an accuracy metric that reflects the distance between the estimated trajectory (using the competing or proposed techniques) and the ground truth trajectory (derived by using distinctive visual markers). In the results below, we measure it as the average Euclidean distance between the ground truth and estimated detections over time - other commonly used distance metrics are MSE (mean squared error), RMSE (root mean squared error), etc. In measuring the distance between trajectories, we only use timestamps (frames) for which both the estimated and ground truth trajectories report the target to be within the field of view.

The second offline metric, referred to as *overlap error*, measures the overlap between the ground truth and estimated trajectories with respect to the camera's field of view. Let FP (false positives) be the ratio of frames in a window of size W for which the estimated trajectory reports a detection in the FOV, whereas the ground truth trajectory suggests that the target is outside the FOV. Let FN (false negatives) be the ratio of frames in which the estimated trajectory reports an empty detection (outside the FOV), whereas the ground truth trajectory includes a detection within the FOV. The *overlap error* between the two trajectories is defined as the sum of false positive (FP) and false negative (FN) ratios, and it reflects the percentage of time in which the estimated algorithm misclassifies the target to be in the FOV when they are not or vice versa.

For the online case, our accuracy metric is referred to as *online location error*, and is the Euclidean distance between the last pair of detections of the estimated and ground truth trajectories (those at the end of the historical window).



(a) (b)

Figure 3.7: (a) Cumulative distribution function of the offline location error. (b) Impact of the frame window size on the offline location error

3.6.2 Results

3.6.2.1 No a priori knowledge of the radio model

The first set of experiments is conducted to validate our assumption that we do not require a priori knowledge of the parameters of the radio propagation model. Our proposed algorithm allows us to learn the model from camera and WiFi data, by considering different radio models (from a prior distribution) and choosing the one that maximizes the fitness function described in Eqn. (3.3). This is in contrast to EV-Loc which assumes a priori knowledge of the radio propagation model and obtains it by collecting training data from the environment (WiFi signal strength data at known distances from the access points). We refer to the radio model learnt from our proposed approach as *Learned model* and the model derived from training data as *Fitted model*.

Our proposed algorithm, RAVEL, assumes a non-environment specific raised cosine prior distribution on the two parameters (P_0 and n) of the radio propagation model (see Eqn. (3.6)), with a generous variance for each parameter. The prior distributions for P_0 and n are based on smartphone-based WiFi specs and a number of studies on typical values of n, e.g. [65]. Fig. (3.6a) shows the received signal strength measurements (600 samples are shown) as we vary the log-distance from four access points and the fitted ground truth radio propagation model. The fitted model ($P_0 = -64$ and $n = 2.2$) is obtained by averaging the P_0 and n values derived from the training data of the four access points.

Fig. (3.6b) shows that the fitness of the best hypothesis (given by Eqn. (3.3)) as we vary the parameters of the radio propagation model. Note that the Learned model, derived from the proposed RAVEL algorithm by maximizing the fitness of the best hypothesis, is very

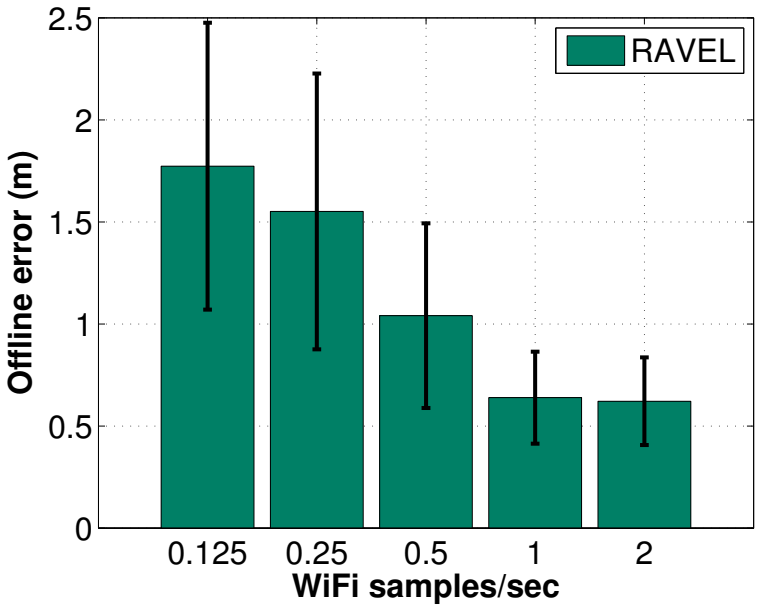
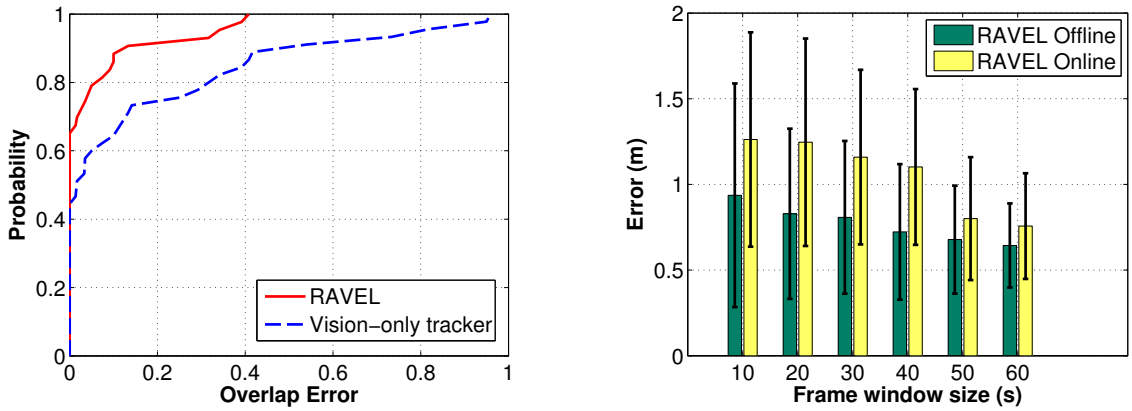


Figure 3.8: Impact of the WiFi sampling rate on the offline location error.

close to the *Fitted model* obtained from a wealth of training data (an expensive survey). More importantly this Learned model only assumes a rather broad and uninformative prior, the mode of which (referred to as *Prior model*) is quite far from the Fitted model. Hence, with very little prior knowledge of the radio propagation model, we are able to accurately infer it and avoid the time consuming step of environment-specific training. However, reliable prior knowledge of the radio model would constraint the search space thus reducing the power requirements of our algorithm making it more suitable for mobile devices and real-time processing.

3.6.2.2 Offline location error

The second set of experiments focuses on the offline case, and aims to compare the offline location error of RAVEL and the competing Vision-only tracker. We first use default values for the window size (120 frames taken over 60 secs) and for the WiFi sampling rate (2 WiFi signal strength values per sec), and plot the CDF of the offline location error over 160 windows (40 windows for each of the four people carrying a smartphone with WiFi). Fig. (3.7a) shows that RAVEL, which leverages WiFi information, achieves a median error of 0.56 m and a 90 percentile error of 1 m, significantly outperforming the competing Vision-only tracker, which has a median error of about 1 m and a 90 percentile error of 4.6 m. Note that to give a fair chance to the Vision-only tracker, we initialise it with the correct visual detection corresponding to the person that we are tracking. However, even with a



(a) (b)

Figure 3.9: (a) Cumulative distribution function (CDF) of the overlap error. (b) Online location error.

correct start the Vision-only tracker diverges from the correct path due to crossing, splitting and other ambiguities. These ambiguities are typically resolved by RAVEL with the help of WiFi data; illustrative examples of how this happens are provided later in the evaluation.

We then proceeded to examine the impact of the window size on the offline location error. We kept the WiFi sampling rate to the default value of 2 samples per sec, and varied the window size (10-60 secs corresponding to 20-120 frames). Fig. (3.7b) shows that the proposed RAVEL approach improves its accuracy as the window size increases. This is reasonable since different people will most likely walk different paths over long periods of time resulting in unique WiFi sequences which act as powerful discriminative signatures for each person. Thus as the frame window size increases the accuracy tends to increase. In Fig. (3.7b) we compare our approach with the best configuration for the Vision-only algorithm, i.e. the Vision-only algorithm achieves the best accuracy for the window size of 10s. This is because for long periods of time a lot of ambiguities are introduced especially in crowded environments which degrades the performance of Vision-only tracking.

The next step was to explore the impact of the WiFi RSS sampling rate on the offline location error of RAVEL with a default window size of 120 frames. As expected, Fig. (3.8) shows that the performance of RAVEL improves as we increase the sampling rate. Note that no significant benefits are observed by sampling above 1 WiFi RSS per second, which means that radio sampling does not need to be intensive (i.e. the algorithm can run on battery operated devices) to obtain accurate trajectories. This however may be also due to the fact that people walk relatively slowly in museums; it is possible that different environments may benefit from higher sampling rates.

3.6.2.3 Offline overlap error

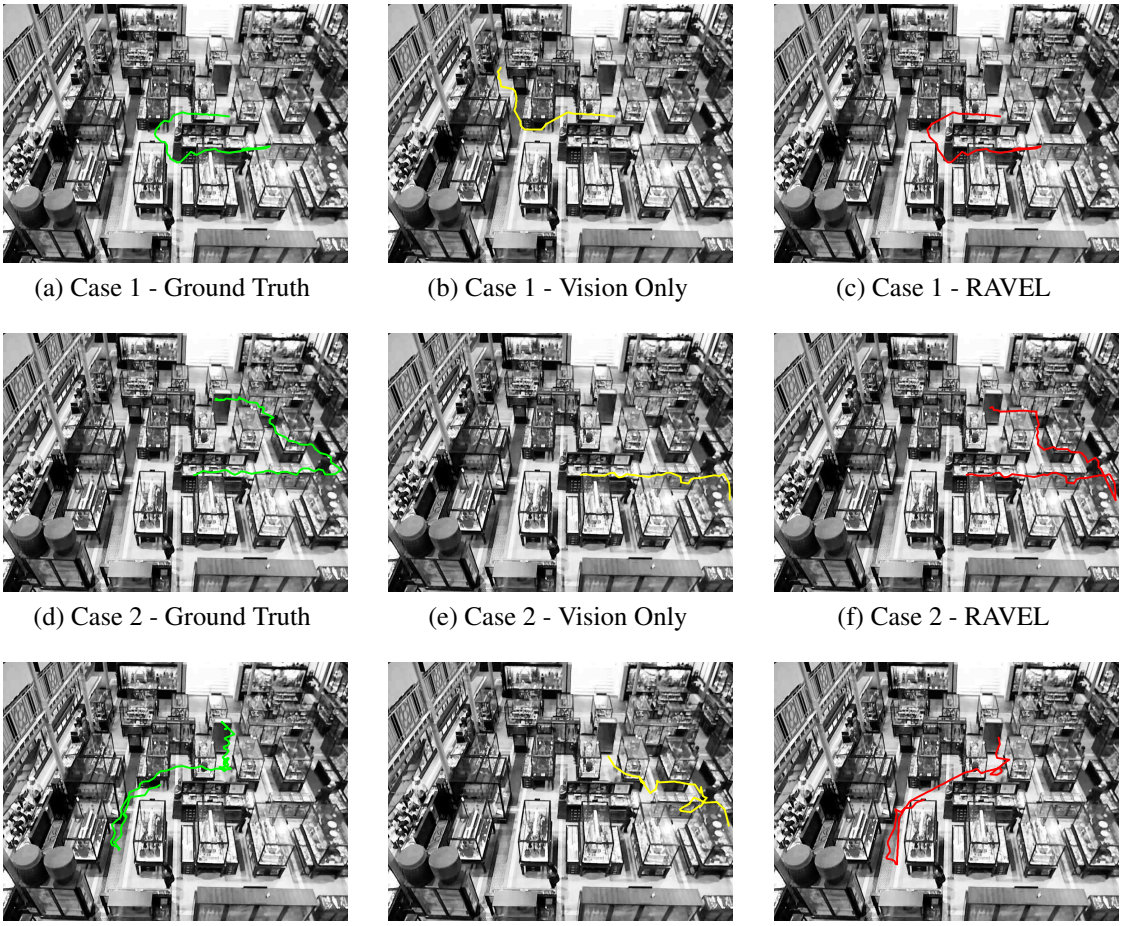
The third set of experiments aims to compare RAVEL with Vision-only tracker in terms of their ability to correctly determine whether a target is inside or outside the camera's FOV. To do so, we use the *overlap error* metric defined above as the percentage of frames in which the estimated trajectory incorrectly places the user inside or outside the FOV when compared to the ground truth. Fig. (3.9a) shows the CDF of the overlap error of the two algorithms over 160 windows (40 non-overlapping windows times four people). Observe that up to 70% of these windows have no overlap error for RAVEL, as opposed to 50% for the competing approach. The maximum overlap error of RAVEL is 40% as opposed to 100% for Vision-only tracker. This shows that RAVEL's superior performance in resolving ambiguities that result from people entering and leaving the FOV and re-appearing into the FOV from different entry points. Illustrative examples of such cases will be examined below.

3.6.2.4 Online location error

The last set of experiments focuses on the online case, where we are interested in the most recent location of a target. We still assume that we have access to a historical window of visual and WiFi detections, and we investigate the impact of the window size on the online (most recent) location error. As expected, Fig. (3.9b) shows that the online location error decreases as we increase the window size. This behavior is similar to that of the offline location error (also shown in Fig. (3.7b)), which measures the average error across all positions of the window. Notice in Fig. (3.7b) that the online location error is always slightly higher than the offline location error for a given window size, because the last position estimate in a given window has no future detections to benefit from.

3.6.2.5 Illustrative examples

In Fig. (3.10) we present three illustrative examples showing the performance of the proposed and competing (i.e. vision-only) algorithms: In the first example (Fig. (3.10a)), the person stops at particular museum exhibit, and stays there for a long period of time. Then she starts walking again, and performs a U turn. While the person is looking at the exhibit without moving, no camera detections are generated. Once she starts moving again it becomes difficult to distinguish in which direction she actually moves using camera detections due to ambiguities with other people's trajectories. One hypothesis is that her path merged with a second person's path moving north as illustrated in Fig. (3.10b); another hypothesis is that she makes a U turn as in Fig. (3.10c). By taking into account her radio data,



(a) Case 1 - Ground Truth (b) Case 1 - Vision Only (c) Case 1 - RAVEL

(d) Case 2 - Ground Truth (e) Case 2 - Vision Only (f) Case 2 - RAVEL

(g) Case 3 - Ground Truth (h) Case 3 - Vision Only (i) Case 3 - RAVEL

Figure 3.10: Illustrative examples showing the performance of the proposed and competing algorithms.

RAVEL can identify the correct trajectory. Another interesting case where WiFi measurements are beneficial is illustrated in Fig. (3.10d). In this case the person leaves the camera's FOV and after some time he re-enters. Fig. (3.10f) shows that RAVEL can re-establish the identity of the person when he re-enters the FOV and provide the correct trajectory. Finally, Fig. (3.10g) illustrates the situation of two targets splitting paths which is quite common in crowded environments. More specifically, two people start walking together (i.e. one detection contains two targets) moving south when at some point they split following different paths. In this case the Vision-only tracker has a 50% chance of following the wrong path, whereas RAVEL can use radio data to select the correct path. Although RAVEL allows us to resolve a number of common ambiguities, it is by no means a perfect tracking system; we have observed certain cases in which ambiguities are not correctly resolved either due to a very limited window size, or to a highly crowded and complex scene. In addition, if there

are regions where either WiFi or vision are unavailable, in its current implementation no positions are returned. In the future this could be solved by incorporating other modalities such as inertial sensing.

Our qualitative and quantitative results though suggest that RAVEL is significantly more accurate and robust than existing Vision-only tracking systems, requiring only a negligible overhead to extend their implementation.

3.6.2.6 Comparison with EV-Loc

Finally, we close our evaluation by comparing RAVEL with the state-of-the-art EV-Loc system [43] which also uses visual and radio signals for positioning. More specifically, EV-Loc system assumes that people carrying WiFi-enabled mobile devices are moving inside an area which is covered by CCTV infrastructure. The workflow of this technique is as follows: In each time-step the visual and WiFi signals of all people in the scene are first acquired and converted to distances from known locations (i.e. WiFi access points). The log-normal radio propagation model is being used by Eqn. (3.6) to convert the WiFi RSS readings to distances from the access points.

In the next step a cost matrix which corresponds to the similarity between each pair of converted distances from the visual and WiFi signals is built. The Hungarian algorithm [74] is then used to find the best matching pairs i.e. given the cost matrix it finds the associations between visual and WiFi signals, thus it identifies a person by their mobile device id. More formaly, let the set of visual signals (converted to distances) denoted by $\mathbf{x} = (x_1, ..., x_n)$ and the set of WiFi signals (converted to distances) denoted by $\mathbf{y} = (y_1, ..., y_n)$ where n is the number of people in the scene EV-Loc aims to solve the following:

$$\underset{\pi_i}{\operatorname{argmin}} \sum_{i=1}^{n} ||x_i - y_{\pi_i}|| \tag{3.9}$$

where π_i is a permutation of the vector $\mathbf{y} = (y_1, ..., y_n)$ and $||.||$ is the Euclidian distance. Once the above optimisation problem is solved with the Hungarian algorithm the two signals (i.e. the matching visual and WiFi signals) are fused into a new signal z_i given by:

$$z_i = \alpha x_i + \beta y_{\pi_i} \tag{3.10}$$

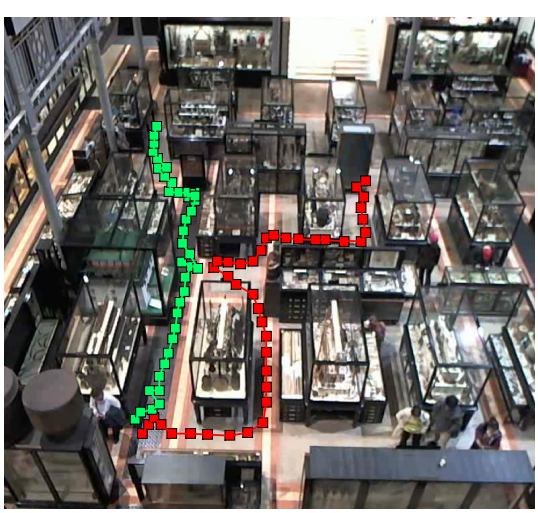
where the parameters α and β reflect the measurement confidence of the visual and WiFi signals respectively. This new signal z_i provides the location estimate using both visual and WiFi signals. In practice, however, the authors state that the accuracy of visual signals is far better than the accuracy of the WiFi signals and so the parameter β is usually close to 0 and they take the visual localisation result alone as the final fusion result. In other words

in EV-Loc WiFi signals are primarily used for tagging a visual trajectory with a mobile id in order to identify a user.

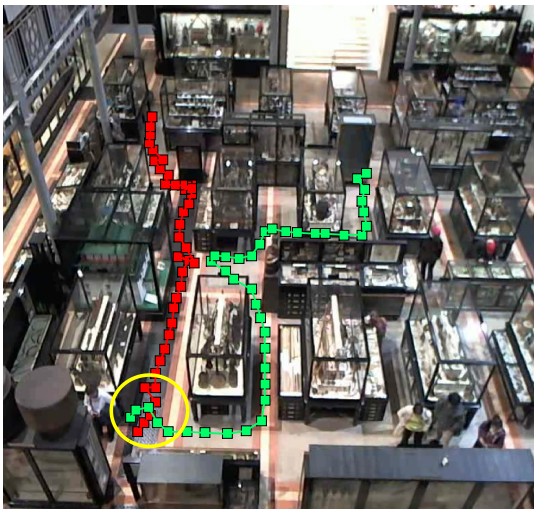
In addition, EV-Loc first performs visual object matching, i.e. first it finds the correspondence of objects across frames using visual tracking techniques and then applies the fusion method described above. To do so it relies on appearance features and it generates similarity matrix between every pair of consecutive visual frames. Then it uses the Hungarian algorithm to find the best match between every pair of neighbouring visual frames. In other words EV-Loc uses a global nearest-neighbour data association technique which accounts for the visual appearance of objects. After this step, it uses fusion technique to localise an object as discussed in the previous paragraph. Compared to the proposed system EV-Loc differs in the following points:

- The fusion of WiFi and visual signals happens after the visual trajectories are formed (i.e. post-fusion method) instead of using WiFi signals to guide the tracking process like RAVEL.
- The fusion technique of EV-Loc (Eqn. (3.10)) practically does not help improve the positioning accuracy and it is used mainly for target identification.
- The visual tracking is done on a frame-by-frame basis and requires the use of visual (i.e. appearance) features. Instead RAVEL provides a multi-hypothesis tracking framework which incorporates WiFi signals into the tracking process and it does not require visual features.

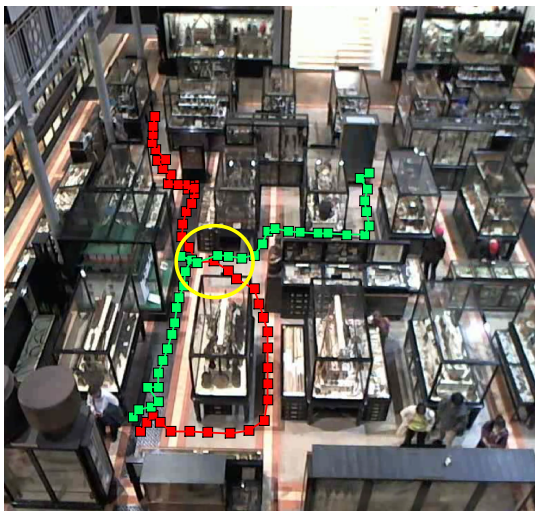
To compare with EV-Loc, we have implemented a frame-by-frame data association technique and more specifically the global nearest-neighbour data association filter (GNNDAF) to deal with the visual object tracking part. More specifically, we assume that people are moving according to a constant velocity model, thus they can be tracked using a Kalman filter. Unlike in the original EV-Loc, in this implementation no appearance features have been used so that we can compare it with RAVEL which does not make use of any appearance features. The GNNDAF uses the measurement likelihood from the Kalman filter to form a similarity matrix between all pairs of objects in two consecutive frames and then it finds the associations between tracks and measurements using the Hungarian algorithm. Once we form the visual trajectories we convert them into distances and together with the distances obtained from the WiFi signals we use again the Hungarian algorithm to find the association between visual and WiFi signals as described by EV-Loc. Finally, we use Eqn. (3.10) with $\alpha = 0.7$ and $\beta = 0.3$ to estimate the position of the target.



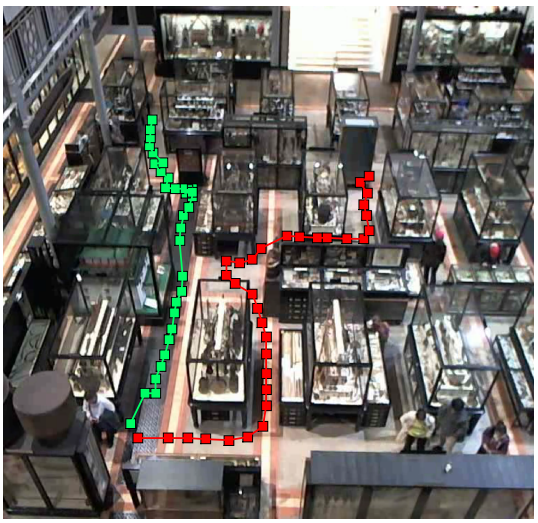
(a) Ground Truth



(b) EV-Loc tracking error - Case 1



(c) EV-Loc tracking error - Case 2



(d) Correct tracking with RAVEL

Figure 3.11: The figure shows tracking problems associated with the EV-Loc system, (a) The ground truth trajectories of two people walking north, (b, c) Tracking problems associated with the EV-Loc system. The yellow circle indicates id-switches due to visual ambiguities. (d) RAVEL is able to resolve these visual ambiguities and maintain tracking which is very close to the ground truth.

Before comparing RAVEL and EV-Loc in terms of accuracy we first outline the tracking problems associated with the EV-Loc system. These are shown in Figs. (3.11a)-(3.11c). More specifically, Fig. (3.11a) shows the ground truth trajectories of two people (red and green) moving north. As we have already mentioned EV-Loc applies first visual tracking techniques on a frame-by-frame basis to create visual trajectories and in a second step it uses WiFi and vision to localise the targets. This method however is prone to association

errors as shown in Fig. (3.11b) and Fig. (3.11c). As we can see once the wrong association is made at a particular point (indicated by the yellow circle) the tracking process diverges significantly from the ground truth and it can never recover as opposed to RAVEL which keeps track of multiple hypotheses. Moreover, it is very interesting to note here that the WiFi signals cannot be used to separate close spaced targets. In the illustrated scenario, even if we had used the EV-Loc system with tracks of length one (i.e. in each time-step we use the Hungarian algorithm to associate single visual locations and not trajectories with their corresponding WiFi signals) still these associations could not be made because the WiFi signals are not accurate enough to distinguish among close spaced targets. From our experience we have observed that WiFi signals become discriminative enough over a given period of time (e.g. 10-60 sec). This is because people will most likely follow different paths resulting in different WiFi signatures over those paths. For this reason we have designed RAVEL to use multi-hypothesis tracking in order to create long enough and distinct enough visual trajectories so that WiFi signals can be used to glue these trajectories together. This is the main reason why RAVEL works best on long time windows, because WiFi is discriminative enough to stitch fragmented trajectories together and identify people. The RAVEL output for the example in Fig. (3.11a) is shown in Fig. (3.11d). As we can see RAVEL successfully identifies the two people and manages to maintain tracking.

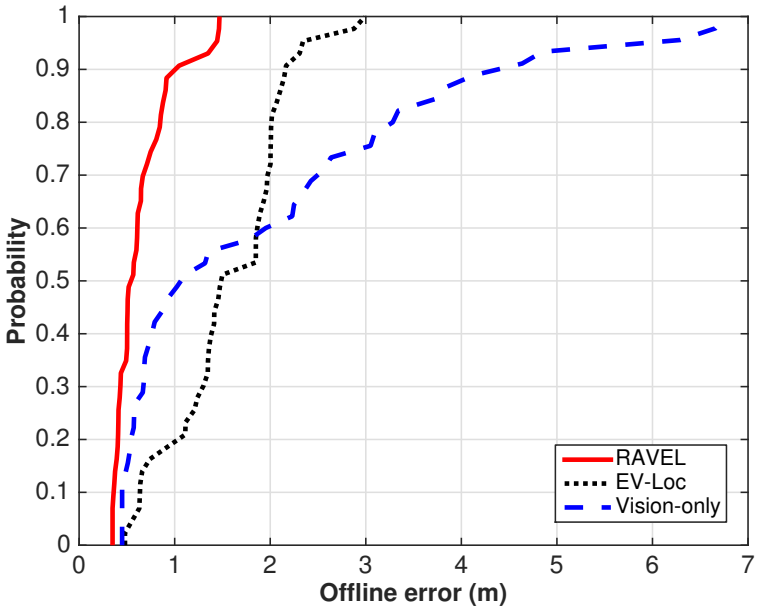


Figure 3.12: Accuracy comparison, RAVEL vs EV-Loc vs Vision-only.

Finally, Fig. (3.12) shows the off-line error CDF of RAVEL and EV-Loc over 160 time windows, of 120 frames each of our museum dataset. Compared to RAVEL's sub-meter

accuracy, EV-Loc has a 90 percentile error of about 2.1 meters. EV-Loc is error prone to association ambiguities and it does not use WiFi signals efficiently as opposed to RAVEL.

3.6.2.7 A note on computational complexity

The main factor which determines the computational complexity of the proposed algorithm is the number of possible hypothesis generated (i.e. number of branches of the tracklet trees) that need to be evaluated. All possible path hypotheses of a single target are encoded in a tracklet tree. Thus as the number of targets increases the number of tracklet trees that we need to process and maintain also increases (linearly with respect to the number of targets). In essence the proposed approach maintains a tracklet forest which encodes all possible path hypotheses for all targets. Now, as the number of tracklet trees increases the total number of tracklet tree branches also increases. The total number of tree branches can very quickly become very large due to simple combinatorial explosion especially in crowded environments. In essence, as the number of visual ambiguities increases the number of tree branches also increases in order to account for the possible path hypotheses. In challenging environments with false and missing camera detections, close spaced moving targets, etc. we expect to have an increase in the total number of tree branches. On the other hand, in ideal scenarios where no visual ambiguities arise, the tracklet trees are being reduced to linear lists.

In order to reduce the computational complexity it is necessary to reduce the number of possible hypotheses generated. This however is a tradeoff between accuracy and computational cost. To reduce the number of generated hypotheses a number of heuristics and techniques can be applied [99, 56] including deleting unlikely hypotheses, merging similar hypotheses and maintaining only the M most likely hypotheses.

3.7 Discussion

A large class of applications that require high positioning accuracy are typically deployed in large public or commercial buildings with open plan architecture, such as museums, airports, stations and shopping centres. The idea is that by knowing the exact location of a person, it is possible to provide them with useful information about the object they are currently focusing on, whether this is an exhibit in a museum, a tool in an industrial plant, or an item that they are interested in buying. Such environments are typically outfitted with camera systems, which were originally installed for safety and security purposes, and their output is manually handled by security personnel. In addition, in many scenarios such environments are equipped with WiFi infrastructure which is used for communication purposes.

Combining these two sensing modalities has the great potential in providing accurate and cost effective positioning without the need of investing in bespoke infrastructure.

In this chapter, we propose the use of existing camera infrastructure for the originally unintended task of indoor positioning. We show that the challenges in visual (camera based) tracking are numerous: cameras are poor at identifying people, especially when the room is not well lit (in a museum), or when the camera is pointing down at them making it difficult to distinguish facial features (bird's-eye view cameras in large stations). In addition, environments such as museums and airports are typically crowded with people crossing paths, and hiding behind dynamically changing obstacles. These conditions create ambiguity and make visual tracking a particularly challenging task.

On the other hand, radio-based systems, which also rely on existing infrastructure have their own limitations. For example, the strength of received WiFi radio signals fluctuates over time at a given position, due to multi-path signal propagation that is hard to model in busy and dynamically-changing environments. Thus radio signal strength is not an accurate enough indicator of location. Therefore, in open-plan indoor environments, existing solutions based on radio also fail to satisfy the accuracy requirements needed for a practical positioning solution.

Motivated by the noisy nature of both visual and non-visual sensor modalities, we have explored whether we can effectively combine them to overcome each other's weaknesses. More specifically, in this chapter we have introduced RAVEL, a new positioning system that integrates anonymous visual tracking with radio measurements to provide accurate tracking information. Although noisy radio measurements are inadequate for positioning on their own, we demonstrated how they can greatly enhance the performance of a relatively simple visual tracker by handling challenging cases like occlusion, trajectory splitting and entry/exit from the camera field-of-view. The additional advantage of our approach is that it automatically learns the radio propagation model, thus not requiring site-specific calibration.

As we have already mentioned RAVEL is an off-line tracking system i.e. the tracking is postponed until all measurements inside a time window have been observed. We have chosen to design RAVEL as an off-line tracking system in order to take advantage of the discriminative power of WiFi RSS signals over long periods of time. This allows us to use WiFi signals to resolve motion and visual ambiguities and connect together anonymous and fragmented visual trajectories. In the Evaluation section we have shown that the accuracy of RAVEL increases as we increase the window size. This is because the WiFi signals of multiple people become distinct over long periods of time, which help us identify and track more accurately. On the other hand as we decrease the window size the accuracy of

RAVEL approximates the accuracy of Vision-only tracking (Fig. (3.7b)). This is because the WiFi signals over short periods of time cannot be used to differentiate among close spaced moving targets. Thus, in this case RAVEL reduces to visual tracking.

Depending on the application scenario a moderate amount of delay can be tolerated. However, for many applications this is not possible. In the next chapter (Chapter 4) we will show how to design an on-line tracking system by incorporating more sensor modalities.

Chapter 4

Tracking with Camera, Radio and Inertial Data

4.1 Introduction

In the previous chapter we presented RAVEL, a new low-cost positioning system that integrates anonymous visual measurements with radio observations to provide accurate positioning in large public or commercial spaces. We have shown that although the WiFi measurements are not by themselves sufficiently accurate, when they are fused with camera data, they become a catalyst for pulling together ambiguous, fragmented, and anonymous visual tracklets into accurate and continuous paths, yielding typical errors below one meter.

As we have shown in the previous chapter RAVEL provides off-line tracking i.e. the trajectory of each person is reconstructed after all camera detections and WiFi measurements for a period of time have been observed. This architecture however, is not suitable for all types of applications. In certain scenarios real-time tracking of multiple people is required. To motivate this we consider the scenario of tracking people in industrial settings such as construction sites. For example, we would like to have a system which can monitor the location of workers to indicate working hazards (e.g. red and green zones), which can be individually tailored (Fig. (4.1)). For example, a steel-worker has the training to operate in areas which might not yet be poured with concrete whilst forming the steel rebar. Conversely, a general construction worker should not venture into regions where steel-work has not been completed. This level of safety requires positioning precision beyond the majority of indoor positioning solutions, with desired sub-meter accuracy and real-time operation.

Moreover in today's large and complex industrial environments the need of advanced planning and scheduling, careful coordination, efficient communication and reliable activity monitoring is essential for productivity purposes. The US National Research Council

found in [6] that construction lags behind other industries in terms of productivity, and blamed the situation on problems with planning, coordination, and communication.

Figure 4.1: Monitoring the location of workers in real-time to indicate working hazards. The figure shows the installation of a new wall during one of our experiments in a construction site. Our aim is to provide a system that can monitor the location of workers providing them with notifications on working hazards (i.e. red and green zones).

Motivated by the above problems we will show in this chapter how to design a real-time tracking system to be used in industrial settings and other challenging environments which require real-time operation and at the same time high accuracy. Real-time operation and high accuracy however are not the only requirements for a practical positioning system. For such a positioning system to be used and adopted in the construction industry, it should be low-cost as well and exploit existing infrastructure. We note that a pervasive feature of construction environments is the use of site-wide closed-circuit television (CCTV) to provide security. CCTV alone however is not useful for positioning, as we have to identify what each object in the video sequence actually is. This is especially challenging for low-resolution, grainy footage typically obtained from CCTV, which is unable to accurately distinguish between different people based on facial recognition.

In order to address all the above problems we propose a novel multi-modal multi-hypothesis tracking framework. More specifically, we will show how we can use Rao-Blackwellized particle filtering and the fusion of three sensing modalities (i.e. visual, radio and inertial) to track and identify multiple construction workers in real-time accurately and

robustly. We assume that the workers carry devices (i.e. smart-phones) which emit radio signals (WiFi/BTLE) and capture inertial measurements. We use these measurements to assign identities to the visual trajectories and improve the tracking accuracy as the visual detector is affected by occlusions, changing light conditions and challenges in detection when targets coalesce e.g. if workers are standing together.

4.2 Problem Definition

In this chapter we tackle the problem of tracking people in environments equipped with one or more stationary calibrated cameras. We assume that people that desire to be tracked carry a mobile device, such as a smart-phone or customised worker safety equipment, and move freely in and out of the field of view (FOV). We divide time into short time intervals, and at each time t we receive a number of camera detections of the moving objects denoted as $C_t = \{c_t^1, c_t^2, ..., c_t^j, ...\},\ 1 \leq j \leq |C_t|$. A camera detection c_t^j represents the bounding box of the j_{th} object generated by a foreground detector. Note that at time t we could be receiving camera detections not only from people but also from other moving objects (i.e. vehicles); false positive detections are also received due to illumination changes, shadows, etc. In order to reduce the number of false positive detections and concentrate on detecting only people we apply a head detector to the output of a foreground detector. A camera detection c_t^j is projected into the ground plane via a projective transformation which will be denoted as $\hat{c}_t^j$.

At time t we also receive a collection of radio measurements $R_t = \{r_t^k\},\ 1 \leq k \leq K$ where K is the total number of people with mobile devices who wish to be tracked and $r_t^k = [\text{rss}^1, ..., \text{rss}^m]_t^k$ is a vector of received signal strength (RSS) measurements of the k_{th} device from m access points. Additionally, we assume that each mobile device is equipped with an inertial measurement unit (IMU) containing an accelerometer and a magnetometer. This allows us to generate at time t a collection of inertial measurements denoted as $S_t = \{s_t^k\}$ where $s_t^k = [b_t^k, d_t^k, \theta_t^k]$ is a vector that contains the step indicator, step-length and heading of the k_{th} person respectively. Each index k uniquely identifies a person and corresponds to a unique MAC address of the mobile device.

The problem to solve is the following: *Given anonymous camera detections $C_{1:t}$, id-linked radio measurements $R_{1:t}$ and id-linked inertial measurements $S_{1:t}$ estimate the trajectories of all users carrying mobile devices and moving inside the camera FOV.*

4.3 Contributions

The major contributions of this work are as follows:

1. We are investigating the problem of real-time tracking in challenging industrial settings and we are proposing a positioning framework explicitly designed for these environments. Our particle-filter based multi-hypothesis tracking framework utilises three different sensor modalities (i.e. vision, radio and inertial) to allow for accurate multi-target tracking in challenging conditions (e.g. long-term occlusions, missing detections, noisy measurements, etc.)

2. We proposed a CCTV and smartphone based positioning system. We show how to combine anonymous measurements from CCTV with id-linked measurements from smart-phones; thus leveraging the positioning accuracy of visual measurements with the identification accuracy of radio/inertial measurements.

3. We demonstrate the impact of applying the social force model to improve tracking. In a construction site the environment changes rapidly with the addition of new walls, corridors, etc. These changes define the walkable area by restricting human motion in certain locations. In this work we show how to take advantage of these environmental changes with social forces to significantly increase the tracking accuracy.

4. We have conducted extensive experiments in a real construction site with the help and guidance of our industrial partners.

4.4 System Architecture

An overview of the proposed system architecture is shown in Fig. (4.2). Our system consists of the following modules :

- **Foreground Detector:** In order to detect the moving people in the scene we use the Mixture of Gaussians (MoG) [91] method. This is one of the most popular foreground detection techniques used by many object detection systems. This is because it is lightweight and capable of running in real-time embedded camera networks [92]. Additional improvements (e.g. [100]) to the original algorithm allow for shadow detection and more efficient real-time detection.

- **Radio Model:** Similarly to RAVEL, in order to model the WiFi received signal strength at a specific distance from the transmitter we make use of the log-normal path loss model [65, 101]. According to this model the received signal strength $P_R(d)$ at distance d from the transmitter is given by the following equation:

$$\begin{aligned} P_R(d) &= P_T - PL(d_0) - 10nlog_{10}\frac{d}{d_0} + X_{\sigma^2} \Rightarrow \\ P_R(d) &= P_0(d_0) - 10nlog_{10}\frac{d}{d_0} + X_{\sigma^2} \end{aligned} \tag{4.1}$$

  where P_T is the transmit power, $PL(d_0)$ is the path loss for a reference distance d_0, n is the path loss exponent, P_0 is the received power at the reference distance d_0 and $X_{\sigma^2} \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable which models the path loss variations due to shadowing. Our choice of using a radio propagation model over fingerprinting is because: a) the above radio model is linear with just 2 free variables which can be easily learned and b) the proposed system uses WiFi, inertial and camera measurements for positioning thus the accuracy of the radio model is sufficient for the problem that we are trying to solve. On the other hand, fingerprinting usually requires intensive surveying of the environment to built radio maps and systematic recalibration to account for temporal changes.

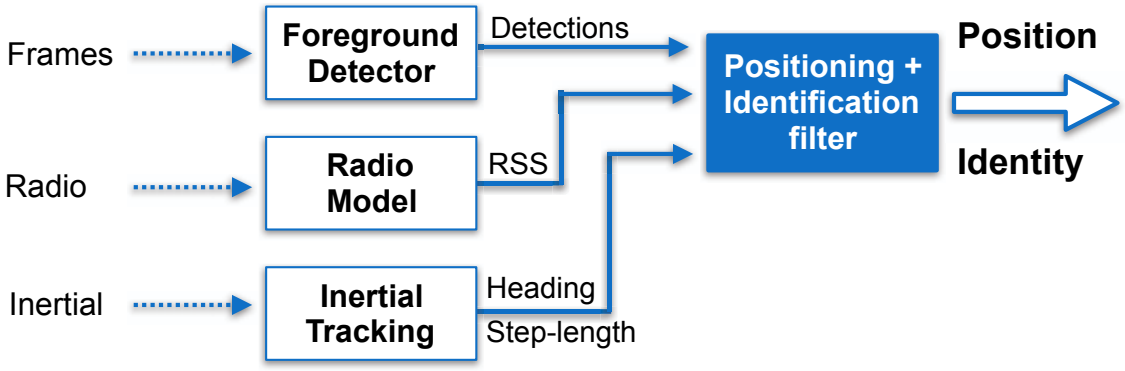


Figure 4.2: Overview of the proposed system architecture.

- **Inertial Tracker:** This module takes as input the accelerometer and magnetometer data from a user's smart-phone and uses them to predict the user's motion. Compared to the traditional motion models (e.g. constant velocity/acceleration) the inclusion of inertial measurement for motion prediction allows us to achieve higher tracking accuracy and maintain tracking behind obstacles and occlusions. The fusion of camera and inertial measurements in our system provides robustness in situations where we

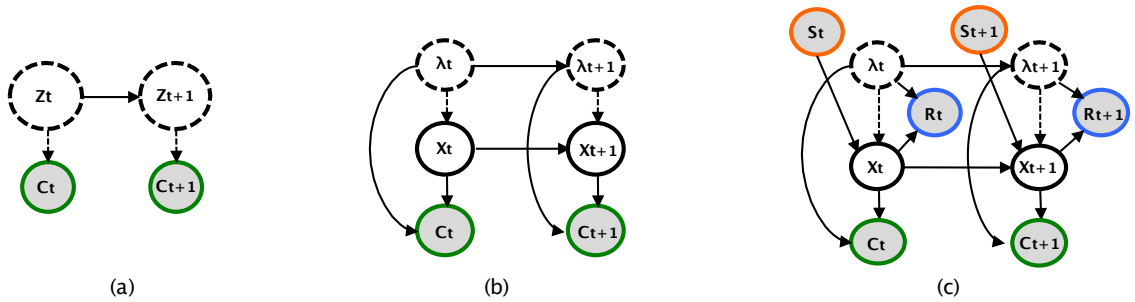


Figure 4.3: Graphical models for the multiple target tracking problem. Shaded nodes indicate observations and clear nodes indicate hidden variables. Nodes with dashed lines use particle filtering estimation. Dashed arrows indicate that the association between target states and measurements must be recovered before updating a target with a specific measurement. (a) Standard particle filter: uses sampling to estimate the joint posterior distribution of states and data associations ($P(Z_t|C_{1:t})$), (b) In Rao-Blackwellized particle filter (i.e. RBMCDA) the data-association is decoupled from the state estimation. The filter samples only from the data-associations distribution $P(\lambda_t|C_{1:t})$. The distribution of target states conditioned on the association $P(X_t|\lambda_t, C_{1:t})$ is calculated analytically. (c) Proposed approach: id-linked radio (R_t) and inertial (S_t) measurements are incorporated to RBMCDA in addition to the camera detections (C_t). The data association problem is changed compared to (a) and (b) as now we need to recover the association between id-linked measurements and tracks in addition to the anonymous measurements-to-tracks association.

have missing detections and allows us to maintain tracking for all targets even without any camera observations.

- **Positioning and Identification filter:** This module obtains anonymous camera detections, radio and inertial measurements from multiple people and is responsible for solving three problems. Firstly, it establishes the correspondences of camera detections across frames, that is, links together anonymous camera detections that correspond to the same person. Secondly, it finds the mapping between anonymous camera detections and id-linked smartphone (radio and inertial) measurements. Finally, it identifies and estimates the positions of multiple targets.

4.5 Rao-Blackwellized Monte Carlo Data Association

Traditional multi-target tracking methods such as MHT [79], JPDA [78] and derivatives [56, 55] have certain limitations which makes them impractical for specific applications. For instance, the MHT has not been designed for on-line tracking (i.e. it uses batch processing), on the other hand JPDA suffers from track coalescence for closely spaced targets

and because it does not maintain multiple hypotheses this method is unable from recovering upon tracking errors.

In contrast, particle filtering fits better to the tracking problem that we tackle in this chapter. It provides a more general and flexible framework for the problem of online multi-target tracking and it can be considered as a generalisation of MHT. Instead of maintaining N most probable data association hypotheses, the joint tracking and data association problem is modelled as a Bayesian estimation problem and the posterior distribution is estimated with Monte Carlo methods. Sequential Monte Carlo based multiple target tracking methods have been presented in [102, 103, 104, 105]. These methods use particle filtering to approximate the joint posterior distribution of states and data associations. However, this type of particle filter modelling for the MTT problem can be inefficient and computationally expensive since in high dimensional state-spaces a large number of samples is needed in order to represent the joint posterior distribution.

Rao-Blackwellization can be used in certain scenarios in order to reduce the computational complexity of these methods. The Rao-Blackwellized Monte Carlo Data Association filter (RBMCDA) [106, 107] is a sequential Monte Carlo MTT method that uses Rao-Blackwellized particle filtering (RBPF) to estimate the posterior distribution of states and data associations efficiently. More specifically, instead of using a pure particle representation of the joint posterior distribution of states and data associations (see Fig. (4.3a)), RBMCDA proceeds by decomposing the problem into two parts: a) estimation of the data-association posterior distribution and b) estimation of the posterior distribution of target states. The first part is estimated by particle filtering and the second part is computed analytically using Kalman filtering (Fig. (4.3b)). The aforementioned decomposition is possible, since in RBMCDA the dynamic and measurement model of the targets are modelled as linear Gaussian conditioned on the data association and can thus be handled efficiently by the Kalman filter.

A high level overview of the RBMCDA algorithm is shown in Alg. (1). The algorithm maintains a set of N particles and each particle corresponds to a possible association of anonymous measurements (y_t) to tracks. Each particle maintains for each target its current state x_t (e.g. location) and state uncertainty (i.e. posterior distribution $p(x_t|y_{1:t})$). In the first step (line 4), a Kalman filter is used to predict the next state of a target based on its previous state ($p(x_t|y_{1:t-1})$). Then, the algorithm considers associating each anonymous measurement with each one of the targets in the particle and estimates the probability of each candidate association event (lines 5-6). The association events are modelled with the association indicator λ_t (e.g. $(\lambda_t = 0) \implies$ clutter association at time t, $(\lambda_t = j) \implies$ target j association at time t, etc). The association probability $\hat{\pi}_j$ for target j is computed

from the measurement likelihood $\hat{p}(y_t|\lambda_t)$ and the prior probability of data associations $p(\lambda_t|\lambda_{t-1})$. By sampling the resulting importance distribution, the algorithm selects only one of the candidate associations (line 7) and updates the state of the respective target with the anonymous measurement (line 8). This is repeated for each anonymous measurement (e.g. for each camera detection in the camera frame). The particle's weight is then updated taking into account its previous weight and the probabilities of selected associations (line 9). Once all particles have been updated and their weights normalised (line 11), they are re-sampled based on their normalised weights (line 12). At the end of each iteration, the positions of the targets are estimated as a weighted average (i.e. mixture of Gaussians) across all particles (line 13).

Note that the algorithm above allows us to enforce data association constraints. For instance, we can express that each track is updated by at most one visual measurement, by suitably modelling association priors in line 5.

1: **Input:** N particles, a measurement vector y_t.
2: **Output:** $p(x_t, \lambda_t|y_{1:t})$: the joint distribution of target states and target-to-measurement associations at time t given measurements up to time t.
3: **for** each particle $i \in (1..N)$ **do**
4: For all targets run Kalman filter prediction step.
5: Form the importance distribution as:
For all association events j calculate the unnormalized association probabilities:
$\hat{\pi}_j^{(i)} = \hat{p}(y_t|\lambda_t^{(i)} = j, y_{1:t-1}, \lambda_{1:t-1}^{(i)})p(\lambda_t^{(i)} = j|\lambda_{1:t-1}^{(i)})$
6: Normalize the importance distribution.
7: Draw new $\lambda_t^{(i)}$ from the importance distribution.
8: Update target $\lambda_t^{(i)}$ with y_t using Kalman correction step.
9: Update particle weight.
10: **end for**
11: Normalize particle weights.
12: Resample.
13: Approximate $p(x_t, \lambda_t|y_{1:t})$ as:
$p(x_t, \lambda_t|y_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta(\lambda_t - \lambda_t^{(i)}) \mathcal{N}(x_t|M_t^{(i)}, P_t^{(i)})$ where $(M_t^{(i)}, P_t^{(i)})$ are the means and covariances of the target states of the i_{th} particle.

Algorithm 1: A high-level description of the RBMCDA filter

The existing RBMCDA algorithm is designed to work with anonymous observations. In the next section we point out how we extend it in order to exploit radio and inertial observations that are inherently linked to unique device IDs (i.e. MAC addresses).

4.6 Proposed Approach

We are now in a position to describe how we extend the RBMCDA framework to address the identification and tracking problem in a construction site setting. The key difference here is that we introduce id-linked observations in addition to the anonymous camera observations (Fig. (4.3c)). This impacts a number of steps in the algorithm above as explained in this section.

4.6.1 State Prediction and Update

As in the original algorithm, each particle uses a set of Kalman filters to track targets. However, in our case, we are not interested in tracking all targets within FOV; we only track people equipped with mobile devices and we continue to do so when they temporarily come out of the FOV. We extend the framework in [106, 107], in order to use id-linked observations in the prediction and correction steps of the Kalman filter. In particular, we use inertial sensor measurements to predict the next state of a person (instead of only relying on the previous state as in line 4). Furthermore, we use WiFi/BTLE and camera measurements to correct the person's state (instead of only anonymous camera measurements as in line 8). More specifically, the target's dynamics in our system are modelled by the following linear equation:

$$x_t = x_{t-1} + B_t \begin{bmatrix} d_{\Delta t} \cos(\theta_{\Delta t}) \\ d_{\Delta t} \sin(\theta_{\Delta t}) \end{bmatrix} + w_t \tag{4.2}$$

where t denotes the time index, $x_t = [x, y]^{\mathrm{T}}$ is the system state i.e. a 2-D vector of the target's position on the ground plane and the pair ($d_{\Delta t}$,$\theta_{\Delta t}$) represents the target's step-length and heading respectively calculated within the tracker's cycle time (Δt). The control input B_t is the output of a HMM-based step classifier which takes as input the accelerometer data from the user's device and returns a step indicator that shows whether a step has been taken or not. A more detailed description regarding our step detection approach is discussed in Appendix A. Finally, w_t is the process noise which is assumed to be normally distributed with mean zero and covariance matrix Λ (i.e. $w_t \sim \mathcal{N}(0, \Lambda)$). In order to calculate the step-length of a person we use an empirical model that takes into account the step frequency obtained from the accelerometer data and is given by the following formula:

$$s = h(a' f_{step} + b') + c' \tag{4.3}$$

where s is the estimated step-length, h denotes the user's height, f_{step} is the step frequency obtained from the device's accelerometer and (a', b', c') are the model parameters. The model above describes a linear relationship between step-length and step frequency

weighted by the user's height. The height value is set to the country's average for men of ages between 25 and 34 years old.

In addition, as we mentioned in Section 4.2, our objective is to track all people that carry mobile devices. Thus, once we associate a camera measurement to a person ID (i.e. device ID), Eqn. (4.2) is used as the predictive distribution of a Kalman filter to model the motion of the identified person using his/her inertial measurements.

Compared with existing techniques (i.e [108]) that use heuristics to model the human motion, we will show in the evaluation section that the use of inertial measurements in our approach results in more accurate tracking. In addition, we have observed that in a construction site workers do not walk regularly, instead they often make big, small and irregular steps depending on the task performed. This makes motion prediction even more challenging since it makes it harder for the step detector/classifier to detect some of the steps correctly. It is worth noting here that the proposed system can correct these step misclassification errors in many situations with the help of visual observations. For instance, when our step classifier predicts wrongly for a specific target that a step has been taken, our system can still correct the final estimated position using the location of the camera measurement. Under the assumption of unambiguous tracks the proposed technique can handle similar situations very efficiently. Figure (4.4) illustrates the scenario discussed above.
We should note here that Eqn. (4.2) is event-based (i.e. based on step events) and events among the different targets are inherently not synchronised. In other words, the steps of different people do not take place at the same time. However, because we need to know the predicted locations of all targets at a specific time, we process Eqn. (4.2) in a time-based manner. We run the prediction equation for all targets on fixed intervals (i.e. every second) and during that time we find the number of steps taken by each person and we calculate the step-length accordingly. Incomplete steps are handled by accounting only for a percentage of the step-length. In other words, d_t is not necessarily the length of a single step. However, throughout this paper we are going to refer to d_t as step-length.

Unlike the original RBMCDA filter that only uses anonymous observations to update the target's state (line 8), in our system a measurement y_t at time t is a vector containing an anonymous location measurement (2D image coordinates transformed to the world plane via a projective transformation [109]) from the camera system and multiple id-linked radio signal strength measurements from people's mobile devices. More formally, the measurement vector is defined as $y_t = [\hat{c}_t,\ rss_t^1,\ ...,\ rss_t^m]^{\mathrm{T}}$ where $\hat{c}_t$ is a camera observation which contains the 2D target coordinates on the ground plane and $rss_t^1,\ ..., rss_t^m$ denote the received radio signal measurements from m access-points of a particular mobile device.

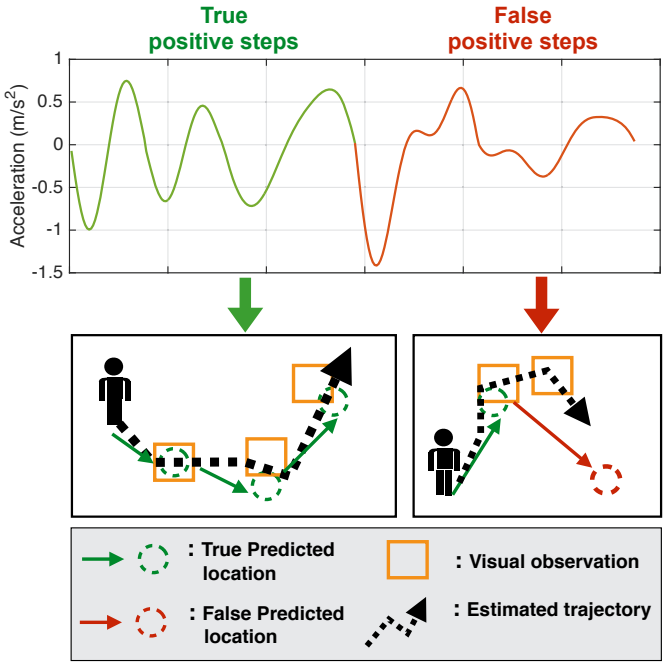


Figure 4.4: Fusing camera and inertial measurements. The dotted circles show the predicted location using inertial measurements (i.e. a step classifier, shown in the top picture, indicates if a step has been taken or not.) Square boxes indicate a camera detection (i.e. the location of a person). When a step is classified correctly the predicted location is collocated with the camera detection (picture on the left). The tracking accuracy can be decreased significantly when the step detector misclassifies a step. However, in the proposed system, the fusion with camera measurements allows to navigate towards the right path in cases where we have unambiguous trajectories (picture on the right).

Thus, the state vector x_t of a target is related to the system measurements y_t according to the following model:

$$y_t = f(x_t) + v_t = \begin{bmatrix} x_t \\ \text{RSS}_1(x_t)) \\ \text{RSS}_2(x_t)) \\ \vdots \\ \text{RSS}_m(x_t)) \end{bmatrix} + v_t \tag{4.4}$$

where f is a non-linear function that translates the true system state vector to the measurement domain and v_t is the measurement noise which follows a normal distribution with zero mean and covariance matrix R ($v_t \sim \mathcal{N}(0, R)$). The function RSS_i is given by:

$$\text{RSS}_i(x_t) = P_i - 10n_i log_{10}\|A_i - x_t\|_2 \ , \ i \in [1..m] \tag{4.5}$$

where m is the total number of WiFi/BTLE access points and $\text{RSS}_i(x_t)$ is the expected

signal strength at location x_t with respect to transmitter A_i. P_i is the received power at the reference distance of one meter and n_i is the path loss exponent. In order to meet the requirements of the RBMCDA filter, i.e. calculate analytically the posterior distribution of the target states with a Kalman filter, Eqn. (4.4) must be linear Gaussian. The non-linearity of the measurement model in our case is handled via the unscented transformation [57, 58]. Thus, the state estimation can be computed analytically using the unscented Kalman filter (UKF) and each particle contains a bank of UKFs; one filter for each target.

4.6.2 Tracking and Identification

In this section, we show how we modified the association steps in lines 5-7 to leverage id-linked measurements.

Suppose for instance that at time t we receive camera detections $C_t = \{c_t^j\},\ 1 \leq j \leq |C_t|$ and radio measurements $R_t = \{r_t^k\},\ 1 \leq k \leq K$ where K is the number of people with a mobile device. Each one of the $|C_t|$ anonymous camera detections could be one of the following three types: (a) a person with a device, (b) a person without a device or (c) clutter (e.g false camera detection caused by illumination changes). Our objective is to associate the type (a) camera detections with the correct radio measurements. In order to do that we use the following procedure. We enumerate all possible combinations $\Omega = |C_t| \times K$ between the camera detections and the id-linked measurements and we create new measurements $y_t^i, i \in [1..\Omega]$ with the following structure:

$$y_t^i = \{\hat{c}_t^m, r_t^j\},\ m \in [1..|C_t|],\ j \in [1..K] \tag{4.6}$$

where $\hat{c}_t^m$ is the camera measurement c_t^m projected into the ground plane. Now, a measurement y_t^i which contains a correct association will have the following property $\text{RSS}(\hat{c}_t^m) \approx r_t^j$ for the correct (m, j) pair, where $\text{RSS}()$ is the function in Eqn. (4.5). In other words, if a person is detected by the camera, then his/her radio measurements (i.e. received signal strength) at that location should match the predicted radio measurements at the same location. Camera detections of type (b) and (c) would normally not exhibit the same property. From our experiments in a real construction site, we have observed that the radio measurements are reasonably stable but only for short periods of time depending on the environmental dynamics. As we discuss in Chapter 5 by periodically re-learning the radio model, we make our system adaptive to the changing environment and thus we can use the procedure above to track and identify the people in the scene.

Moreover, the proposed algorithm can handle the creation and termination of tracks. For instance, when a new person (i.e. a new mobile device) is entering the FOV, we initiate

a new track by initializing the system state with the camera location that best matches the received radio measurements. Additionally, we allow a target to die when for a fixed period of time no camera observation has been used to update its state. The above procedure runs continuously thus new tracks are created and others are terminated dynamically as people enter and leave the FOV.

As we have already mentioned the association probability is computed as the product of the measurement likelihood and association prior. The measurement likelihood of associating y_t^i with target j, $\hat{p}(y_t^i|\lambda_t = j)$ is computed as $\hat{p}(y_t^i|\lambda_t = j) = \mathcal{N}(y_t^i; \hat{y}_t, V_t)$ where $\hat{y}_t$ is the expected measurement of target j at the predicted state and V_t is the innovation covariance obtained from the UKF.

Given m simultaneous measurements within a scan the predictive distribution of data associations can be defined as an m_{th} order Markov-chain $p(\lambda_t^m|\lambda_t^{m-1}, ..., \lambda_t^1)$ which allows us to enforce certain association restrictions. In our system this predictive distribution is defined (i.e. assigns zero probability to unwanted events) so that the following conditions are met:

1. A track can be updated with at most one measurement.

2. A measurement can only be used to update at most one track.

3. An already established track (with a specific sensor ID) can only be updated with a measurement of the same sensor ID.

4. Once a camera detection is assigned to a track all other measurements which include the latter camera detection are classified as clutter.

5. A new target is not born if there is an existing target with the same sensor ID as the newborn target. This means that each particle maintains only targets with unique sensor IDs.

Some of the above restrictions can be relaxed depending on the application scenario. For instance, when two people are close to each other they can be detected as one object. In this case the 4th restriction can be relaxed in order to allow two tracks (i.e. two people with different sensor IDs) to be updated with the same camera detection.

To summarize, a particle represents states only for people carrying mobile devices - not for all people in the field of view. Inertial data of each person's device are used to predict their next state. Anonymous camera data are associated with a person's track only if they *agree* with both their inertial and radio data.

Input: N particles, camera (C_t), radio (R_t) and inertial (S_t) measurements.
Output: $p(x_t, \lambda_t | y_{1:t})$.
Apply Eqn. (4.6) to C_t and R_t to create y_t.
for each measurement $m \in (1..|y_t|)$ **do**
  for each particle $i \in (1..N)$ **do**
    For all targets in i run prediction step (Eqn. (4.2)).
    Form the importance distribution and draw new association event ($\lambda_t^{(i)}$).
    Update target $\lambda_t^{(i)}$ with m using UKF correction step. Update particle weight.
  end for
end for
Normalize particle weights.
Resample.
Approximate $p(x_t, \lambda_t | y_{1:t})$ as in Algorithm (1)

Algorithm 2: A high-level work-flow of the proposed system.

First a foreground detector is used to detect the moving people in the scene and then the 2D image coordinates of the detected people are projected into the world plane (i.e. ground plane) via a projective transformation (i.e. homography). More specifically, given a set of points p_i in the projective plane $\mathbb{P}^2$ and a corresponding set of points $\hat{p}_i$ likewise in $\mathbb{P}^2$ we would like to compute the projective transformation that takes each p_i to $\hat{p}_i$. In our case we consider a set of point correspondences $p_i \leftrightarrow \hat{p}_i$ between the image plane and the world ground plane and we need to compute the projective transformation $H_{3\times3}$ such that $Hp_i = \hat{p}_i$ for each i. The matrix H can be computed using the Direct Linear Transformation (DLT) algorithm [109] which requires at least 4 point correspondences. Additional points can improve the estimation by minimizing a suitable cost function such as the geometric distance between where the homography maps a point and where the point's correspondence was originally found, i.e. we would like to find the matrix H which minimizes $\sum_i d(\hat{p}_i, Hp_i)^2$ where $d(.,.)$ is the Euclidean distance between the two points.

Once we calculate H we can use it to project the targets from the image plane into the ground plane and obtain their location on the ground plane. We can then use inertial and radio data using Eqns. (4.2) and (4.4) as explained earlier in this section.

Finally, we should note here that when at some time-step a particular target does not receive radio measurements then if the target is a new target the identification and creation of a new track is postponed until radio measurements are available. Otherwise, if the target is an existing target, tracking proceeds by only considering the motion model of the target (Eqn. (4.2)). A high-level work-flow of the proposed technique is shown in Alg. (2).

4.7 Integration of Social Forces

Traditionally, tracking systems use motion models to predict human motion [96, 97]. Examples of these models include the constant velocity and acceleration model, the Brownian model, the Gauss-Markov model, etc. These motion models however, make weak assumptions on human motion. For instance, the Brownian motion model makes no assumptions about the target dynamics and the constant velocity model assumes linear target motion. Such assumptions however are not suitable for all types of scenarios (i.e. tracking the workers in a construction site). Moreover in the case of long term occlusions and missing detections these motion models fail to maintain tracking which causes divergence from the true trajectory and significant losses in the tracking accuracy. In the previous section we have introduced a measurement-driven motion model based on inertial observations which helps us increase the accuracy of motion prediction and mitigate some of the challenges mentioned above. Although the motion prediction has been improved significantly by the fusion of inertial observations in the motion model of our system, there is still room for improvement. More specifically, we have observed that in a construction site workers do not walk regularly, instead they often make big, small and irregular steps and they change their speed and direction significantly depending on the task performed.

In this section we describe how we have modified our system to make use of the social force model (SFM) [95, 110] in order to improve motion prediction and achieve higher tracking accuracy in real-world and demanding scenarios. In the social force model the interactions between people are described with social forces. These forces model different aspects of motion behaviours, such as the motivation to reach a goal, the repulsive effect on walls, etc. In other words the social force model assumes that the behaviour of human motion is affected by the motion of other people and also by obstacles from the environment. Thus the SFM aims to describe and predict the behaviour of human motion with the introduction of repulsive forces exerted on people by modelling the interactions between people and the influence of the environment on human motion. Figure (4.5) shows the architecture of the proposed system with the inclusion of social forces. In addition to the inertial observations we have now introduced the use of social forces in the target dynamics. The rest of this section explains in more detail how we have augmented social forces to our system.

4.7.1 The Social Force Model

In the Social Force Model a person p_i with mass m_i aims to move with a certain desired speed $\hat{\nu}_i$ in a desired direction $\hat{\epsilon}_i$. In our system the desired direction is taken from the IMU measurements (i.e. heading) so that $\hat{\epsilon}_i = \theta^i$ and the desired speed $\hat{\nu}_i$ is calculated as

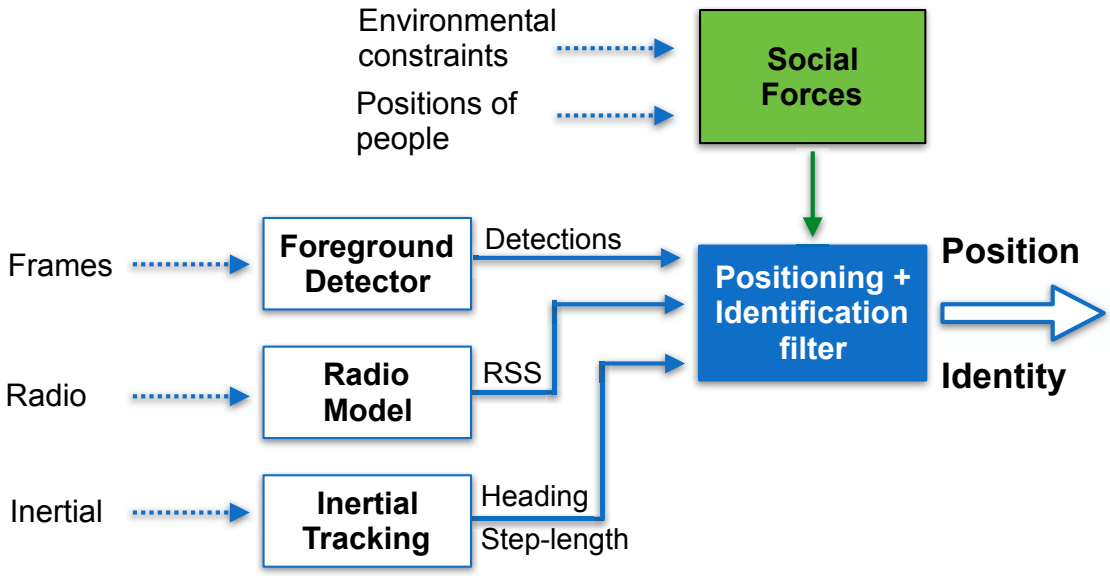


Figure 4.5: The proposed system architecture with the integration of social forces. We utilise inertial measurements and social forces in the target dynamics. We use social forces to account for the interactions between people and the effect of the environment on human motion.

$d_{\Delta t}/\Delta t$ where $d_{\Delta t}$ is the step-length from the IMU and Δt the tracker's cycle time. At each time step the motion of people is described by the superposition of repulsive and physical forces exerted from other people and the environment.

4.7.1.1 Repulsive Forces

Human motion is affected by environmental constraints (i.e. obstacles) and the motion of other people. Thus in the presence of other people or obstacles a person might not be able to maintain the current heading and speed. These disturbances are described by repulsive forces which prevent a person from moving along the desired direction. More specifically, the repulsive force F_i^R is modelled as the sum of social forces $f_{i,k}^{\text{soc}}$ exerted by other people or obstacles according to:

$$F_i^{\text{R}} = \sum_{j \in P \setminus \{i\}} f_{i,j}^{\text{soc}} + \sum_{j \in O} f_{i,j}^{\text{soc}} \tag{4.7}$$

where $P = \{p_j\}_{j=1}^{N_p}$ is the set of all people (i.e tracks) and $O = \{o_j\}_{j=1}^{N_o}$ is the set of all environmental constraints (i.e. obstacles). The above social repulsive forces are described

as:

$$f_{i,j}^{\text{soc}} = \alpha_j e^{\left(\frac{r_{i,j}-d_{i,j}}{b_j}\right)} n_{i,j}\gamma(\lambda, \phi_{i,j}) \tag{4.8}$$

where $j \in P \cup O$ and a_j, b_j denote the magnitude and range of the force respectively. People and obstacles are assumed to be circular objects with certain radii, thus $r_{i,j}$ denotes the sum of radii of entities i and j and $d_{i,j}$ is the Euclidean distance between their centers. The term $n_{i,j}$ describes the direction of the force, (normalised vector) pointing from entity j to entity i. Finally, the social forces are limited to the field of view of humans, therefore the anisotropic factor $\gamma(\lambda, \phi_{i,j})$ is added to the model and is given by:

$$\gamma(\lambda, \phi_{i,j}) = \lambda + (1-\lambda)\frac{1+\cos(\phi_{i,j})}{2} \tag{4.9}$$

where λ denotes the strength of the anisotropic factor and $\cos(\phi_{i,j}) = -n_{i,j} \cdot \hat{\epsilon}_i$ is the cosine of the angle between the desired direction and the direction of the force.

4.7.1.2 Physical Forces

Environmental constraints (i.e. walls, obstructions, etc) define the walkable area by restricting human motion in certain locations. These hard constraints can be modelled as physical forces exerted from the environment onto people and can be defined as follows:

$$F_i^{\text{phys}} = \sum_{j \in O} f_{i,j}^{\text{phys}} \tag{4.10a}$$

$$f_{i,j}^{\text{phys}} = c_j g(r_{i,j} - d_{i,j}) n_{i,j} \tag{4.10b}$$

where c_j denotes the magnitude of the force and $g(x)$ is defined as $g(x) = x$ if $x \geq 0$ and 0 otherwise, making $g(x)$ a contact force. We should note here that physical forces can also be applied between people if desired (i.e. so that different people would not occupy the same space). This can be done by adding an additional term in Eqn. (4.10a) to account for forces between people as we did in Eqn. (4.7).

4.7.2 Social Forces for Motion Prediction

We are now in a position to describe how we model the target dynamics in our system to make use of social forces and inertial measurements. More specifically, the total force F_i^{tot} exerted on a particular person p_i is the superposition of all repulsive and physical forces given by:

$$F_i^{\text{tot}} = F_i^{\text{R}} + F_i^{\text{phys}} \tag{4.11}$$

We can now include F_i^{tot} to our motion model (Eqn. (4.2)) by making use of Newton's second law given by $F_i^{\text{tot}} = m_i \frac{\mathrm{d}v_i}{\mathrm{d}t}$ so that Eqn. (4.2) becomes:

$$x_t = x_{t-1} + B_t \begin{bmatrix} d_{\Delta t}\, cos(\theta_{\Delta t}) \\ d_{\Delta t}\, sin(\theta_{\Delta t}) \end{bmatrix} + \frac{1}{2}\frac{F^{\text{tot}}}{m}\Delta t^2 + w_t \tag{4.12}$$

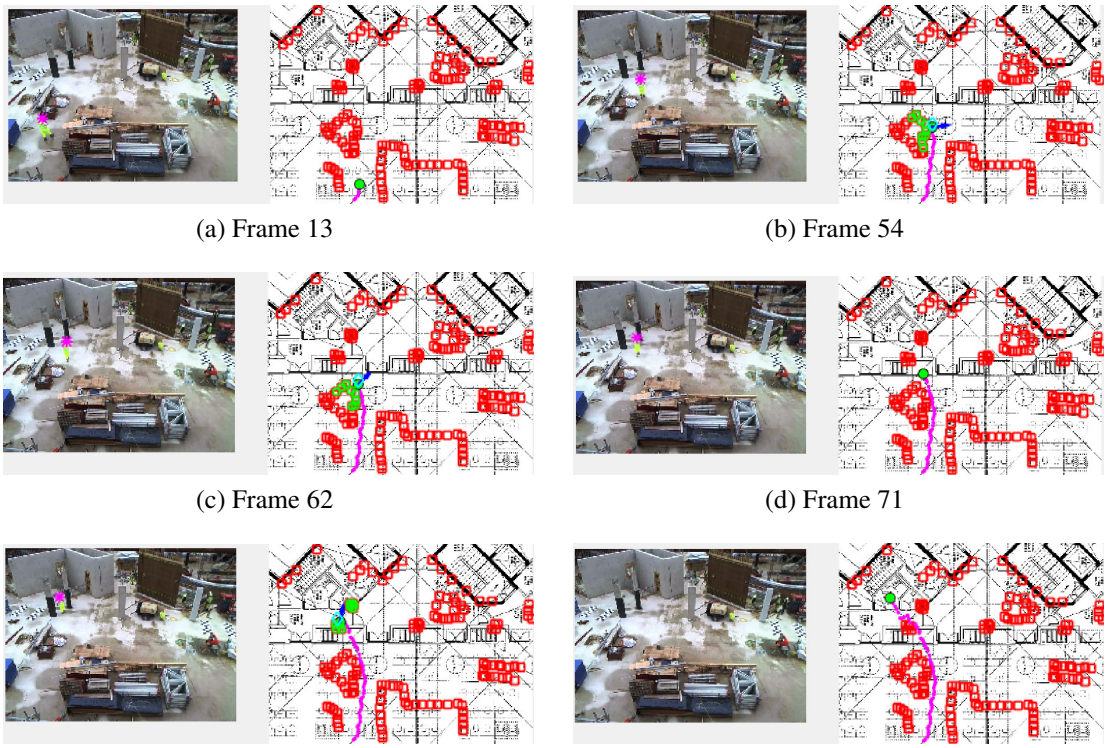
As we can see from Eqn. (4.12) the predicted motion of a person is calculated by taking account the previous position, inertial measurements (i.e. step-length and heading), and the forces exerted to this person by other people and the environment. Equation (4.12) can now be used in our tracking framework as the predictive distribution of the Kalman filter. This predictive distribution is given by:

$$p(x_t|x_{t-1}, S'_t, P_t, O_t) = \mathcal{N}(x_t; \psi(x_{t-1}, S'_t, P_t, O_t), J_\psi \Sigma_{t-1} J_\psi^{\mathrm{T}} + \Lambda) \tag{4.13}$$

where $S'_t = [d_{\Delta t} cos(\theta_{\Delta t}),\ d_{\Delta t} sin(\theta_{\Delta t})]^{\mathrm{T}}$ is a vector that contains the step-length and heading at time t, P_t is the set of all people tracked and O_t is the set of all obstacles from the environment. The function $\psi(x_{t-1}, S'_t, P_t, O_t) = x_{t-1} + B_t S'_t + \frac{1}{2}\frac{F^{\text{tot}}}{m}\Delta t^2$ is the mean of the predicted location, Σ_{t-1} is the covariance matrix of the estimate, Λ is the covariance matrix of process noise and $J_\psi = \frac{\partial \psi(\cdot)}{\partial x}$ is the Jacobian of $\psi(\cdot)$.

Finally, with the above motion model, in each time step in addition to the inertial measurements we can now use repulsive and physical forces exerted on targets in order to improve the predicted location estimates. We will show in the evaluation section that the integration of social forces in our motion model allows us to make better motion predictions and improve the accuracy of our tracking system. We should note here that the set of all people $P = \{p_j\}_{j=1}^{N_p}$ (i.e. the location of people) that is required in order to calculate repulsive forces is already maintained and provided by our tracker and the only additional piece of information needed is the set of environmental constraints $O = \{o_j\}_{j=1}^{N_o}$. For this it is necessary to build and maintain a map which contains the location of these constraints (e.g. obstacles, inaccessible areas, walls, etc). We will show in the next chapter (Chapter 5) how these environmental constraints can be learned in real-time.

Figure (4.6) shows the application of social forces during the tracking of one person. The red squares indicate the presence of obstacles (e.g. walls, equipment, etc.). At this stage these obstacles have been manually set. However, we will show in Chapter 5 that we can automatically learn this obstacles in real-time. When the target is not near any of these obstacles in the scenario depicted in this figure social forces have no effect (i.e. no forces from the environment affect the motion of the target). This is shown in Fig. (4.6a). However, we can see from Fig. (4.6b) that in the presence of obstacles the forces acting on the target push the target away from the obstacle, illustrated by the blue arrow,



(a) Frame 13 (b) Frame 54

(c) Frame 62 (d) Frame 71

(e) Frame 90 (f) Frame 113

Figure 4.6: The figure shows the use of social forces as a person walks through the area. Forces are exerted by the environment to the targets and affect their motion. Red squares indicate obstacles, green squares indicate the obstacles which exert forces to the target shown in the above figure. The total force is shown with a blue arrow. Purple dots show the estimated trajectory.

in order to avoid collisions with the obstacles. The rest of the figures, Figs. (4.6c)-(4.6f), show with green rectangles which of the environmental constraints exert forces onto this target at each time instance. The use of social forces allows us to achieve more accurate motion prediction (e.g. by avoiding obstacles, going through walls, etc.) and the predicted locations are more aligned with the actual observations which improves the final position estimates.

4.8 Evaluation

In this section we will show through a real-world setup the performance of the proposed system. In particular we will show how the fusion of camera, inertial and radio measurements and their integration into the Rao-Blackwellized particle filter allows us to achieve high accuracy with a relatively low cost system which uses only existing infrastructure and

smart-phones. We will demonstrate that the proposed system outperforms existing techniques and finally we will demonstrate its robustness in challenging conditions such as occlusions, missing detections and sensor noise.

(a) (b) (c)

Figure 4.7: We have conducted real-world experiments in a construction site in order to evaluate the performance of the proposed system. (a) The construction site on day 1, (b) the construction site on day 36, (c) floor-plan of the site.

4.8.1 Experimental Setup

In order to evaluate the performance of the proposed approach we have conducted two real world experiments in a construction site (Fig. (4.7)). In both experiments we placed two cameras with non-overlapping FOV at approximately 8 meters above the ground facing down. In the first experiment the two cameras were covering an area of approximately 11m $\times$ 9m each and in the second experiment an area of 14m $\times$ 4m each. The duration of each of the experiments was approximately 45 minutes with the cameras recording video at 30fps with a resolution of 960 $\times$ 720 px. We should also mention here that each camera was processed separately (i.e. we do not consider the multi-camera system scenario). The area of the site was outfitted with 12 WiFi access points and 5 workers were supplied with smartphone devices. The total number of people in the scene was varying from 3 to 12 as workers were entering and exiting the field of view. The objective of the experiment was to identify and track the workers who were carrying a smartphone device using camera, radio and inertial measurements. The radio measurements were obtained by their smartphones receiving WiFi at 1Hz and 10Hz respectively. The inertial measurements (i.e accelerometer and magnetometer) obtained from their smartphones had a sampling rate of 100Hz.

To obtain the ground truth of people's trajectories we followed the same approach as in RAVEL. We supplied all people to be tracked with helmets of different colours and their ground truth trajectories were obtained using a mean-shift tracker [98] to track the coloured helmets. We have decided to use the procedure above for obtaining the ground

truth trajectories since with GPS we could not get the required accuracy (i.e. GPS achieved a room-level accuracy during our experiments at the construction site) for this specific task.

In our implementation we have used RGB images as input to the MoG foreground detector, however we have not used any colour features for people identification and our filter tracks only the position of targets. Any target detector which outputs target coordinates can be used with the proposed technique without any changes to the algorithm. It is also worth mentioning that the proposed system can also be extended to utilise visual features (i.e. colour) for target identification, however these features are not always available and therefore cannot be relied on to uniquely identify the workers.

4.8.2 Results

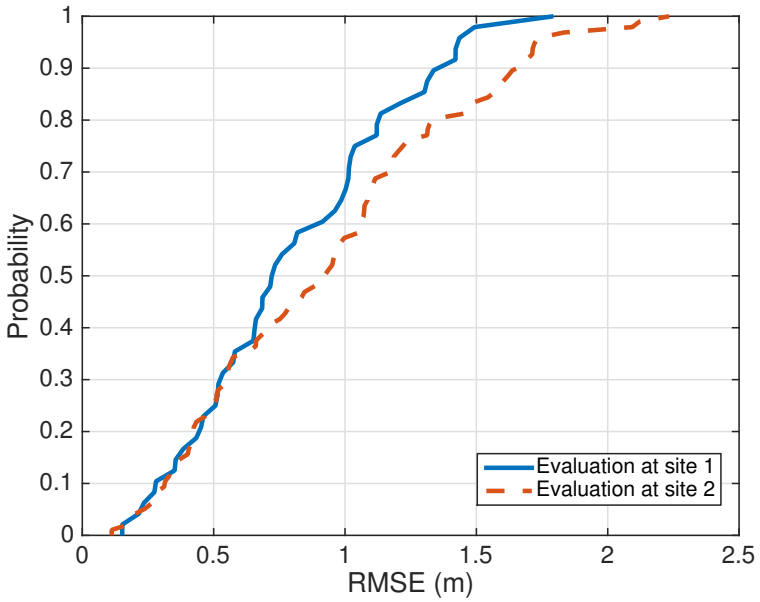


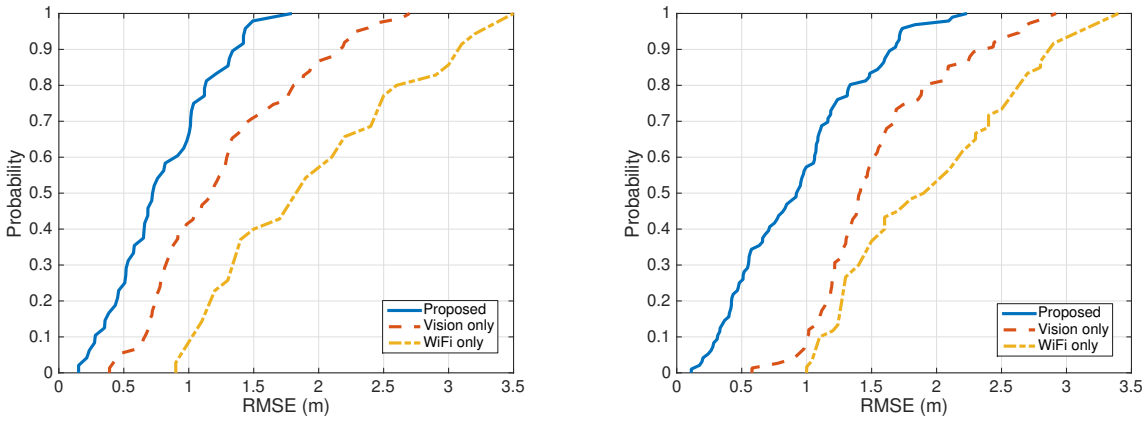
Figure 4.8: The figure shows the error CDF over over all targets in a construction site at two different days.

Accuracy at different sites:

The first set of experiments evaluates the tracking accuracy of our system (i.e. how well we can identify and track people with smartphone devices among all people in the FOV). Our performance metric in this experiment is the root mean square error (RMSE) between the ground-truth and the estimated trajectory. In all the experiments shown here we have used 100 particles to estimate the filtering distribution. For this test we have divided our dataset into time-windows of one minute (i.e. 1800 frames) each. Figure (4.8) shows the error CDF over over all targets in a construction site at two different days (Fig. (4.8)). As we can see from Fig. (4.8) the proposed techniques show similar results in the two tests. More

specifically, we were able achieve a 70 percentile error approximately below 1 meter in both tests. Furthermore, the proposed technique has a 90 percentile error of 1.4 meters and 1.6 meters for the first and second trial respectively. We would like to note here that these results were achieved with just 6 (out of the 12) WiFi access points and with one monocular stationary camera which makes the proposed system highly accurate and practical given its low-cost. The effect of the number of WiFi access points on the performance of our system will be discussed later in this section.

Comparison with competing techniques:



(a) Trial one

(b) Trial two

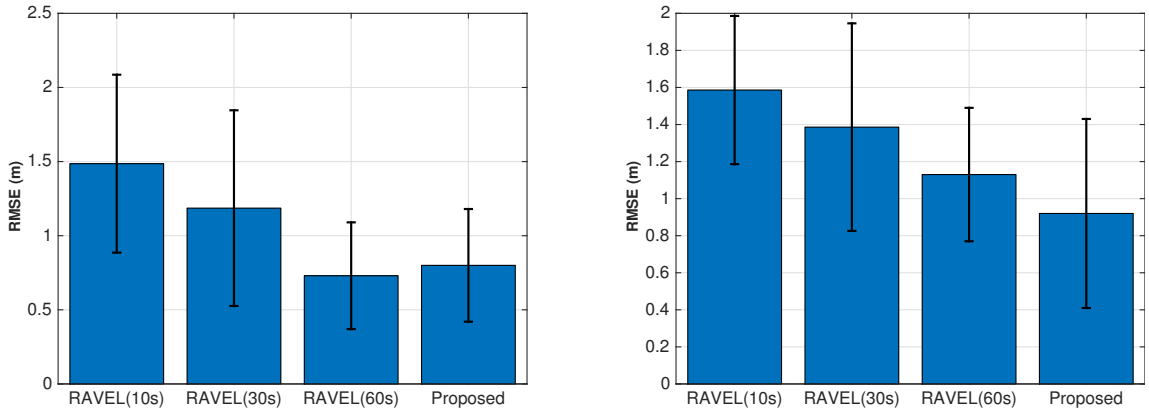
Figure 4.9: The figure shows the error CDF of the proposed system in comparison with vision-based and WiFi-based system. Our multi-modal multi-hypothesis tracking framework outperforms existing techniques.

Next we wanted to see how the proposed system performs compared to existing techniques which nowadays are widely used. For this reason we compare the proposed approach with the state-of-the-art visual tracking and fingerprinting techniques. The competing visual tracking technique is the original RBMCDA algorithm (referred to as vision-only tracker in this section) which uses Rao-Blackwellized particle filtering taking account only visual observations for multi-target tracking. For the competing WiFi-based positioning we have implemented the continuous space estimator of the Horus [22] fingerprinting system (termed as Radio only) by taking into account the 12 WiFi access points in the construction site environment. We should note here the the vision-only tracker does not support target identification whereas the proposed method does.

Figure (4.9) shows the results that we have obtained on our two trials in the construction site. More specifically, whereas our system achieves a 90 percentile error of about 1.4

meters, the vision-only tracker has an equivalent error of 2.3 meters and the WiFi-only system has a 90 percentile error of 3.1 meters. This is shown in in Fig. (4.9a).

In our second trial, (Fig. (4.9b)), the accuracy of the vision-only system decreased further achieving a 90 percentile error of 2.5 meters whereas the WiFi-only system has an equivalent error of about 2.9 meters. In this test the proposed approach performs significantly better. The sources of error for the vision-only tracker are due to the missing camera detections which lead to loss of tracking after just few iterations. The algorithm uses a constant velocity Kalman filter for tracking human motion which fails easily when we have missing camera detections and occlusions. Additional sources of error are due to association ambiguities such as when people cross paths. On the other hand our approach does not have the above problems as it uses a more accurate motion model and WiFi observations to identify and localise the targets which improves the tracking accuracy. Finally, the fusion of camera, inertial and WiFi observations used by our approach significantly outperforms the fingerprinting method. In these tests we have used just 6 WiFi access points for the proposed approach whereas the WiFi only method uses all 12.



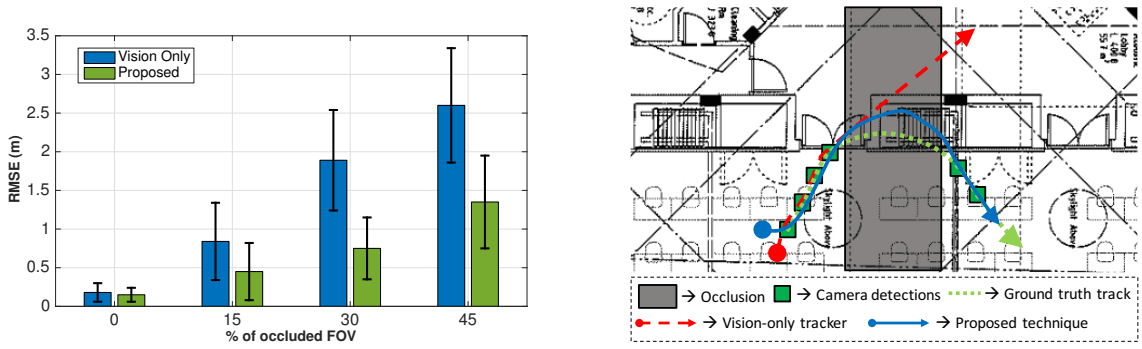
(a) Trial one

(b) Trial two

Figure 4.10: The figure shows the accuracy of our Rao-Blackwellized multi-fusion tracker compared to RAVEL for the two trials.

The next step is to compare the proposed technique with the RAVEL system which is also a multiple hypothesis tracking and identification system. RAVEL exploits the smoothness of motion and radio signal strength data in order to track and identify targets. Unlike our technique proposed in this chapter, RAVEL is more of a reconstruction technique (i.e. performs off-line tracking) as it requires to observe all measurements over a time window (W) in order to provide the trajectories of each target. We have tested RAVEL using time windows of sizes 10, 30 and 60 seconds in our construction site trials and we have compared

it with the proposed online system. We should note here that both systems use the same foreground detection technique and radio model with the same settings.



(a) (b)

Figure 4.11: (a) The figure shows the RMSE between the proposed technique and the vision-only tracking for different amounts of occlusion. The use of inertial measurements by the proposed technique improves tracking significantly in noisy scenarios. (b) Illustrative example showing the difference between vision-only tracking (red line) and the proposed approach (blue line) in the presence of occlusions (grey area). In cases of prolonged missing camera detections (green squares) the constant velocity model of the vision-only tracker is not sufficient enough to maintain tracking. On the other hand the proposed technique with the aid of inertial measurements is capable of closely following the target despite the presence of long-term occlusions.

In Fig. (4.10) *RAVEL(10s)*, *RAVEL(30s)* and *RAVEL(60s)* shows the accuracy of RAVEL for window sizes of 10, 30 and 60 seconds respectively. *Proposed* denotes the proposed system of this chapter which uses inertial measurements in addition to the camera and radio observations that RAVEL uses.

In the first trial shown in Fig. (4.10a) the average error of RAVEL decreases from 1.5m to 0.7m as we increase the window size. This is reasonable since it acquires more information over time and it can use this information to solve visual and motion ambiguities. On the other hand the proposed system is slightly less accurate from RAVEL(60s) however, it operates in real-time. In our second trial, shown in Fig. (4.10b), the proposed technique outperforms RAVEL(60s). The reason behind this is because the workers in this trial walk slower or they do not walk at all (i.e. they perform a task in place) which makes them invisible to the foreground detector (i.e. they cannot be detected as moving objects and they become part of the background). In these cases the motion model that RAVEL utilises is not efficient enough since it relies heavily on camera observations. In contrast the proposed technique does not have this problem as it used other sources of information (i.e. inertial) for motion prediction.

Robustness:

This set of experiments aims to demonstrate the robustness of the proposed technique. First we wanted to see how our technique performs on difficult trajectories (i.e. various amounts of occlusions and missing detections). In order to simulate occlusions we remove a specific area of the field of view (FOV) by disabling the camera detections inside that area.
More specifically, we generated occlusions at random locations that occupy a rectangular area of specific size inside the FOV. Then we evaluated the accuracy of the proposed approach compared to the vision-only tracker on 50 trajectories of variable length generated from our ground truth data. Fig. (4.11a) shows the RMSE over all trajectories between the proposed system and the vision-only tracker for different configurations of occlusions (i.e. shown as the percentage of occluded FOV). For each configuration we run the test 10 times; each time the occlusion was positioned to a different location.

The two methods achieve a comparable performance when there are no occlusions. However, the proposed approach significantly outperforms the vision-only tracking in scenarios with long-term occlusions and large amounts of missing detections. In the presence of long-term occlusions the constant velocity/acceleration motion model utilized by most visual tracking techniques fails and cannot be used to reliably model the inherently complex human motion. On the other hand Fig. (4.11a) shows that the use of inertial measurements by the proposed technique provides a more accurate model of human motion. An illustrative example is shown in Fig. (4.11b).

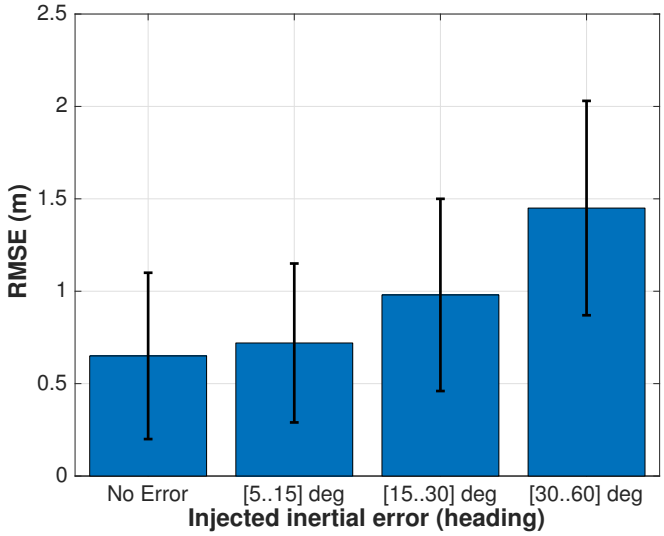


Figure 4.12: The RMSE of the proposed technique under different amounts of injected heading error.

Additionally in order to study how our approach can cope with variable noise from the inertial sensors we followed a similar procedure as in the previous paragraph and we generated 50 trajectories from our ground truth data. At each time-step and for each trajectory we inject a random bias error to the heading estimator. More specifically we sample a heading error uniformly from a specific range of the form $[a..b]$ $degrees$ and we add it to the output of the heading estimator. By doing this we can get an idea of how our approach performs in environments with disturbed magnetic fields. Fig. (4.12) illustrates the results of this experiment for different amounts of injected noise. As we can see the proposed technique can cope with moderate amounts of inertial noise; achieving a sub-meter accuracy for bias up to 30 degrees.

Moreover, we wanted to see how the number of people not being tracked in the scene affects the performance of our system and in addition what is the impact of visual noise on the tracking accuracy. In order to study this, we used 10 minutes worth of data (i.e. 18000 frames) from our construction site dataset. For each frame in this dataset we have superimposed visual objects from future timestamps in order to increase the visual noise and the number of people in the scene.

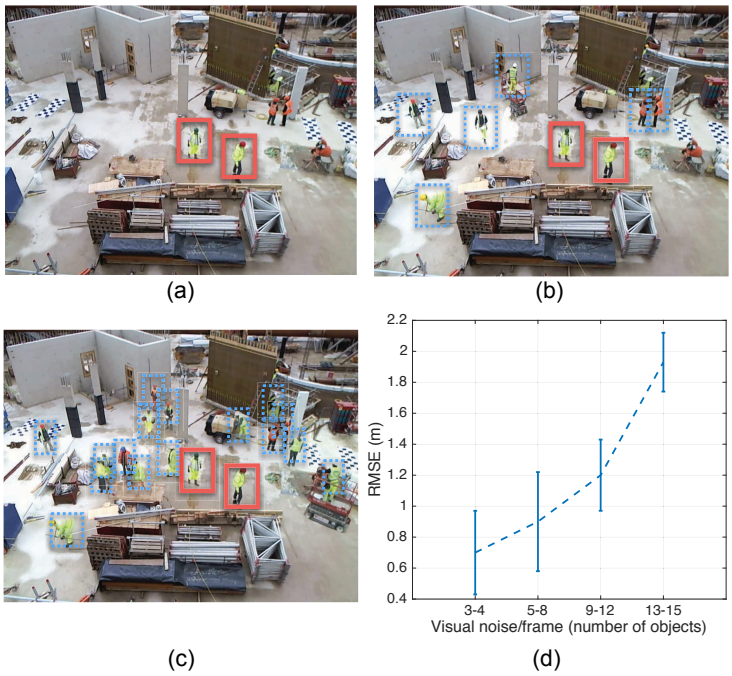


Figure 4.13: The RMSE of the proposed technique under different amounts of visual noise. (a) Camera snapshot without visual noise where we track the people in red rectangles. (b) Visual noise is injected in the scene (i.e. objects in blue rectangles). (c) Additional visual noise is injected in the scene. (d) The impact of visual noise on the performance of the proposed approach.

We have split the dataset in windows of 1 minute each (i.e. 1800 frames) and we recorded the RMS error for different number of visual objects as shown in Fig. (4.13). It is worth noting that the number of people that we track includes only the people which carry mobile devices (i.e. 5 people). As we can see from Fig. (4.13) as we increase the number of visual objects in the scene the accuracy drops. More specifically when we have relatively small number of objects in the scene (i.e. 3-4 per frame) the error is approximately 0.7 meters and increases to approximately 1.9 meters when the number of objects increases to 13-15 per frame. The reason behind this is due to the fact that the WiFi cannot distinguish close-spaced targets and despite the use of IMU data for motion prediction a track can incorrectly be updated with the wrong visual observation (i.e. visual noise). A possible solution to this problem, is to consider the evolution of the WiFi signal over multiple frames as opposed to the on-line filtering approach that we have currently implemented, this however precludes the used of on-line tracking. Additionally multiple overlapping cameras can also help but this will increase the system's cost and complexity. It is also worth noting that since our method only tracks the people with mobile devices the injected visual-noise (i.e. people not being tracked) does not affect the time complexity of our method. In these scenarios we only observe a slight processing overheat during the data association step since now we have more visual measurements to process (i.e. decide to which track to assign each measurement).

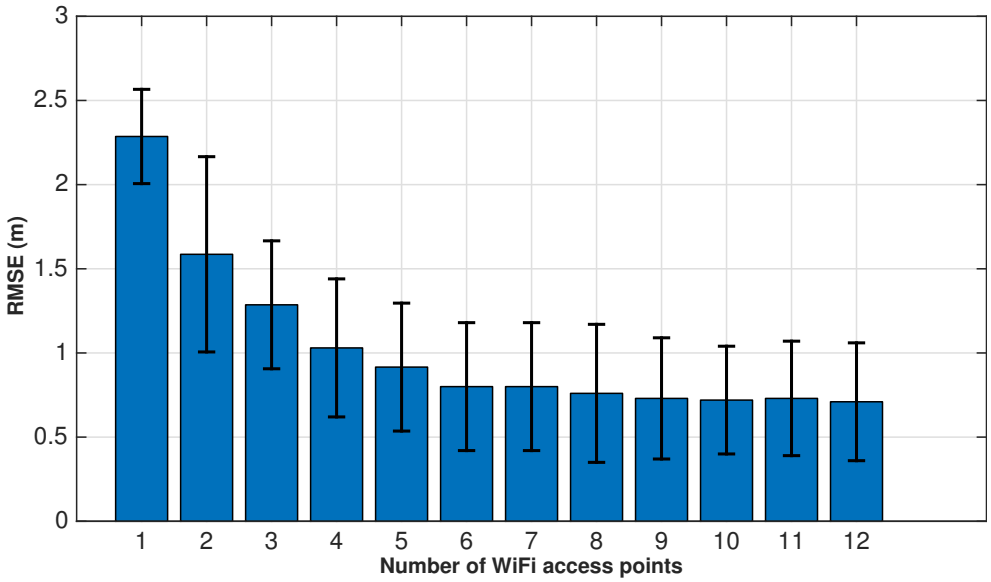


Figure 4.14: The figure shows the impact of the number of access points on the accuracy of our system.

Next, we investigate the impact of the number of WiFi access points on the accuracy. Figure (4.14) shows that as we increase the number of access points the error decreases

until it reaches a plateau at about 6 access points. As we can see from the figure, 1 to 3 WiFi access points are not enough in order to achieve sub-meter accuracy. This is because we do not have sufficient radio information to distinguish the different targets. In other words we cannot make reliable assignments between the id-linked (camera and inertial) measurements with the anonymous camera detections. The accuracy increases significantly with 4-5 access points reaching sub-meter level accuracy.

One of the requirements of this thesis is to design cost-effective positioning systems that rely on the existing infrastructure. However, the amount of existing infrastructure found in todays buildings varies depending on factors like location, building type, country, etc. Figure (4.14) provides us with useful insights on the performance of the proposed approach with respect to the existing infrastructure. We believe that the requirement for 4-5 WiFi access points in order to achieve sub-meter accuracy is reasonable for the building types we target in this thesis (i.e. airports, shopping malls, museums, large construction sites). Unfortunately, in environments with less infrastructure (i.e. 1-2 WiFi APs) we will not be able to achieve sub-meter accuracy with the proposed system. This is a tradeoff between accuracy and available infrastructure.

Impact of Social Forces:

The set of experiments shown in Fig. (4.15) aims to investigate the impact of social forces on the performance of our system. For this experiment we are using our extended system architecture as shown in Fig. (4.5) which includes the SFM module and uses the improved motion model given by Eqn. (4.12). With this system we can now take into account the interactions between people and the influence of the environment on human motion. Accurate human motion prediction is required in our system mainly due to the missing and noisy camera and inertial measurements.

Our intuition is that the addition of social forces will improve the motion prediction; thus increasing the overall tracking accuracy. Social forces essentially help us avoid predictions through obstacles (i.e. walls) and also help us model the interactions between people. In these tests we assume that people have a mass of 70Kg and a radius of 0.2m. The rest of SFM parameters are as follows $a_j = 50$N, $b_j = 0.5$m, $\lambda = 0.5$ and $c_j = 250$N/m. Figure (4.15a) shows the impact of social forces on the accuracy of our system for our two trials. As we can see the social forces improve the overall accuracy by approximately 20% on both trials.

The reason behind these improvements is due to the more accurate motion prediction, which more realistically models human behaviour. A second reason for these improve-

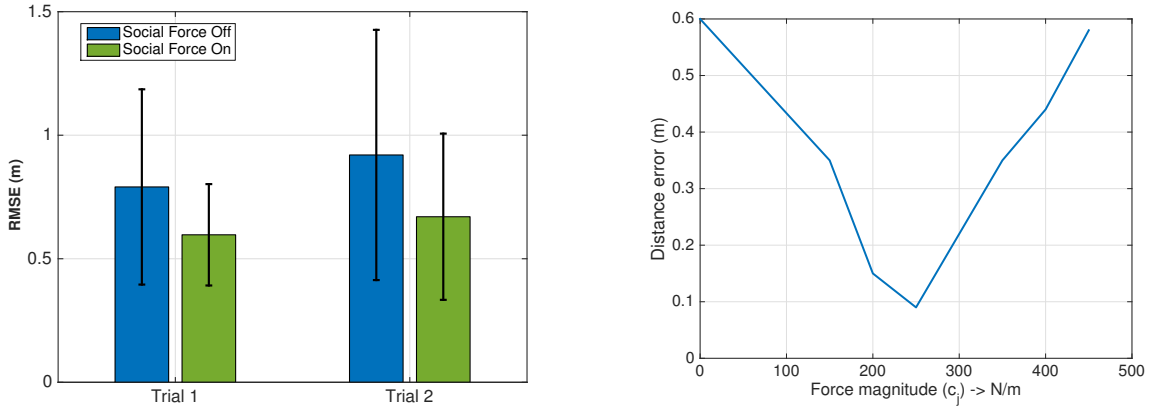


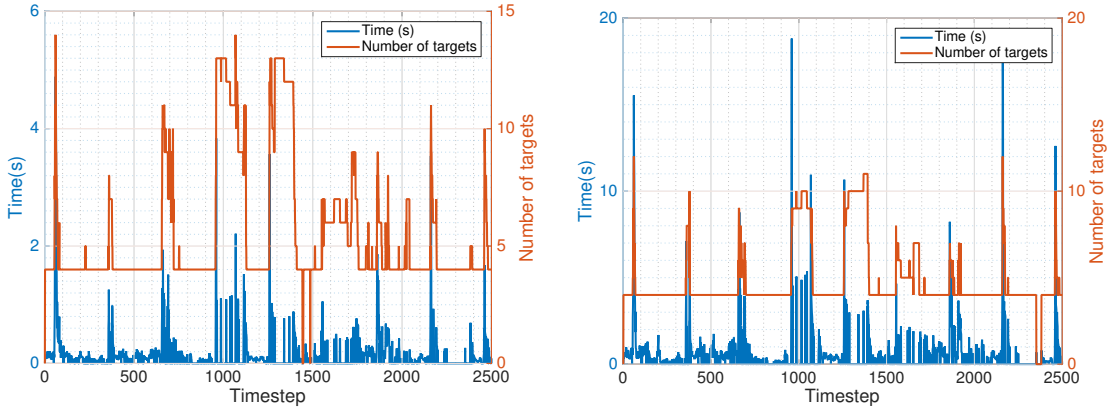
Figure 4.15: (a) Impact of Social Forces on the performance of our system. (b) Tuning the parameters of the social force model. The graph shows the impact of the force magnitude (c_j) from Eq. (4.10a) on the accuracy of the system.

ments is that now the predicted locations are more closely aligned with the actual observations, which improves the final position estimates.

Finally, we should note that correctly selecting the parameters of the social force model is very important if we would like the SFM to be beneficial and improve tracking accuracy. Figure. (4.15b) shows the impact of the force magnitude parameter (c_j) from Eq. (4.10a) on the accuracy of the system in the case where an obstacle blocks the trajectory of a person. More specifically, the graph shows that when $c_j = 0$ N/m the social forces have no effect on the human motion, as a result the predicted position goes through the obstacle which results in an error of about 0.6 meters from the correct location. As we increase the force magnitude social forces exerted onto this person which are opposite to his/her motion direction prevent them from going through the obstacle. This results in an improved position estimate as shown in the figure for $c_j = 250$ N/m. However, when the force magnitude is too large i.e. 450 N/m the repulsive forces push the person too far away from the correct location which results in errors as shown in (4.15b). We have found experimentally that the SFM works best if it is tuned so that it would point towards the right direction but without causing significant repulsion. This strategy allows us to have improved location predictions that align better with the actual observations.

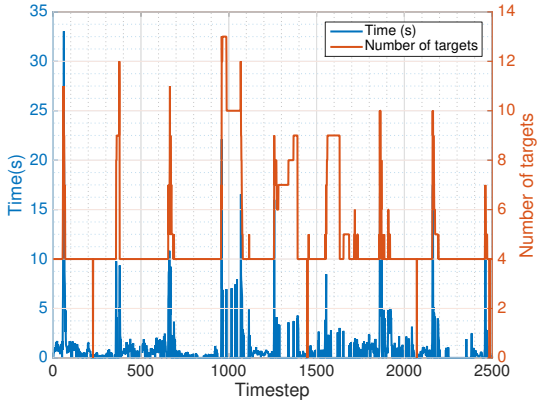
Profile analysis:

Finally, we have performed a timing analysis of the proposed system in order to determine if it is capable of real-time operation. We should note here that all tests were performed on a mobile Intel CPU i.e. Intel Core M 5Y51 clocked at 1.10GHz with 4MB of cache. Our system had 8 GB of DDR3 RAM clocked at 1600 MHz.

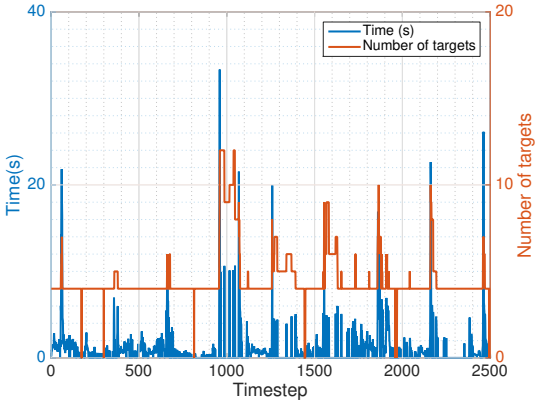


(a) 10 particles

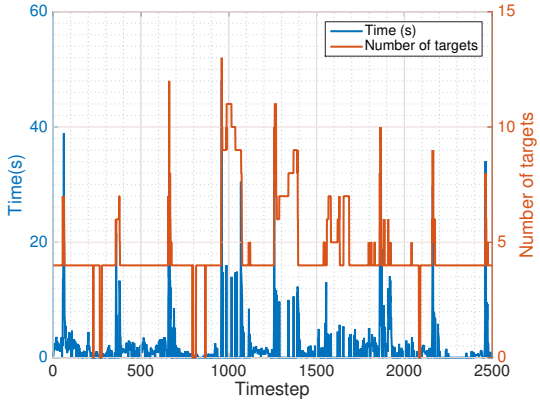
(b) 50 particles



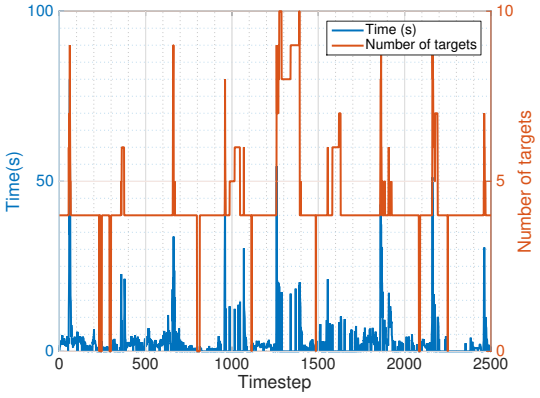
(c) 70 particles



(d) 100 particles



(e) 150 particles



(f) 250 particles

Figure 4.16: Computational time with respect to the number of particles and number of targets

All the analysis shown in this section concern our unoptimised Matlab implementation. Before going through the tests we would like to remind here that in our Rao-Blackwellized particle filter each particle tracks all targets thus each particle maintains k Kalman filters,

where k is the number of targets. In other words at each time step each particle encodes a different realisation of the whole system which coexists with the others and progresses parallel in time.

The first set of experiments shows the computational time of the proposed system with respect to the number of particles and the number of targets. More specifically, Fig. (4.16) shows the computational time of the proposed system over a period of 2500 steps with respect to the number of particles and the number of targets. We should note here that the number of targets shown in this figure is estimated based on the particle with the highest weight, i.e. it is not necessary that all particles contain the same number of targets, still however this provides us with an estimate for the number of targets with high probability. As we can see from Figs. (4.16a)-(4.16f) the computational time spikes whenever the number of targets increases in all six plots. This is reasonable since the size of each particle increases, with the addition of Kalman filters in order to track the new targets. In addition, we observe that the computational time it takes for one iteration also increases as the number of particles increases. This is because the number of particles represents the number of copies of our system which coexist at the same time, which makes each iteration of the system computationally more expensive.

Profile Analysis			
Number of Particles	**Min time (s)**	**Max time (s)**	**Avg time/iteration (s)**
10	1.37E-04	4.88	0.13
30	2.94E-04	12.79	0.33
50	4.24E-04	18.82	0.52
70	6.06E-04	33.03	0.76
100	8.43E-04	33.33	0.93
130	0.0011	56.48	1.41
150	0.0012	48.10	1.44
200	0.0016	79.91	1.74
250	0.0019	91.38	2.12
300	0.0020	183.03	2.84

Table 4.1: Profile analysis of the proposed system. The table shows the computational time with respect to the number of particles over a period of 2500 steps. The first column indicates the fastest iteration over this period, the second column indicates the slowest iteration over this period and the last column shows the average time per iteration over the 2500 steps.

Table (4.1) shows in a more compact way the timing results of this experiment. As the number of particles increases the average time per iteration increases. More importantly, we can see that the slowest iteration can take 4.88 seconds to complete for the case of 10 particles whereas in the case of 300 particles this number becomes as high as 183.03 seconds. This behaviour is due to the large number of targets that need to be tracked during those iterations (also shown in Fig. (4.16)). These numbers have been obtained on a low-frequency, ultra-low-voltage mobile processor with our unoptimised Matlab implementation. In principle, the proposed algorithm is highly parallelizable as we can compute each particle in parallel. Each particle provides a complete realization of the system and tracks all targets thus the particles are independent of each other and the computation can take place on a different core/thread. This would reduce the computational time significantly and with the appropriate system can achieve real-time operation. If we exclude the maximum iteration time, we can see that we can run the system as is for 100 particles faster than 1Hz.

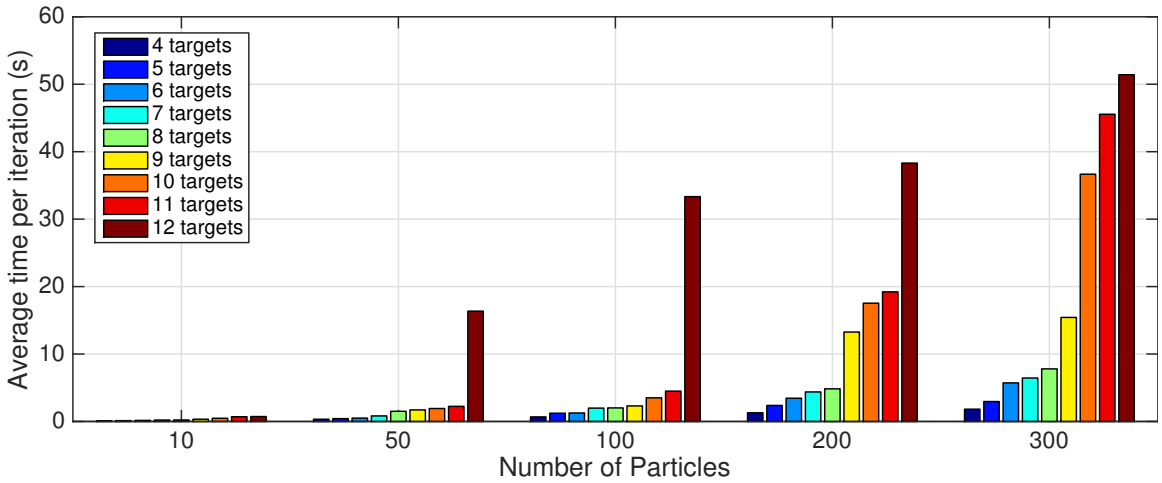


Figure 4.17: The figure shows the impact of the number of targets and the number of particles on the time it takes for our algorithm to complete one iteration.

Figure (4.17) shows more clearly the average time it takes for one iteration to be completed with respect to the number of targets and the particles used. For instance, when we use 300 particles, each time the system tracks 12 targets it requires on average approximately 51 seconds to process the data. We can observe an increase in the computational time with respect to the number of targets and the number of particles. This however is an expected outcome that often appears in multi-target particle-based systems. A parallel implementation of this system can help us to reduce the processing time and achieve real-time operation in all circumstances, although the current implementation allows for a near real-time operation for a moderate number of particles and targets.

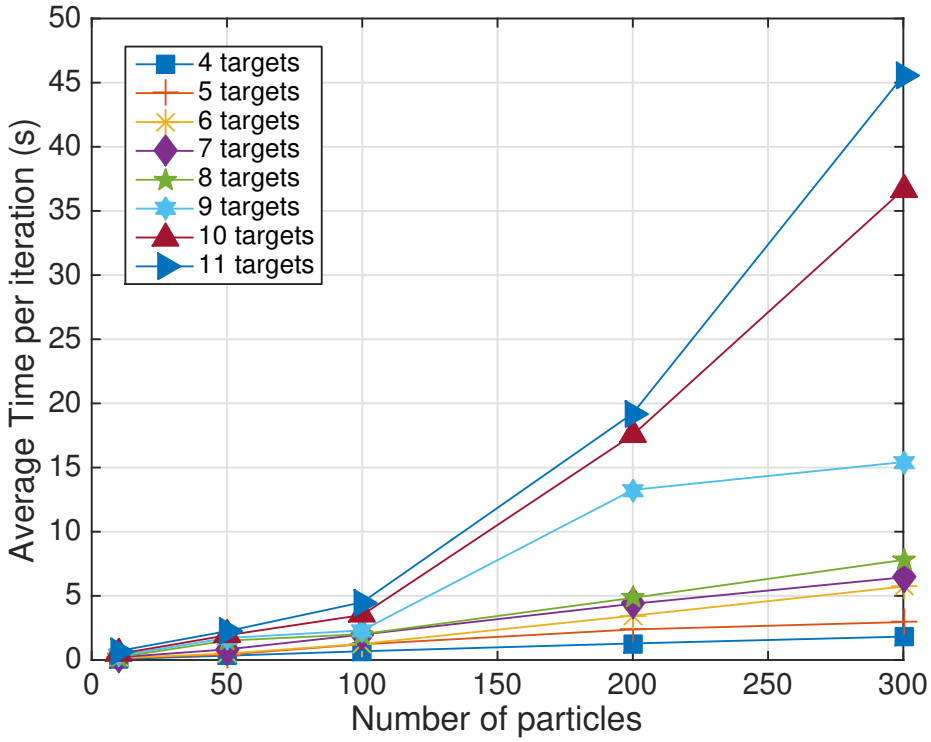


Figure 4.18: The figure shows the increase in processing time of the proposed technique as the number of target increases.

Figure (4.18) provides more insight on the computational complexity of the proposed technique and how this method scales as number of targets and number of particle increases. As we can see from this figure the processing time for one iteration increases significantly as the number of targets to be tracked increases. This trend is more clear when the number of target increases from 9 to 11. Because, in the proposed system each particle tracks all targets multiple copies of the same target co-exist in time and this increases the computational complexity of this approach. In essence, when we need to track X targets with N particles our approach creates $X \times N$ Kalman filters which run the time update and measurement correction equations. This increases the computational time for one complete iteration as shown in the graph. One additional reason for this increase in the time complexity is due to the larger number of data associations that arise as the number of targets increases. In other words as the number of targets increases the number of associations that we need to enumerate increases. This adds an additional overheat as in each time-step we need to enumerate all association events and create the proposal distribution from which we sample association probabilities. As we have already mentioned a parallel implementation of the proposed technique in which each particle is executed on a different cpu/thread will make things more manageable. However, reducing the computational com-

plexity of particle filter methods is an open research problem. In this study we have used Rao-Blackwellized particle filters which allows us to perform some parts of the computation analytically and others by sampling. This is the state-of-the-art particle filter method in terms of computational efficiency.

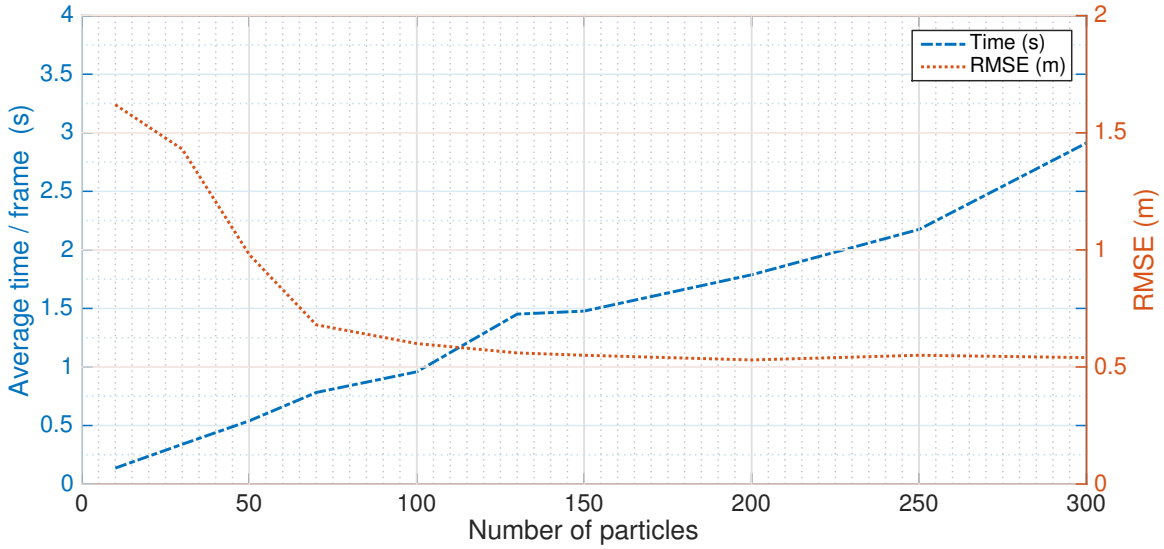


Figure 4.19: The figure shows the impact of the number of particles on the accuracy and the timing requirements of the proposed system.

Finally, we examine the impact of the number of particles on the accuracy of the system. In order to do that we have tested the proposed system over the same 2500 iterations and we measured the accuracy with respect to the number of particles. In addition we measure the average time per iteration with respect to the number of particles. Our findings are presented in Fig. (4.19). As expected the number of particles improves the accuracy whereas at the same time increases the computational time. It is however interesting to see that we can achieve a sub-meter accuracy with just 50 particles. Moreover, the effect of the number of particles reaches a plateau at approximately 150 particles. We should note here that on a different dataset and/or problem the number of particles that are required for optimal performance might be different. However, from the real-world experiments we have conducted, our intuition says that the proposed system can reach near optimum performance with approximately 100-150 particles in other tracking scenarios as well.

4.9 Discussion

In this chapter we proposed a multi-modal positioning system for multiple-target tracking. We showed that it is possible to adapt Rao-Blackwellized particle filters - traditionally

used to discern tracks using anonymous measurements - in order to both identify and track people being monitored by CCTV and holding mobile devices. Additionally, we showed that the use of social forces is highly beneficial and improves the tracking accuracy. We have further investigated the feasibility of real-time operation through a profile analysis of the proposed system. According to the timing analysis performed, the system is capable of real-time tracking without requiring heavyweight modifications for a moderate number of targets and particles. Our experiments showed that our online approach achieves similar positioning accuracy to the existing offline RAVEL approach. We also showed that the proposed technique is robust in scenarios with visual and inertial noise. Lastly, with the integration of social forces we improved the accuracy by 10-20%.

In the next chapter we extend further our system with learning capabilities. More specifically we are interested in robust and accurate tracking in challenging conditions and environments. For this reason we have developed cross-modality training techniques to automatically learn the internal parameters of our system and make it adaptive to changing conditions.

Chapter 5

Cross-modality Learning

5.1 Introduction

In the previous chapter we proposed a novel multi-modal multi-target tracking framework for applications that require real-time tracking and high accuracy. Moreover, we have shown how to improve human motion prediction with the fusion of inertial measurements and social forces into the target dynamical model. Additionally, the proposed system is capable of maintaining tracking under long term occlusions and it is robust to inertial noise (i.e. heading bias).

However, during our experiments in the construction sites we have observed some additional challenges which make the problem of tracking in these kinds of environments even more interesting. More specifically, tracking workers in a construction site and in similar industrial settings is much more challenging compared to other environments (e.g. museums, parks, etc.) mainly due to the many moving parts and the fast large-scale changes that occur in these complex environments. For instance, in most environments, the positions of walls and floors remain constant over time. In construction sites however this is not true. For example, Fig. (5.1) shows the effect of a wall being installed in the middle of one of our tracking experiments. The received signal strength of a worker's smartphone from one of the access points dropped considerably after the installation of the wall, in a matter of minutes. In addition to these short changes, during our experiments we observed significant long term changes within periods of a few weeks, the scene changed dramatically, staircases or entire floors were added, obfuscating the view to the first floors and creating additional layers where people needed to be tracked. Moreover, the radio and magnetic maps proved unstable with the movement of large structures.

Existing positioning systems have been designed to operate within environments that exhibit long-term stable macro-structure with potential small-scale dynamics. They leverage this environmental stability to provide accurate location services. In contrast, in highly

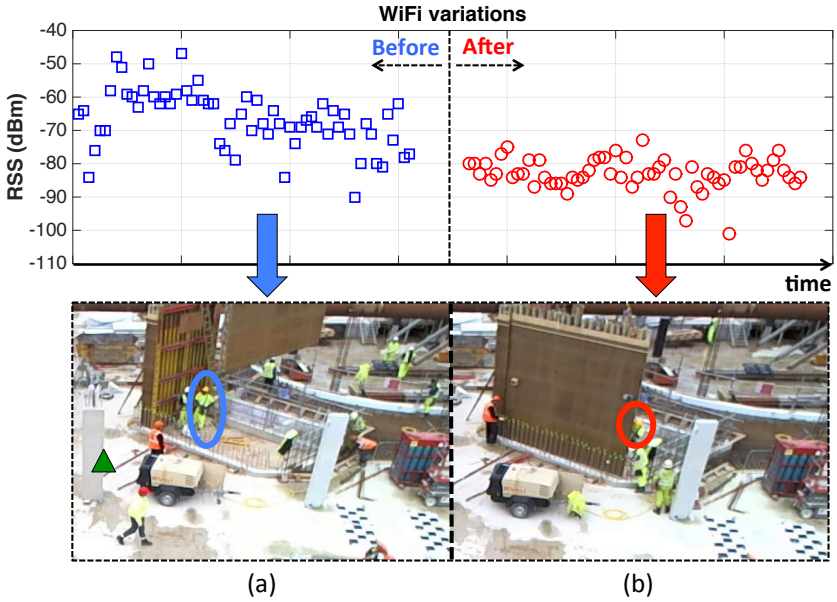


Figure 5.1: The WiFi signal strength received by the worker in circle is affected by the installation of a new wall (a) Before the installation there are direct WiFi signals from the access point (shown as triangle) to the worker, (b) The worker is blocked by the new wall, which affects the propagation properties of the WiFi signals as shown in the graph above.

dynamic environments these assumptions are not valid. Currently, there is no system that can provide reliable and accurate tracking in such environments.

In this chapter we propose a cross-modality learning framework which extends our system with learning capabilities and makes it suitable for highly dynamic industrial environments.

5.2 What to learn

To make our system adaptive and robust in challenging and dynamic environments we have looked further into the real-world problems to identify the major improvements needed in order to build such a system. During our trials in construction sites we have identified the following sub-systems in our proposed technique that need to be extended with learning capabilities:

- **Radio model:** As shown in Fig. (5.1) the changing environment in a construction site greatly affects the propagation properties of radio signals precluding the use of systems which rely on stable, long-term maps or use pre-trained radio models for positioning. Because of this we need to be able to learn the parameters of the radio propagation model in our system so that it can be used reliably in all conditions.

- **Visual detection:** One of the most important components of our system is the foreground detector which we use to detect the moving people in the scene. Without reliable object detection all tracking systems fail miserably, ours as well. We looked further into our foreground detector implementation and we have identified which parameters play significant role on the object detection rate. It turns out that the *learning rate* of the Mixture of Gaussians background subtraction technique is responsible for how well the algorithm learns to distinguish the foreground pixels from the background i.e. segmenting out the moving targets from the background. This parameter requires careful and labor intensive tuning in order to acquire its optimal value. Thus we need to make our system be able to learn the MoG learning rate automatically without relying on manual and time consuming tuning.

- **Step-length:** Human motion modelling and prediction is an essential component of modern tracking systems. Our system uses inertial observations for the target dynamics in order to achieve better motion modelling. This however depends on how well we use our data. To be more specific we calculate the step-length of people using a pre-trained empirical model. This model combines the user's step frequency and height with a set of three parameters for estimating the step-length. In our initial implementation we use a pre-trained universal model (i.e. same parameters for all users). We will show however in this chapter how we have extended our system in order to learn automatically a personal step-length model for each person and improve the motion prediction and accuracy.

- **Maps:** Finally, we have extended our approach to learn *occlusion* maps i.e. maps that show inaccessible areas, obstacles and obstructions. These maps model the environmental constraints and can be used dynamically by our system to aid tracking via the use of social forces.

5.3 System Architecture

Figure (5.2) shows the complete architecture of our system with the addition of the cross-modality technique which we are going to discuss next in this chapter. At each time step the final system receives as input camera, radio and inertial observations of all people to be tracked (i.e. people with smart-phone devices). It has three pre-processing modules: a) a foreground detector, b) radio model and c) an inertial tracker which it can be further decomposed into a step detector, step-length estimator and heading estimator. The social

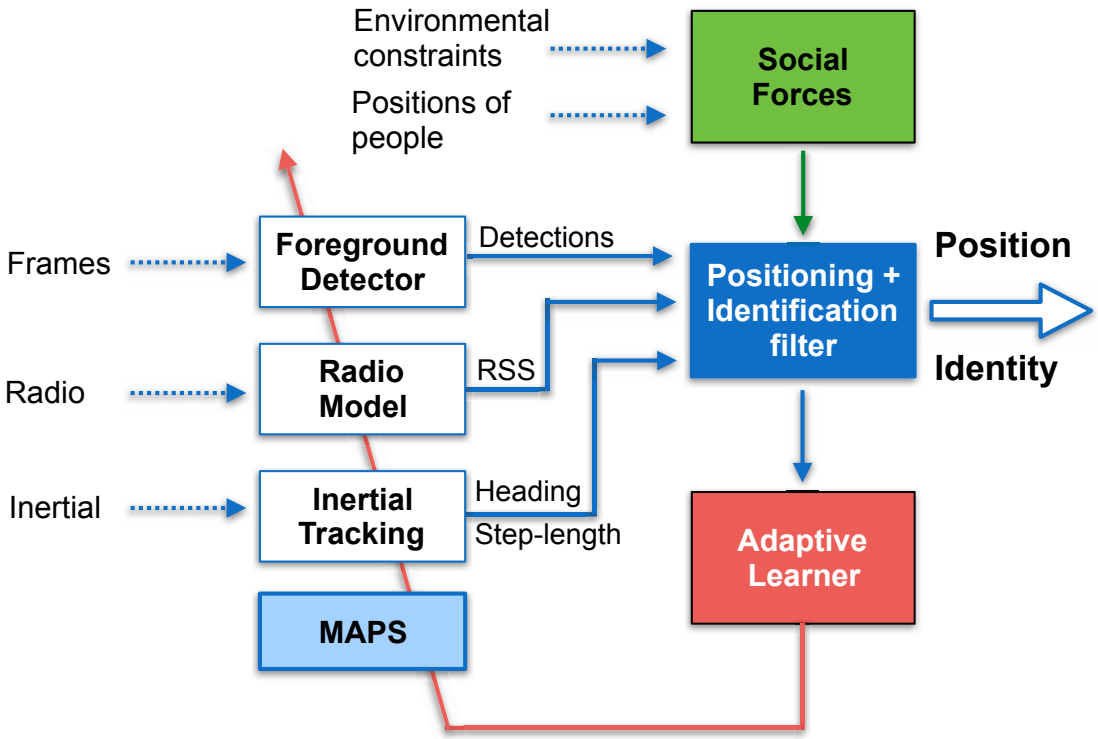


Figure 5.2: The figure shows the proposed system architecture after the inclusion of cross-modality learning. We use the output of our tracker to learn different parameters of our system's components.

forces module is responsible for modelling the interactions between people and the influence of the environment on human motion. Next, the Positioning and Identification filter which we have described in Chapter 4 in detail obtains all input data and performs the core process of tracking. The new addition in this revised architecture is the *Adaptive Learner* module. This module uses the output of the filter as shown in Fig.(5.2), in combination with the input observations, and performs *cross-modality training*. More specifically it performs the following tasks:

- It configures the foreground detectors' internal parameters by taking into account available motion measurements.
- It tunes the step-length estimation method by leveraging reliable camera measurements.
- Finally, it exploits camera measurements to learn the radio model, occlusion and magnetic maps which can be used to further improve the systems' accuracy.

5.4 Track Quality Estimation

As we have briefly mentioned the output (i.e. track) of our *Positioning and Identification filter* can in certain cases be used to learn the parameters of various internal components of our system. To motivate this we first consider the case of a single target moving inside the field of view. If we assume that the visual detector has 100% detection rate we can obtain the track of this target by simply taking the camera detections without even requiring the use of a tracker (e.g. Kalman filter, particle filter, etc.).

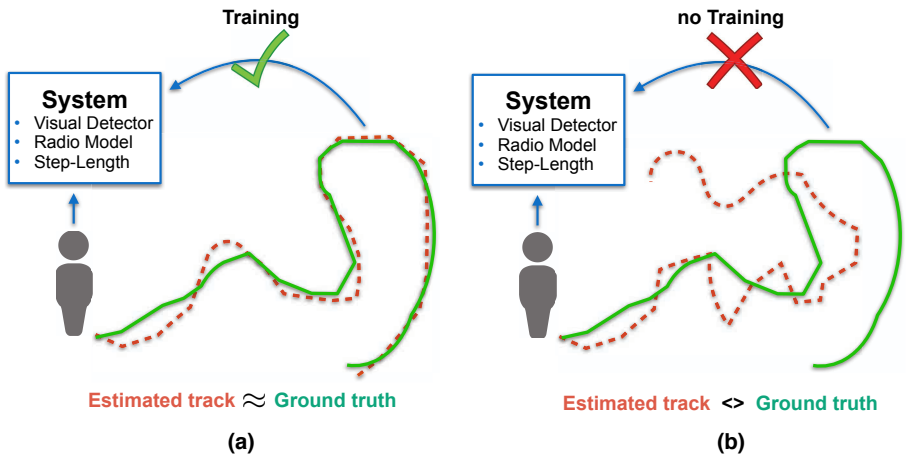


Figure 5.3: The figure illustrates the idea of cross-modality training. The output track of our system can be used in combination with the input observations to learn the internal parameters of our system. (a) In order to use cross-modality training we need to assess the quality of the output track i.e. how close to the ground truth this track is, (b) the output track cannot be used for cross-modality training if it diverges significantly from the ground truth.

Under this assumption the "estimated" trajectory which we have obtained by looking at the visual detector is an exact copy of the ground truth trajectory (assuming the camera system provides accurate positioning). Now, since we have the track of this target we can use it to train other parameters of our system. We should note here that the track of a target has the radio and inertial measurements linked to it which are specific for this target. This is however an ideal scenario since the visual detector is not 100% accurate. In addition in the real-world we have sensor noise, occlusions and multiple targets which makes things harder. The idea however remains the same that in certain situations the output track can be used to learn other parameters of the system which we call cross-modality training. What remains is to find a way to assess the quality of a track i.e. how close the output track is to

the ground truth *without requiring the knowledge of actual ground truth trajectories* (Fig, (5.3) illustrates this idea).

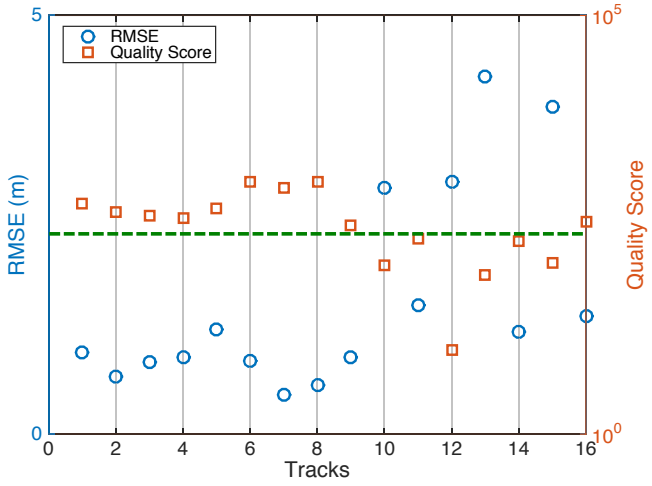


Figure 5.4: Track quality estimation: The figure shows the quality score of 16 tracks along with their RMSE. Tracks with quality score above the horizontal dotted line are considered qualifying and can be used for cross-modality training

Our objective is to assess the quality of output tracks to make sure that they qualify for the training process. Thus, the goal of the *Track Quality Estimation* phase is to find candidate tracks which can be used for cross-modality training.

Let us assume that at time-step (or scan) t we receive m measurements $\{y_t^1, y_t^2, ..., y_t^m\}$. In addition y_t^0 is defined for each time-step to be a dummy variable indicating the possibility of a missed detection. Then the incremental quality score of a track j during this time-step is defined as:

$$\Delta L_t^j = \begin{cases} \log\left(\dfrac{\hat{p}(y_t^i|\lambda_t = j)p_d}{\hat{p}(y_t^i|\lambda_t = 0)}\right) & , \text{ if } \exists\ i \in [1..m] \text{ s.t } \lambda_t = j \\ \log\left(1 - p_d\right) & , \text{ otherwise} \end{cases}$$

where the quantity $\hat{p}(y_t^i|\lambda_t = j)$ is the likelihood of the measurement assigned to track j. The term $\hat{p}(y_t^i|\lambda_t = 0) = p(clutter)$ is the likelihood of the measurement originating from clutter which has a uniform probability density over the measurement space of area S (i.e. $p(clutter) = S^{-1}$) and finally p_d is the probability of detection. Then, the cumulative quality score of track j is given by:

$$Q_j = \sum_{t=1}^{T} \Delta L_t^j \tag{5.1}$$

where T is the total length of the track. As we can see the quality score Q of a track penalizes the non-assignments due to missing detections while favoring the correct measurement-to-track associations. Fig. (5.4) shows that the quality score is negatively correlated with the root mean square error. Finally, in order to mark a track as a high confidence track that *qualifies* for cross-modal training its quality score is tested against a pre-determined threshold Q_{Th}. If $Q_j \geq Q_{\text{Th}}$ then the track is qualified (i.e. *high quality track*) and it can be used for cross-modality training, otherwise the track is rejected (Fig. (5.4)).

5.5 Foreground Detector Training

The mixture of Gaussians (MoG) [91] foreground detection which is used by our system is one of the most popular approaches for detecting moving targets from a static camera. This approach maintains a statistical representation of the background and can handle multi-modal background models and slow varying illumination changes.

In the original algorithm the history of each pixel is modeled by a mixture of K (typically 3-5) Gaussian distributions with parameters $(\beta_k, \mu_k, \sigma_k I)$ for the mixture weight, mean and covariance matrix of the k_{th} Gaussian component. In order to find the pixels that belong to the background, the Gaussian distributions are ordered in decreasing order according to the ratio (β_k/σ_k); background pixels exhibit higher weights and lower variances than the foreground moving pixels. The background model is obtained as $B^* = \arg\min_B \left(\sum_{k=1}^{B} \beta_k > P_b\right)$ where P_b is the prior probability of the background. The remaining $K - B^*$ distributions represent the foreground model.

On the arrival of a new frame each pixel is tested against the Gaussian mixture model and if a match is found the pixel is classified as a background or foreground depending on which Gaussian component it was matched with. If no match is found the pixel is classified as a foreground and it is added to the mixture model by evicting the component with the lowest weight. When a pixel is matched, the weight of that k_{th} Gaussian component is updated using an exponential weighting scheme with learning rate α as $\beta_{t+1} = (1-\alpha)\beta_t + \alpha$, and the weights of all other components are changed to $\beta_{t+1} = (1-\alpha)\beta_t$. A similar procedure is used to update the mean and covariance of each component in the mixture.
The learning rate (α) controls the adaptation rate of the algorithm to changes (i.e. illumination changes, speed of incorporating static targets into the background) and is the most critical parameter of the algorithm. Fast learning rates will give greater weight to recent changes and make the algorithm more responsive to sudden changes. However, this can cause the MoG model to become quickly dominated by a single component which affects the algorithm's stability. On the other hand slow learning rates will cause a slower

adaptation change which often results in pixel misclassification. Over the years many improvements have been suggested by the research community that allow for automatic initialization and better maintenance of the MoG parameters [111]. More recent techniques [112, 113] address challenges like sudden illumination variations, shadow detection and removal, automatic parameter selection, better execution time, etc .

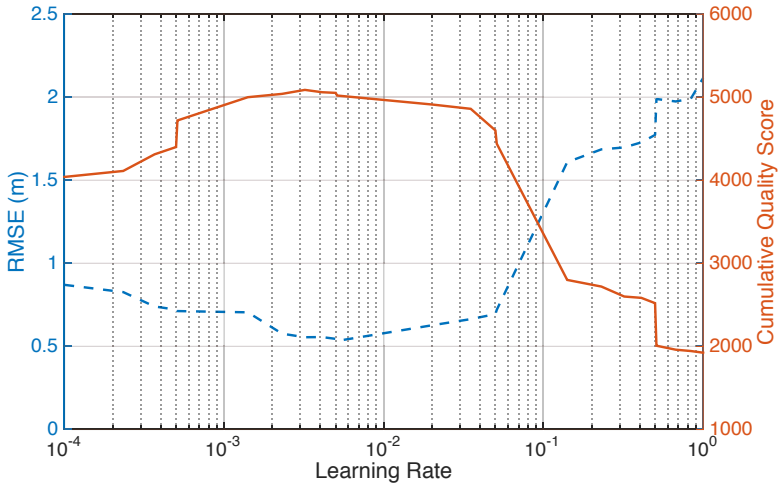


Figure 5.5: The figure shows the cumulative quality score (CQS) over a period of time as a function of the foreground detector learning rate (α). The optimum learning rate according to RMSE maximises CQS; we can use this metric to train the foreground detector.

In this section we propose a novel method for obtaining the optimum learning rate α^* of the foreground detector using the *high-quality* (i.e. pseudo ground-truth) tracks of our filter. Suppose we are given a track $X^j_{1:T} = \{x^j_1, x^j_2, ..., x^j_T\}$ of length T where $x^j_t, t \in [1..T]$ denotes the state of the track at time t. Since, both camera and inertial measurements could have been used to estimate track $X^j_{1:T}$ then its states $x^j_t, t \in [1..T]$ are of two types: type (a) states that have been estimated using camera and inertial measurements and type (b) states that have been estimated only using inertial measurements. A high-quality track ensures that $X^j_{1:T}$ contains the right mixture of type (a) and type (b) states and thus does not deviate significantly from the ground truth trajectory. This is possible, since propagating a track by only using inertial measurements is accurate enough for short periods of time. This key property of the inertial measurements allows us to use a high quality track as if it was the ground truth trajectory to train the learning rate of the foreground detector. In other words the type (b) states of a high quality track tells us that the target is moving to specific locations and the foreground detector does not detect any target at those locations.

The quality score of tracks (Eqn. (5.1)) can be used to find the optimum learning rate by solving the following optimization problem: *Given a time window $\mathcal{T}$ find a learning*

rate α^* so that the cumulative quality score (CQS) $\sum_j Q_j$ of all high quality tracks $j \in \mathcal{T}$ is maximized.

Figure (5.5) shows how our approach can find the optimum learning rate (α^*) of the foreground detector by solving the optimisation problem discussed in this Section. In this example we used 5 minutes of data, running the foreground detector for different values of (α) and calculating the cumulative quality score (CQS) for that period. Our intuition is that the optimum learning rate will reduce the number of missing detections, thus increasing the number of high quality tracks as well as their quality score. This is shown in Fig. (5.5) where the optimum learning rates achieve a high CQS, also evident by the low RMSE.

5.6 Optimizing the Step Length Estimation

Similar to the foreground detector training procedure, *high quality* tracks can also be used to learn the step-length model of each person being tracked. More specifically, the step-length of a user can be obtained from the universal model proposed in [114] as:

$$s = h(a' f_{step} + b') + c' \tag{5.2}$$

where s is the estimated step-length, h denotes the user's height, f_{step} is the step frequency obtained from the device's accelerometer and (a', b', c') are the model parameters. The model above describes a linear relationship between step-length and step frequency weighted by the user's height.

Since the heights of people that we need to track are not known a priori every time a new track is initialized that contains a sensor ID which has not been recorded before, the step-length estimator uses Eqn. (5.2) to provide an initial estimate of the target's step-length. At this point the height value is set to the country's average for men of ages between 25 and 34 years old. The parameters (a', b', c') have been pre-computed with a training set of 8 people of known heights using foot mounted IMUs.

As the tracking process proceeds high quality tracks are obtained periodically for each target. From these tracks the following IMU data are extracted for each step: a) step frequency, b) step start-time and c) step end-time. The start/end times of each step obtained from the IMU data are then matched to camera detections in order to obtain the position of the target during those times which are essentially the step-lengths measured from the camera system. Thus, for each target we obtain a collection of n calibration points $\{S\mathrm{v}^i, f^i_{step}\}_{i=1}^n$ where $S\mathrm{v}^i$ is the visual step-length of the i_{th} step and f^i_{step} its frequency obtained from the IMU. The calibration set of each target is then used to train a personal step-length model of the form $S\mathrm{v} = \varrho_1 f_{\text{step}} + \varrho_0$ using the least squares fitting. Finally, the

step-length estimator can switch to the trained model once the least squares goodness of fit $\left(R^2 = 1 - \frac{\text{residual sum squares}}{\text{total sum squares}}\right)$ exceeds a pre-defined threshold.

5.7 Radio Model

High quality tracks are also being used in order to learn the parameters of the radio propagation model which our system uses as explained in Section 4.6.1. More specifically, from a high quality track $X^j_{1:T} = \{x^j_1, x^j_2, ..., x^j_T\}$ of length T, the type (a) states are extracted. Let us call a type (a) state $\tilde{x}^j_t$; this state has been estimated using camera, radio and inertial measurements. Thus a collection of type (a) states $S = \{\tilde{x}^j_t : j \in K, t \in \mathcal{T}\}_n$ of length n where K is the total number of people with smartphones and $\mathcal{T}$ is the running time of our filter, contains n pairs of (location, RSS) measurements. Now, this collection of (location, RSS) points can be used to estimate the parameters of the log-normal radio propagation model [65] given by Eqn. (4.5) for each access point using least squares fitting. At regular intervals we re-estimate the radio model parameters based on the most recent portion of collected data. We should note here that the parameters of the radio model are initialised empirically based on a number of studies for different environments [65].

5.8 Learning Maps

Additionally, we can follow a similar procedure to learn radio, magnetic and occlusion maps (Fig. (5.6)). The radio and the magnetic maps can be combined and used for localisation in situations where the camera is occluded by an obstacle or they can be used in conjunction with the radio model to improve the system's accuracy. Additionally, the occlusion map, which is derived from the camera detections provides statistics about the environment (i.e. frequent visited areas, inaccessible areas, etc) which our system can use to improve its performance. For instance, suppose that a particular person is not detected by the camera during some time and our filter reverts to IMU tracking; the occlusion map can help us filter out impossible trajectories.

In order to learn the occlusion map we use the following procedure: We first discretize the world plane creating a 2D grid. During a time-window we then project the camera detections into the world plane and we count the number of hits in each cell creating a 2D histogram. The normalised histogram is then thresholded and the cells that are found to be below a predefined threshold are marked as occlusions/obstacles; this is shown in Fig. (5.7). The set of occlusions found $O = \{o_j\}_{j=1}^{N_o}$ are also used to model repulsive forces exerted from the environment onto people.

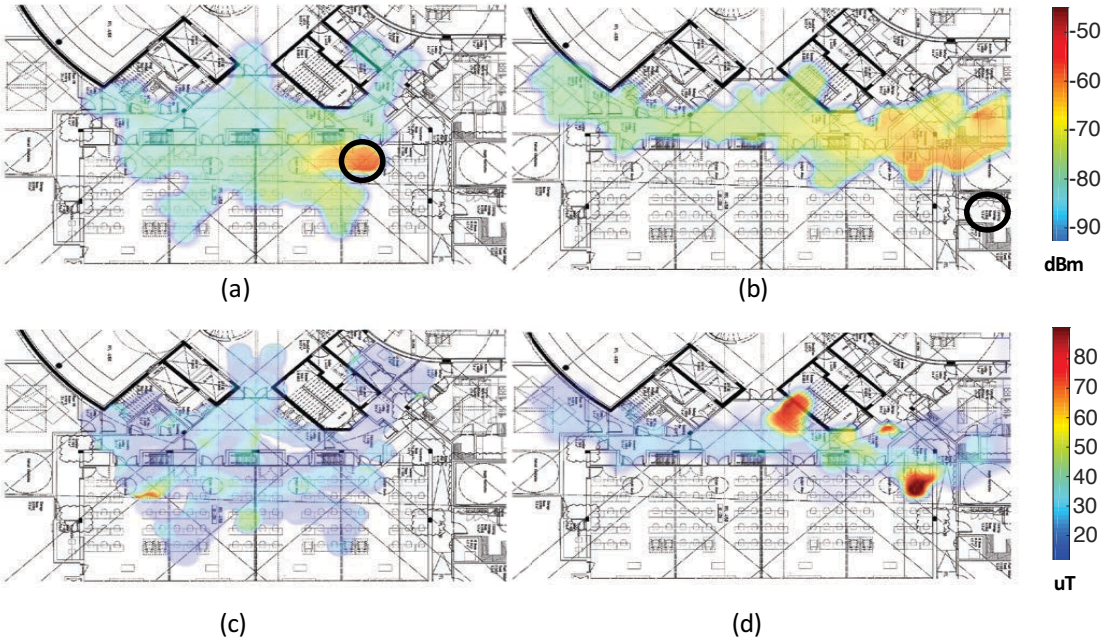


Figure 5.6: Radio and magnetic maps of the construction site: (a) WiFi map learned during the first day of the experiment showing the RSS with respect to one access-point (circle denotes the position of the AP), (b) WiFi map learned 36 days later for the same AP which was moved from its initial location because the ground floor had to been built, (c) magnetic map on day 1, (d) magnetic map on day 36. The environment constantly changes which makes the task of localisation and tracking very challenging.

5.9 System Evaluation

5.9.1 Experimental Setup

In order to evaluate the performance of our cross-modality technique we used the same setup and dataset we have described in Chapter 4. Specifically, we split the dataset into training and test sets. Each training set contains 20 minutes worth of data and the tests are 10 minutes long each. In total we have generated 5 training sets and 5 test sets for this evaluation. We first use the training set to learn the parameters of our system using cross-modality training and then we evaluate the performance of the learned parameters using the test sets.

We should also mention that the ground truth trajectories were obtained with the help of a mean-shift tracker [98] by tracking the colour helmets of the workers. In many cases however we had to intervene with manual labelling since the above technique is not 100% accurate.

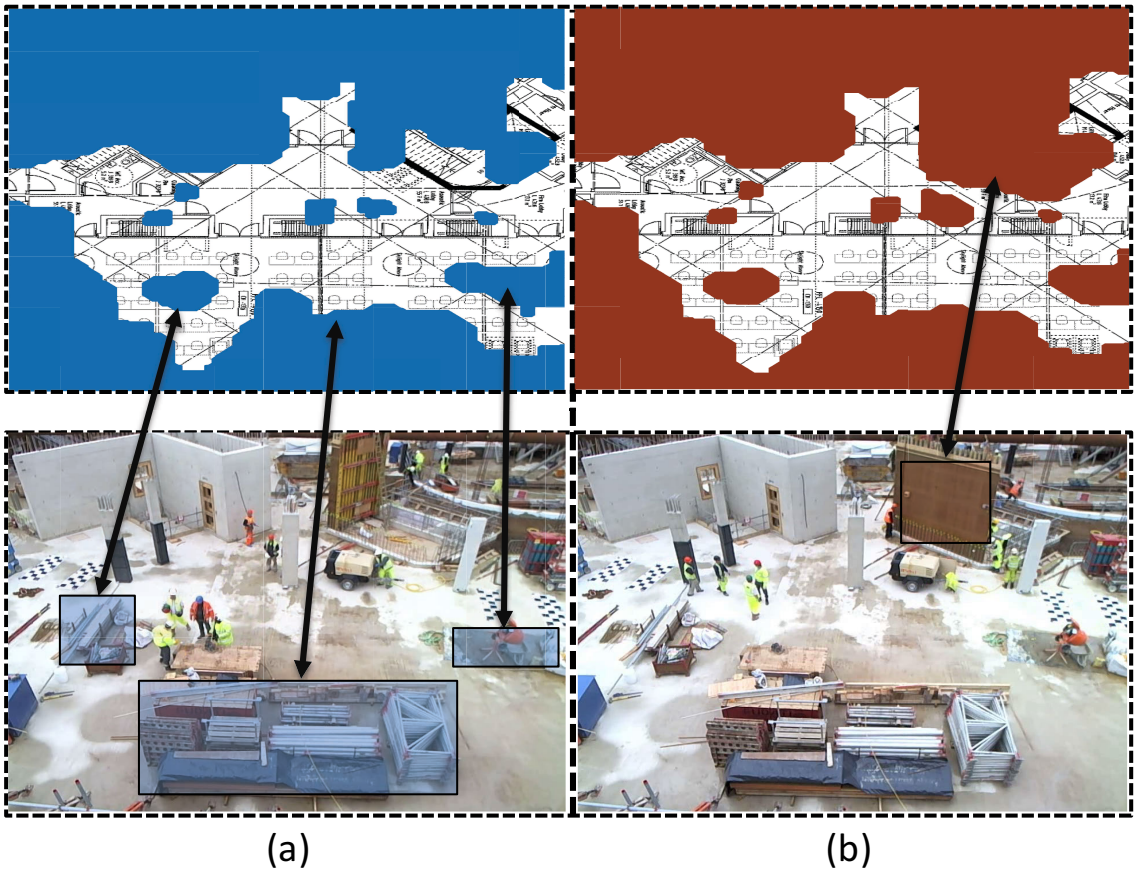


Figure 5.7: The figure shows the occlusion maps learned during a period of 10 minutes for each map. (a) Areas that appear to have no human activity are marked as occlusions, (b) As the construction site evolves new occlusions are created. In this case the installation of a new wall creates a new occlusion. These changes are detected automatically by our system and are used to improve the tracking accuracy via the use of social forces.

In addition, colour features where used only during the foreground detection phase and we are not using any colour or other visual features in our tracking filter. The proposed *Positioning and Identification filter* (Fig. (5.2)) maintains only the x-y position coordinates of a particular target; it can however be extended to use colour features as well.

5.9.2 Results

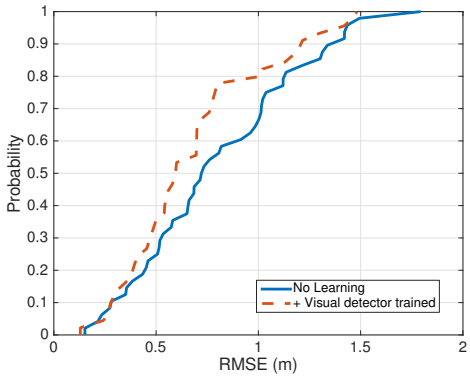
Accuracy:

Figure (5.8) shows the accuracy of the proposed system after applying cross-modality training. The results where obtained by running our filter on the test sets for time-windows of one minute (i.e. 1800 frames) for the two trials. Figure (5.8a) and (5.8b) show the effect of the foreground detector training on the tracking accuracy as function of RMSE in meters.

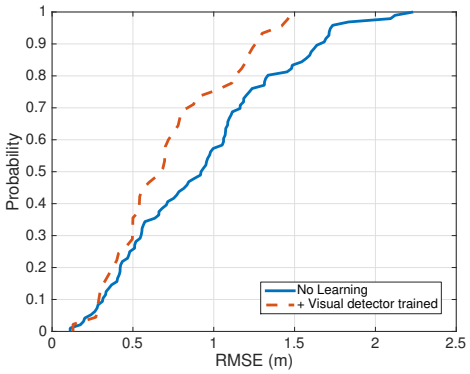
In the first trial the foreground detector training reduces the 70 percentile error of our system from 1m to about 0.7 meters and the 90 percentile error drops to approximately 1.2m. For our second trial the 90 percentile error after the foreground detection training improves by approximately 26% and the 70 percentile error drops to 0.8 meters. This increase in the tracking accuracy is due to the better background/foreground modelling which we achieve after tuning the foreground detector learning rate parameter. This allows the foreground detection module to give better object detection rate which in turn improves the overall tracking accuracy of our system.

Next in Figs. (5.8c) and (5.8d) we observe the accuracy results once we have trained the radio model as well. Figure (5.8c) shows that the 90 percentile error drops from 1.26m to just below 1m. The same is true for the second trial where we achieve sub-meter accuracy. As we can see the error decreases significantly once both the foreground detector and the radio model are learned. This is expected since our system requires both camera and radio measurements in order to determine the correct measurement to track association and update the target states. In the case of excessive missing camera detections, the trajectory of a target is estimated only by inertial measurements which is the main cause of the low accuracy. On the other hand, if the radio model was not trained, camera detections would not be able to be linked with radio measurements, which would also cause identification and tracking errors.

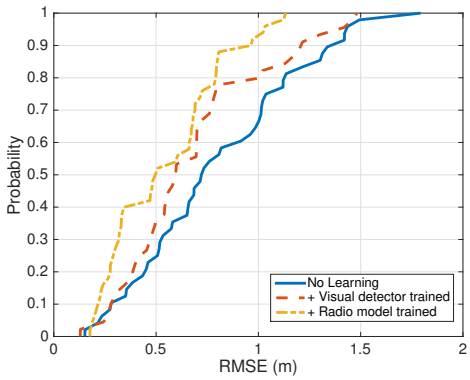
Finally, we show in Figs. (5.8e) and (5.8f) once the step-length estimator is also trained the performance improves further. In the first trial the 70 percentile error is reduced by approximately 18% whereas in the second trial this drops to about 13%. As we can see from the figure the improvement on the tracking accuracy is less obvious in this case. This however is reasonable, since once the foreground detector and the radio model are both trained most of the time the targets are updated with camera observations which are used to correct the IMU predictions. However, once the camera measurements become unavailable, the difference in accuracy between a trained and a universal step-length model is significant.



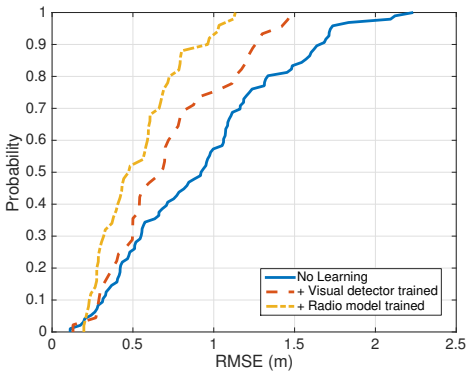
(a) Trial 1 - After training the foreground detector



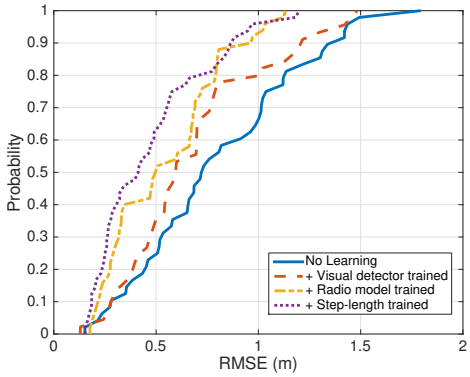
(b) Trial 2 - After training the foreground detector



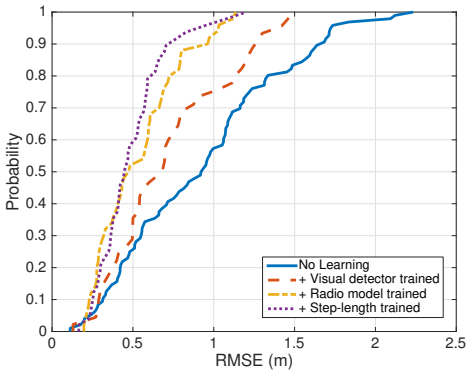
(c) Trial 1 - After training the radio model in addition to the foreground detector



(d) Trial 2 - After training the radio model in addition to the foreground detector



(e) Trial 1 - After training the step-length in addition to the radio model and the foreground detector

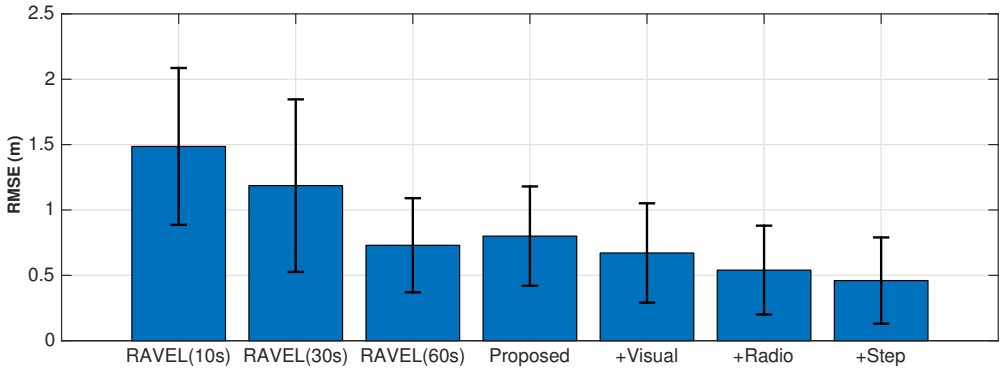


(f) Trial 2 - After training the step-length in addition to the radio model and the foreground detector

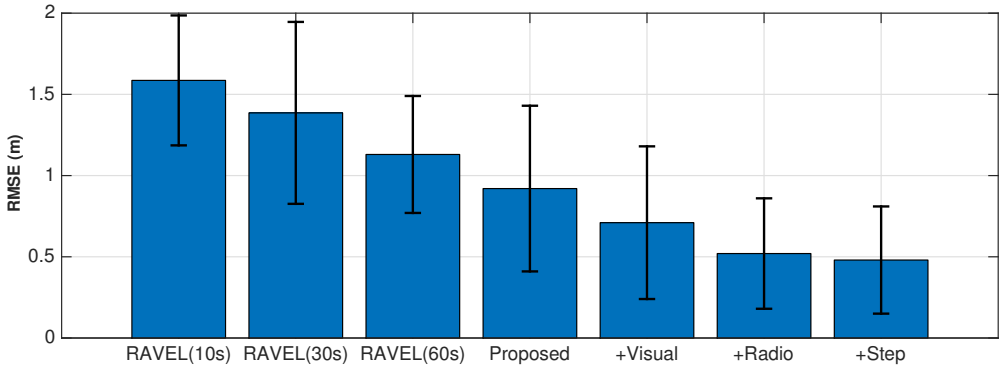
Figure 5.8: The graphs show the performance evaluation of the proposed system after applying cross-modality training on the construction site trials. (a, b) The foreground detection module is trained which makes object detection more accurate leading to better tracking accuracy. (c, d) In addition to the foreground detection the radio model is trained as well. This allows us to make better associations between id-linked and anonymous observations. The plots show a significant boost in accuracy. (e, f) The step-length estimation is trained, which improves the overall accuracy further.

Comparison with RAVEL:

Figure (5.9) shows how the final system architecture (Fig. (5.2)) with cross-modality stacks up against RAVEL. *Proposed* refers to the technique discussed in Chapter (4), +*Vision* is the proposed technique with cross-modality training enabled with the foreground detector trained, +*Radio* is the proposed technique where both the foreground detector and the radio model have been trained. Finally, +*Step* refers to the proposed technique when all sub-systems have been trained (i.e. foreground detector, radio model and step-length).



(a) Trial 1



(b) Trial 2

Figure 5.9: The graphs show the performance evaluation of the proposed system compared to RAVEL after applying cross-modality training on the construction site trials. *Proposed* denotes our approach where the foreground detector, the radio and step-length model are not trained. +*Vision* is our approach after the foreground detector has been trained, +*Radio* refers to our approach with the foreground detector and radio model trained and in +*Step* the step-length model is also trained. RAVEL(10s), RAVEL(30s) and RAVEL(60s) is the competing technique evaluated at window sizes of 10, 30 and 60 seconds respectively.

We should note here that RAVEL is also capable of learning the radio propagation model however it cannot optimise the visual detector and it does not make use of inertial observations which are the reasons for the lower accuracy. More specifically, in Fig. (5.9a) we can see that even though RAVEL(60) beats the proposed technique, once the foreground

detector is trained the accuracy improves beyond RAVEL. Furthermore, once the radio model and the step-length is trained the proposed technique is more accurate than RAVEL by approximately 35%.
Similar results are shown in Fig. (5.9b) as well for our second trial. In this case the proposed technique performs better than the off-line RAVEL, in real-time. Once all the sub-systems of the proposed technique have been trained the accuracy improvement over RAVEL(60) is as high as 56% with a mean RMSE error of just half meter. This increase in the accuracy over RAVEL is due to the following reasons:

- The proposed system utilises inertial measurements for motion prediction as opposed to RAVEL which does not. This allows the proposed technique to be more robust and accurate under long-term occlusions and maintain tracking without requiring visual observations. In addition, when people stay still (i.e. not walking) they become part of the background and so invisible to the foreground detector. This affects the accuracy of RAVEL because in those scenarios it goes "blind", i.e. we are trying to track something we cannot observe. The use of WiFi in RAVEL is used to mitigate this problem but this is not as accurate as the proposed technique which utilises inertial observations. Because we have acceleration data we can estimate the position of a target even in cases where the target becomes part of the background. We do not have the "blind" problem anymore.

- The proposed technique makes use of social forces to account for the interactions between people. This allows improved target dynamics which makes the predictive distribution of our filter more accurate. In addition, with social forces we can also model environmental constraints with occlusion maps. Social forces exerted from the environment onto people allow us to track targets more accurately and robustly in the environment. As the environment is populated with more constraints (i.e. walls, corridors) the gain from using social forces increases

- The proposed technique uses cross-modality training to optimise the parameters of its sub-system. This leads to a significant performance improvement over RAVEL while making the proposed technique suitable for highly dynamic environments.

5.10 Discussion

Existing positioning systems have been designed to operate within environments that exhibit long-term stable macro-structure with potential small-scale dynamics. They leverage this environmental stability to provide accurate location services. In contrast, in highly dynamic environments these assumptions are not valid. Currently, there is no system that can provide reliable and accurate tracking in such environments.

In this chapter we have proposed a novel positioning system to fill this gap. Specifically, we have presented our complete system augmented with cross-modality learning techniques which make this system suitable for highly dynamic environments. We showed that there is significant scope for automatically training the various sensor modalities, and this proved particularly useful in rapidly changing environments.

We have shown how the accuracy of our system improves gradually as the various components are trained using the cross-modality learning techniques presented in this chapter. We have observed in particular that the tracking error significantly decreases once both the foreground detector and the radio-model are learned. This is because our system requires both camera and radio measurements in order to determine the correct measurement to track association and update the target states. In the case of excessive missing camera detections, the trajectory of a target is estimated only by inertial measurements which is the main cause of the low accuracy. On the other hand, if the radio model was not trained, camera detections would not be able to be linked with radio measurements, which would also cause identification and tracking errors. In addition we have gained further improvements in the accuracy by learning a private step-length model for each person.

The learning techniques presented in this chapter make the proposed system adaptive and able to operate in challenging and dynamic environments reliably and accurately.

Finally, our experiments in a real construction site showed that after training, in our online approach the positioning error is decreased by a further 50%.

Chapter 6

Conclusion and Future Work

In this thesis we have studied the problem of practical positioning and tracking in challenging environments for which traditional positioning systems fail to provide a practical and viable solution. The need for practical positioning systems for various applications and environments, e.g. museums, shopping malls, construction sites, etc. strongly motivates this study. Next we have explained what makes this problem challenging i.e. the lack of existing positioning infrastructure, the limitations of the existing infrastructure and the environmental variability. We further described the key requirements which make a positioning system practical. More specifically, a positioning system that can be used in the application scenarios which we are targeting in this thesis should be low-cost, accurate, robust and scalable. To achieve these goals, we have proposed two positioning frameworks which are based on the fusion of id-linked and anonymous sensor streams. The proposed systems take advantage of the positioning accuracy of anonymous measurements (i.e. camera observations) and the identification accuracy of id-linked sensor streams (i.e. radio and inertial observations) in order to provide accurate positioning through the use of multiple-hypothesis tracking architectures. Our first system RAVEL exploits the opportunistic and ubiquitous radio signals (e.g. WiFi/Bluetooth Low Energy) in combination with CCTV measurements to provide sub-meter accuracy while being able to be applied in both receiver-centric (i.e. people carrying smartphones) and transmitter-centric (i.e. low-cost tags in warehouses, industrial or construction sites) applications. On the other hand our Rao-Blackwellized particle filtering system fuses radio, inertial and camera observations to provide accurate real-time tracking in highly dynamic industrial environments. In addition, we have shown how we can use social-forces and cross-modality training to address the challenges due to the environmental variability in industrial settings. Finally, we have conducted real-world experiments in two settings i.e. museum and construction site and we have showed through these real-world evaluations the performance of the proposed approaches. This chapter first

summaries the contributions of this thesis and the functionality of the proposed positioning frameworks and then discusses ideas for improvements and future work.

6.1 Summary of Contributions

This section concludes the contributions of this work and explains how the proposed frameworks meet all key requirements we have set in this thesis that make a positioning system practical and applicable for the application scenarios we have described in this study. In addition, we summarise the functionality and performance of the proposed systems. Concretely, the novel contributions of this thesis are as follows:

6.1.1 Formulation of the positioning problem

We have motivated the problem of multiple target positioning from various real-world scenarios, such as tracking the visitors in a museum to provide location-based information about the museum exhibits, improve navigation in airports, provide location-based advertising in shopping centers and improve productivity and safety in industrial settings. We have also described the key positioning challenges arising in these scenarios with illustrative examples. For instance, in a museum most security cameras are installed to provide a large field of view, typically resulting in a bird's eye view of the scene. This top-down perspective makes it difficult to distinguish facial features and accurate identification becomes even more challenging when the room is not well lit. In addition, the number of people changes over time and tracking becomes very challenging when people move behind obstacles, exit the field of view or cross paths. In other settings such as construction sites the environment changes over time, the workers wear similar uniforms, the sensor readings (e.g. WiFi) are noisy and the limited available infrastructure makes accurate tracking extremely difficult. We have explained the important concepts and assumptions and we have shown how different types of sensor modalities (i.e. id-linked and anonymous) can be used together to build accurate, cost-effective and robust positioning systems suitable for a wide range of applications. In essence we have demonstrated how key concepts of state estimation and data association can be used together to design positioning systems that operate in environments with limited sensing infrastructure and how we can use this concept to make a low-cost, high accurate positioning system from low-cost low-accuracy sensors.

6.1.2 Radio And Vision Enhanced Localisation (RAVEL)

We have proposed a novel positioning framework (in Chapter 3), which provides an accurate, low-cost and ubiquitous solution to the multiple target tracking problem, and which can be used in a wide range of application scenarios. RAVEL fuses anonymous visual detections captured by a stationary camera with WiFi readings to track multiple people moving inside an area with CCTV coverage. The WiFi measurements of each person are used to add context to the trajectories obtained by the camera in order to resolve visual ambiguities (e.g. split/merge paths) and increase the accuracy of visual tracking. The two main components of RAVEL are: a) a visual-based detector and b) a radio-aided tracker. These components can run in the same device (e.g. a server collecting measurements from multiple transmitters) or across different devices (e.g. smart cameras running the visual-based detector and disseminating traces to mobile devices, each running their own radio-aided tracker) according to the application requirements.

The visual-based detector is responsible for the task of tracklet generation whereas the radio-aided tracker is responsible for tracklet merging. RAVEL firstly uses the visual-based detector to collect visual detections over a period of time and to form unambiguous small trajectories (i.e. tracklets). Then these tracklets are used to create tracklet trees for each person (i.e. probable trajectory hypotheses). Finally, the WiFi measurements of each person are then used to search through the tracklet tree in order to find their most likely trajectory. The most likely trajectory is the one that agrees the most with the WiFi measurements. This deferred decision multiple hypothesis tracking approach allows us to use the noisy and inaccurate RSS-based WiFi measurements from people's smart-phones in combination with the anonymous camera detections which are provided by a lightweight object detector to create a highly accurate positioning system with sub-meter accuracy that meets the application requirements set throughout this thesis.

6.1.3 Rao-Blackellized Particle Filtering

We have proposed a novel multi-target multi-sensor tracking framework based on Rao-Blackwellized particle filtering (in Chapter 4), which is more suitable for on-line tracking. In certain scenarios the task of tracking people is very challenging. For instance in a construction site the workers are assigned different tasks and perform different activities which translate in different walking speeds, sudden movements, etc. All these challenges in combination with the noisy camera detections (e.g. missing detections, false detections, etc.) make accurate tracking far harder compared to more stable environments like museums and airports. In order to address the above problems we have designed an on-line tracking

framework (as opposed to RAVEL) which utilises three sensing modalities (i.e. camera, radio and inertial). Because inertial measurements are more informative and accurate they allow us to build a sequential Monte-Carlo online tracking system. The proposed system uses a foreground detector to detect the moving object in the field of view captured by a stationary camera (i.e. CCTV footage). A step detector, step-length estimator and a heading estimator are used to process the inertial measurements of each person and together with the radio measurements, are fed into a positioning and identification filter in order to provide the location estimates of all people in the scene. The Rao-Blackwellized particle filter is used to decompose the tracking problem into two parts a) estimation of the data-association distribution which is performed using sequential Monte-Carlo and b) estimation of the posterior distribution of target states which is computed analytically. This technique allows us to reduce the computational complexity of the tracking task and build an on-line system. Finally, the proposed framework is robust to noisy measurements, occlusions and intersecting paths. For instance, the state-of-the-art systems lose tracking of targets in cases where we have long-term occlusions (i.e. no camera observations). In our first system RAVEL we have addressed the problem of short-term occlusions using WiFi measurements. However, WiFi measurements are not informative enough to be used in cases where we lose visual tracking for a prolonged period of time. This is where inertial observations come into play, since they allow accurate tracking in cases with long term occlusions and make the proposed system significantly more accurate and robust compared to the existing techniques. However, compared to RAVEL the proposed system requires all measurements from all targets to be send to a central server for processing.

Additionally, in order to further improve the tracking accuracy we have incorporated the use of social forces into our tracking framework. The behaviour of human motion is affected by the motion of other people and also by obstacles from the environment. With the integration of social forces we can model the interactions between people and the influence of the environment (e.g. obstacles, walls, etc) on human motion. In order to do that we have introduced repulsive forces into our motion model which helps us to improve the motion prediction and achieve higher overall tracking accuracy.

6.1.4 Cross-modality Training

We have also proposed a novel learning approach which makes our Rao-Blackwellized particle filtering-based approach suitable for dynamic environments (in Chapter 5). When we have identified a target trajectory and linked a set of visual observations with a set of radio and inertial measurements with high probability, we then use this trajectory to learn the parameters of all the components in our system. In order to do that we have

defined a track quality metric which we use to assess the quality of the estimated trajectory i.e. how close is the estimated trajectory to the ground truth. Trajectories that qualify can then be used for cross-modality training. Since the qualifying tracks consist of a set of linked visual, radio and inertial measurements we use them to improve the step-length estimation for each target, train the foreground detector learning rate in order to increase the object detection rate and finally dynamically learn the radio model and different types of maps such as occlusion and magnetic. We have shown in Chapter 5, how the accuracy of our system improves gradually as we learn the different components of our system. Significant improvements are observed once the foreground detector and radio model are learned. In addition, we have shown that cross-modality training significantly boost the tracking accuracy and enables robust and reliable tracking in dynamic environments. Our experiments showed that the use of cross-modality training in the proposed system enables us to achieve sub-meter accuracy in challenging environments.

6.1.5 Real-world Evaluation

Testing and assessing the performance of a positioning system is a big challenge itself. We have talked about positioning and tracking of multiple targets/objects in highly dynamic environments and buildings with limited infrastructure which makes the performance evaluation even more challenging. Existing positioning techniques have been tested only in fairly simple and stable environments (e.g. offices, parks, etc) and do not capture real-world dynamics. In challenging environments maintaining tracking as people move behind obstacles, exit the field of view, or cross paths is an exceedingly difficult task. Industrial environments such as construction sites are characterised by rapid-large scale changes in structure and in combination with the changing lighting conditions and the many moving objects make the problem of tracking even more difficult. In this work we have conducted extensive real-world experiments in museums and construction sites in order to understand exactly the performance and the limitations of the proposed techniques. Only under these real world conditions we can truly understand the factors that affect the performance of our system and gain key insights on how to improve our approaches. More specifically, we have partnered with the Oxford Pitt Rivers museum and the Laing O'Rourke construction company in order to test the proposed system in real world conditions. With RAVEL we have shown that the idea of fusing id-linked and anonymous sensor streams helps us achieve sub-meter accuracy in a museum outperforming the state-of-the-art tracking systems. We have further showed how WiFi signals help improve tracking in the presence of occlusions, when people cross paths and enter/leave the field of view. For more dynamic environments such as construction sites which require real-time and accurate tracking we have showed that the

proposed Rao-Blackellized particle filter approach which utilises three sensing modalities also achieves sub-meter accuracy and outperforms RAVEL with the use of cross-modality training and the integration of social forces.

6.2 Future Work

In this work we have investigated in depth the problem of fusing together anonymous and id-linked measurements to provide a general framework for multi-target positioning. In addition, we have investigated learning techniques in order to make our systems adaptive in dynamic environments and have demonstrated their performance through real-world experiments. However, there are a number of limitations in the current systems that open interesting new directions for future work.

6.2.1 Complex human motion

A state-of-the-art tracking system must have the ability to monitor and predict the motion of the people being tracked. People normally avoid abrupt changes of their direction and speed and tend to have a linear motion pattern. Tracking systems take advantage of this fact and it is very common to use a constant velocity or constant acceleration motion model in order to model human motion. However, in certain situations human motion can become highly complex and unpredictable. For instance, in a construction site we have observed that there are cases where the motion of workers does not fit with traditional motion modelling techniques. For this reason we have incorporated inertial measurements (i.e. accelerometer and magnetometer) into our motion prediction process and we further improved motion prediction by modelling the interactions between people using social forces. Unfortunately, there are still cases where the proposed system can lose track of the targets. There are a number of things we can do in order to improve motion prediction. First, we can incorporate measurements from a gyroscope into our system instead of using only accelerometer and magnetometer and we can also use more sophisticated dead reckoning techniques [115]. This however, will increase the power consumption of the system (i.e. the smart-phone or sensor-tag which is used to sample the sensors) as the gyro sensors consume up to ten times more energy than the accelerometer and magnetometer. To avoid draining the battery we can use such techniques opportunistically and only when needed i.e. in difficult situations where the target is occluded and no camera measurements are available. As a second step we can incorporate deep learning techniques in our system for motion prediction. Instead of using motion models and modelling the interactions of people with the environment using social-forces we can use data-driven approaches like

deep-learning. The Deep Tracking system introduced in [116] demonstrates how recurrent neural networks (RNNs) can be used for object tracking without requiring any feature engineering or sensor models. It would be very interesting to see if we can capture human motion more accurately using deep learning techniques. However, this would require huge amount of ground-truth data which in certain situations (i.e. like in construction sites) is very hard to obtain.

6.2.2 A multi camera system

Another direction of future work is to incorporate additional non-overlapping cameras to our system. Currently we assume that our system takes input for a single camera and multiple copies of the proposed system can be deployed in different locations in order to provide tracking in a wider area. Since the targets that are being tracked can be identified using their device ID independently and still be able to track a unique target through different locations (we can merge the output of each system since we know the target ID). However, we can improve the tracking performance further if we have a central system that can take advantage of multiple non-overlapping cameras. For instance, since we know the target ID and we can track the target through different locations we can incorporate and learn transition probabilities between the different cameras of this system. We can then use probabilistic graphical models like hidden Markov models (HMMs) and conditional random fields (CRFs) to predict future target trajectories, better resolve motion ambiguities and improve the overall tracking accuracy. We need however to consider how such a system would scale, and what are the processing and cost requirements.

6.3 Closing Remarks

Recent advances in sensor technology and wireless communications along with increasingly faster and more power efficient mobile processors is changing the landscape of positioning services by creating a new post-GPS era. The next generation of positioning services will be ubiquitous, completely hidden from our lives and highly accurate. In shopping centers people would be able to know exactly where to find the item they are searching for, in shopping malls people would be able to save time by not getting lost, in museums they will be provided with location-triggered information about the exhibits and many businesses will improve their efficiency, lower their costs and provide safety to their workers. We believe that the positioning systems designed during this study will take us a step closer to the above vision by enabling a new-era of location-aware services.

Appendix A

Step-detection

Using acceleration data from a smart-phone to detect steps is a very important process of our system. In Chapter 4 the step detection is used by Eqn. (4.2) to model the target dynamics and predict the human motion. Here we will give an overview of the most popular step detection techniques and discuss in more detail our step detection approach.

A.1 Background

Over the years a wide variety of algorithms and techniques have been proposed to identify steps from accelerometer readings [117]. Popular step detection techniques include:

- Peak detection: Walking patterns (i.e. heel-strike events) are usually associated with sharp changes in the magnitude of the vertical acceleration. Peak detection algorithms are trained to recognise these events in order to identify and extract individual steps [118, 119, 120, 121].
- Zero-crossings: Another technique is to monitor the acceleration value for zero crossings [122, 123, 124].
- Auto/cross correlation: These techniques take advantage of the strong periodicity of the acceleration signal which occurs when someone is walking and apply auto/cross correlation on the acceleration signal in order to detect this cyclic pattern and identify/extract steps [125, 126, 127].
- Spectral analysis: These methods are the equivalent of the auto/cross correlation in the frequency domain. More specifically, the frequency spectrum of the acceleration signal is computed and then the dominant peak of the signal in the frequency domain is identified as the step frequency [128, 129, 130].

The above techniques however have a number of limitations which deems them impractical for certain applications. In a construction site for instance we have observed that the workers do not walk regularly, instead they often make big, small and irregular steps depending on the task performed. This behaviour often creates multiple peaks in the acceleration magnitude which greatly affects peak detection based methods. On the other hand we have observed that pure zero-crossing techniques are not satisfactory due to the false positive zero crossings from various events other than walking. Finally, the last two techniques (i.e. correlation-based and spectral analysis) depend on identifying periodicity in the acceleration signal. With these techniques however it is very difficult to reject false positives caused by any repetitive movement other than walking.

A.2 Proposed Technique

In order to mitigate some of the above problems we looked further into our data in order to acquire key insights and design a robust step detection algorithm. We should note here that in the proposed technique we have used 3-axis accelerometers sampled at 100 Hz which are commonly found in smart-phones. From our experiments we have identified that from all three axes the vertical acceleration contains the most informative data and so we use this axis to detect the step patterns. Our algorithm is inspired from the step detection techniques described above (i.e. peak detection, zero-crossing, etc) it is however extended in order to handle our real-world construction site data. More specifically the proposed technique includes four steps:

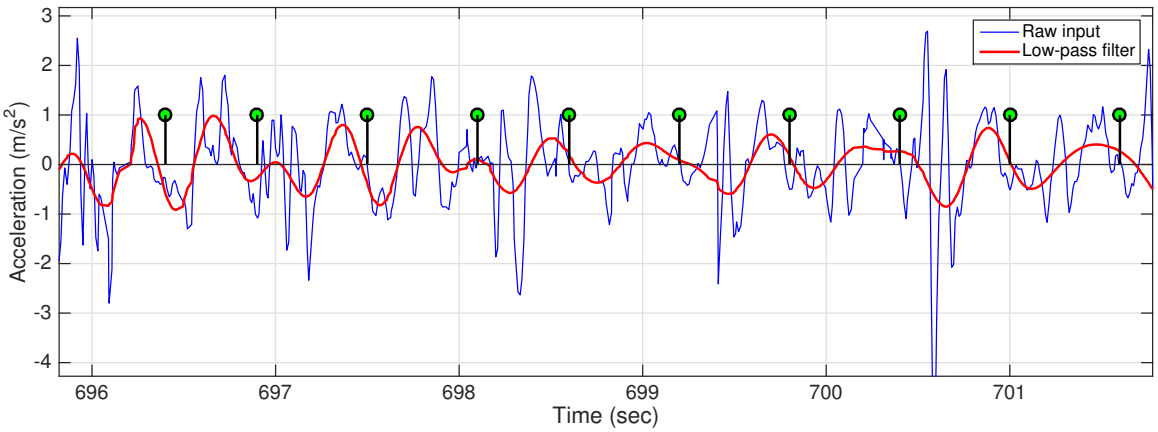


Figure A.1: The figure shows the raw vertical acceleration input of a user for a period of 10 steps (blue line) along with the filtered acceleration signal (red line). The green circles indicate the end of a step.

- Pre-processing: In this step we process the raw acceleration input to remove sensor noise and the influence of gravity on the measured acceleration. In order to eliminate the force of gravity from the measurements we use a high-pass filter which removes the DC component from the signal so that its offset is set to zero. Then we apply a low-pass Butterworth filter (8th order) to reduce the noise of the signal. This is shown in Fig. (A.1).

- Step segmentation: In this step we use the pre-processed data obtained from the previous step as input to a finite state machine which aims to identify and extract step events. This module aims to extract likely step events from the acceleration waveform and pass them to the next modules for further processing.

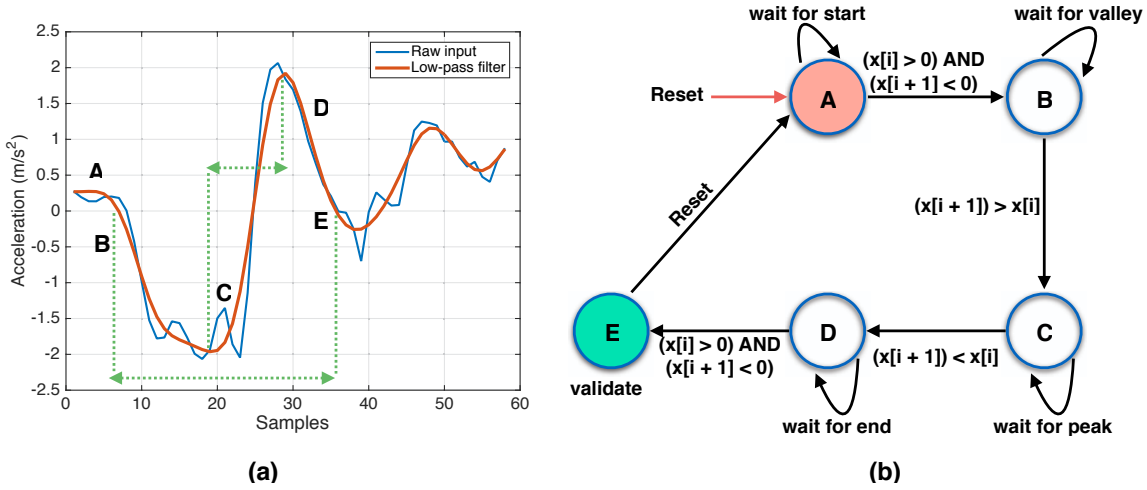
- Step validation: The steps extracted are then validated based on a set of rules which we are going to discuss in more detail next in this section. The objective here is to filter out impossible steps and reduce measurement noise.

- Step classification: Finally, we use the symbolic aggregate approximation technique (SAX) to capture and encode the shape characteristics of the step pattern and then we use a hidden Markov model to classify the SAX output as a step or not.

Now we are in a position to describe in more detail the segmentation, validation and classification steps of the proposed technique.

A.2.1 Step Segmentation

In our experiments the users had their phone in the front pocket of their trousers. From the accelerometer data obtained we have observed the following pattern. A step consists of two distinct phases namely a swing phase, where the foot is brought forward indicating the start of a step and a touch-down phase where the foot is placed on the ground which indicates the end of the step. We have studied what happens to the vertical acceleration during these two phases and we have found out the following pattern: (a) during the swing phase, the vertical acceleration exhibits a negative peak with an absolute acceleration magnitude larger than 0.2 m/s^2, (b) this negative peak is then followed by a positive peak with an acceleration magnitude larger than 0.2 m/s^2 which then reaches 0 m/s^2 as soon as the heel touches the ground and then again the acceleration moves towards a negative peak and so on and so forth resulting in a repetitive pattern. This cyclic behaviour is shown in Fig. (A.1) for 10 steps, where the green circles indicate the end of a step (i.e. heel touch-down phace).
The aim of the step segmentation module is to look at the acceleration data and identify the cyclic pattern described above. In order to do that we have designed a 5-state finite state



(a) (b)

Figure A.2: Step segmentation: We use a finite state machine to identify and extract the acceleration characteristics of a step. (a) The acceleration pattern of a step, (b) we use a 5-state FSM to identify the step pattern. The figure shows in (a) what acceleration patterns we are looking in each FSM state. Step validation: We use heuristics, such as the total duration of a step (indicated by the dotted green lines in (a)) to validate a step and filter out impossible steps.

machine (FSM), shown in Fig. (A.2), which takes as input the low-pass filtered acceleration data and looks for step patterns. First, in state A, our FSM looks for a zero crossing in the acceleration magnitude from positive to negative which indicates the possibility of the beginning of a step. When this happens we transition to state B where we wait for a negative minimum which indicates the swing phase described above. Once we have identified the negative peak we transition to state C waiting for a positive peak. Finally, we move to state D where we wait for a zero crossing which indicates the possibility of the end of a step. The FSM then transitions to state E where the step validation phase is performed.

A.2.2 Step Validation

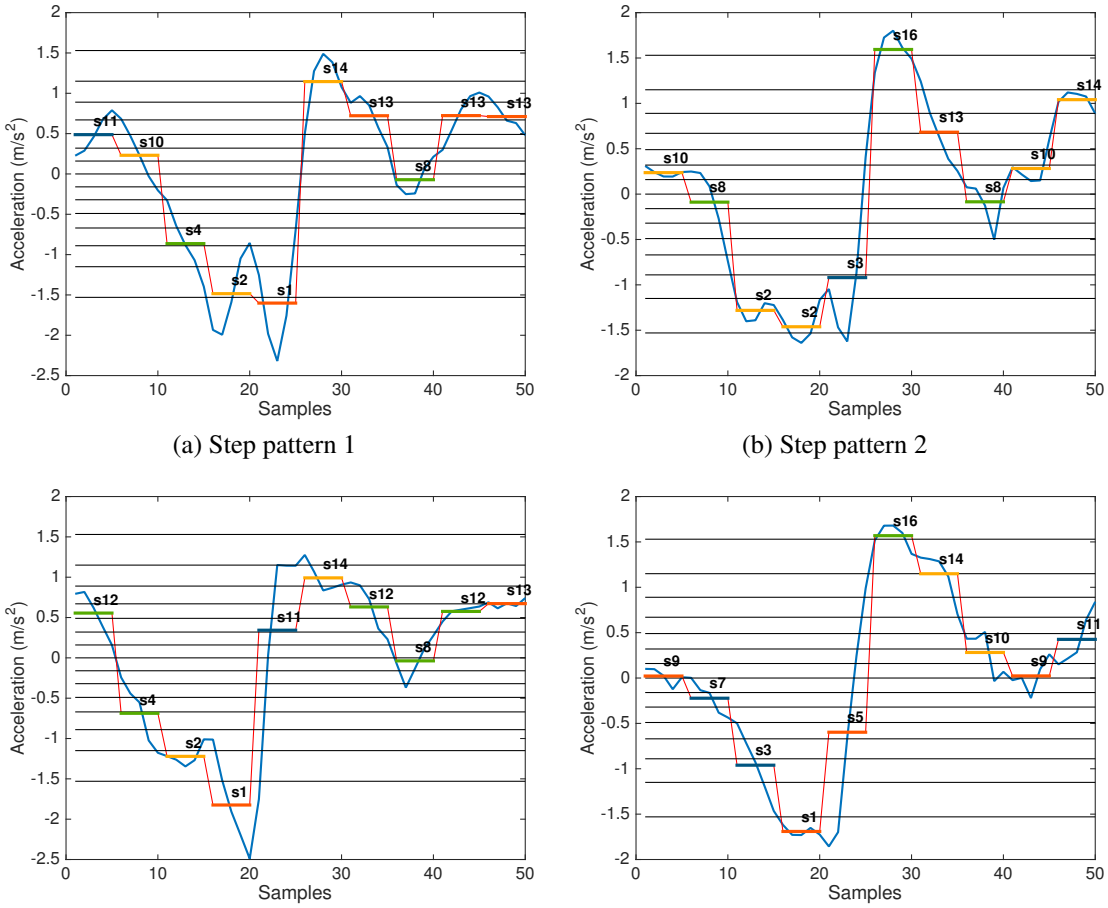
In this phase we conduct a sanity check of the extracted acceleration pattern before sending it to the step classifier. More specifically we expect a step to obey the following general guidelines:

- The total duration of a step should be between 400 and 800 milliseconds.
- The negative and positive peaks should be separated by an amount of time in the range of 200 and 400 milliseconds.
- The absolute magnitude of the negative and positive peak should exceed 0.2 m/s^2.

The above rules were obtained experimentally from our dataset. A step is sent for further processing if and only if all three rules are met. Otherwise, the extracted step pattern is discarded.

A.2.3 Step Classification

This is the core module of our proposed step detection technique. This module takes as input a valid acceleration step pattern and classifies it as a step or not. More specifically the operation of this module can be divided into two parts *encoding* and *classification*.



(a) Step pattern 1

(b) Step pattern 2

(c) Step pattern 3

(d) Step pattern 4

Figure A.3: The figure shows the application of the symbolic aggregate approximation (SAX) method to the extracted accelerometer step data. We use SAX to reduce the dimensionality of the data to symbolic vectors of length 10 using an alphabet of length 16. For instance, in (a) the shape of the step waveform is encoded into the vector $(s_{11}, s_{10}, s_4, s_2, s_1, s_{14}, s_{13}, s_8, s_{13}, s_{13})$. We use this symbolic representation to classify the acceleration pattern as step or not.

Encoding:

The aim of the encoding process is to find a high-level representation of the acceleration time-series step pattern which captures all the important step characteristics in a compact way so that it can be then used efficiently as an input to a classifier. Many high level representations of time series have been proposed, including Fourier transforms, wavelets, piecewise polynomial models, etc. Most of these techniques however suffer from two major problems: (a) the dimensionality of the symbolic representation is usually the same as the original data and (b) the distance measures on the symbolic data have little correlation with the distance measure on the original time series. One technique that overcomes these problems is the Symbolic Aggregate Approximation (SAX) [131] which allows for dimensionality reduction of the original time-series and also allows distance measures to be defined on the symbolic approach that lower bound corresponding distance measures defined on the original series. Finally, the symbolic representation (as opposed to real value representation) allows us to use simpler modelling techniques such as Markov models to define probabilities over discrete sequences.

In this work we use SAX to discretise, tokenise and reduce the dimensionality of the extracted acceleration step patterns and then with this representation train a step classifier. The details of SAX can be found in [131] and in this section we provide a brief overview of the method. SAX allows a time series of length n to be reduced to a string of length w (with $w < n$) using an alphabet s of arbitrary size. To transform the time series from n dimensions to w dimensions, the data is divided into w equal sized "frames". The mean value of the data falling within a frame is calculated and a vector of these values becomes the dimensionality reduced representation. This representation is also known as Piecewise Aggregate Approximation (PAA). The PAA coefficients are then mapped to symbols given predetermined equiprobable breakpoints. Figure (A.3) shows the application of SAX to the extracted acceleration pattern. We should note here that SAX is applied to the acceleration data before the application of the low-pass filtering. This allows us to capture better the shape characteristics of the waveform. In our system the extracted waveforms are reduced to symbolic vectors of length 10 with an alphabet of size 16 (i.e. the number of unique symbols). The aforementioned parameters were chosen experimentally and different configurations give similar results.

Classification:

Finally, we assume that the sequence of SAX symbols e.g. $S = (s_{11}, s_{10}, s_4, ...)$ are generated from an unobservable Markov chain and so we use a collection of these sequences

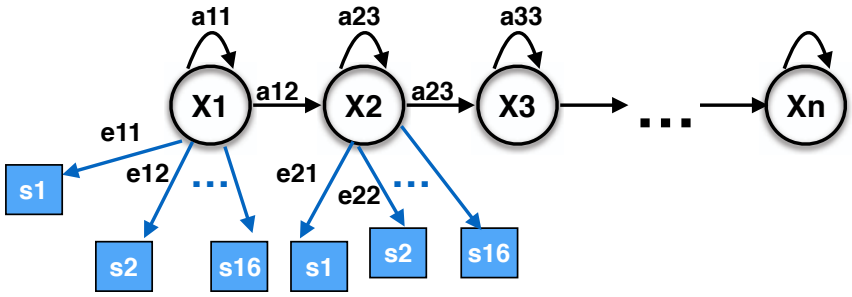


Figure A.4: We use a left-right hidden Markov model (HMM) to classify a given symbolic acceleration representation as being a step or not. We assume that the observations i.e. symbols (s_i) are emitted from unobservable states (x_j). We train the HMM, we find emission (e_{ij}) and transition (a_{ij}) probabilities given a collection of true steps and then we use the trained HMM to find the likelihood of a new sequence of observations given the trained model.

which represent positive steps to train a hidden Markov model [84, 132, 133]. More specifically, we use a left-right HMM with 8 hidden states (X_j) as shown in Fig. (A.4) to calculate the likelihood of an observation sequence (S) given the trained model (Θ) i.e. $p(S|\Theta)$ which is the sum of joint likelihoods of the observation sequence over all possible state sequences allowed by the model.

$$p(S|\Theta) = \sum_{X} p(S|X,\Theta)\, p(X|\Theta) \tag{A.1}$$

where $p(S|X,\Theta)$ is the likelihood of an observation sequence given a state sequence and the model and $p(X|\Theta)$ is the probability of a state sequence given the model. Once we calculate $p(S|\Theta)$ we can then make a final hard decision whether to accept the given sequence S as a step or not.

Figure (A.5) shows the overall step detection accuracy of the proposed technique on a sample of 900 steps from 4 people. In this figure, the first two diagonal cells show the number and percentage of correct classifications by the proposed technique. For example, 567 steps were correctly classified as true steps. This corresponds to 63.0% of all 900 acceleration patterns. Similarly, 264 cases were correctly classified as not steps. This corresponds to 29.3% of all samples. 27 of the acceleration patterns that are not steps were incorrectly classified as steps and this corresponds to 3.0% of all 900 samples. Similarly, 42 actual steps were incorrectly classified as not steps and this corresponds to 4.7% of all data. Out of 594 step predictions, 95.5% are correct and 4.5% are wrong. Out of 306 not-step predictions, 86.3% are correct and 13.7% are wrong. Out of 609 true step cases, 93.1% are correctly predicted as steps and 6.9% are predicted as not steps. Finally, out of 291

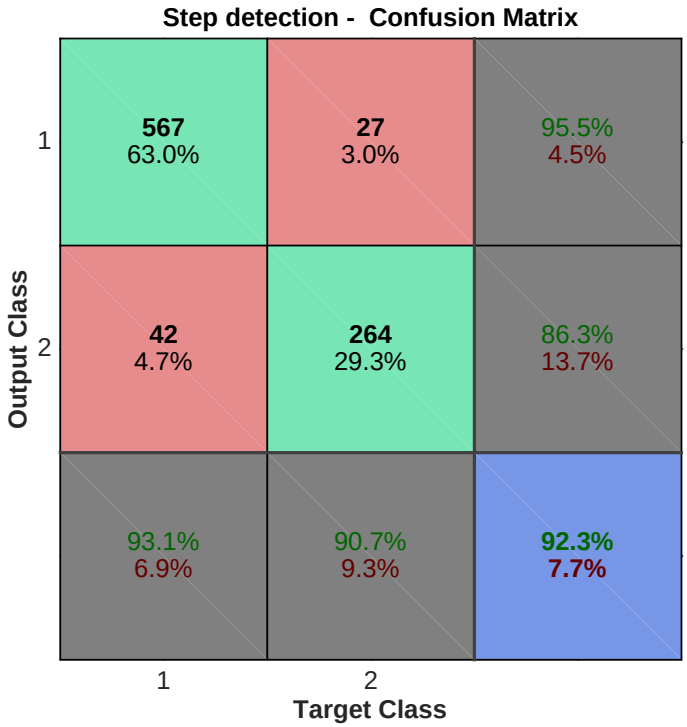


Figure A.5: Step detection accuracy of the proposed technique over 900 steps from 4 people.

not step cases, 90.7% are correctly classified as not steps and 9.3% are classified as steps. Overall, 92.3% of the predictions are correct and 7.7% are wrong classifications. We should note here that the proposed technique significantly outperforms traditional step-detection approaches (i.e. peak detection, zero crossing and correlation based). Table (A.1) shows the step detection accuracy of various competing techniques for the dataset described above and clearly motivates our decision to investigate more closely the step detection problem.

Step Detection Accuray	
Approach	**Accuracy**
Peak Detection [118]	81.2%
Zero Crossing [124]	73.0%
Auto/cross correlation [127]	84.4%
Proposed	92.3%

Table A.1: Comparison of existing step-detection approaches in terms of step detection accuracy.

Bibliography

[1] J. B.-Y. Tsui, *Fundamentals of Global Positioning System Receivers: A Software Approach*. Wiley Series in Microwave and Optical Engineering, 2005.

[2] G. Fischer, O. Klymenko, D. Martynenko, and H. Luediger, "An impulse radio uwb transceiver with high-precision toa measurement unit," in *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, pp. 1–8, Sept 2010.

[3] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The cricket location-support system," in *Proceedings of the 6th annual international conference on Mobile computing and networking*, MobiCom '00, (New York, NY, USA), pp. 32–43, ACM, 2000.

[4] R. Want, A. Hopper, V. Falcão, and J. Gibbons, "The active badge location system," *ACM Trans. Inf. Syst.*, vol. 10, pp. 91–102, Jan. 1992.

[5] "Leantegra inc." https://leantegra.com. Accessed: 2017-06-12.

[6] N. R. Council, *Advancing the Competitiveness and Efficiency of the U.S. Construction Industry*. Washington, DC: The National Academies Press, 2009.

[7] M. Pietrzyk and T. von der Grun, "Experimental validation of a toa uwb ranging platform with the energy detection receiver," in *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, pp. 1–8, Sept 2010.

[8] M. Segura, H. Hashemi, C. Sisterna, and V. Mut, "Experimental demonstration of self-localized ultra wideband indoor mobile robot navigation system," in *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, pp. 1–9, Sept 2010.

[9] H. Kröll and C. Steiner, "Indoor ultra-wideband location fingerprinting," in *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, pp. 1–5, Sept 2010.

[10] A. Ward, A. Jones, and A. Hopper, "A new location technique for the active office," *Personal Communications, IEEE*, vol. 4, no. 5, pp. 42–47, 1997.

[11] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster, "The anatomy of a context-aware application," in *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, MobiCom '99, (New York, NY, USA), pp. 59–68, ACM, 1999.

[12] H. shik Kim and J.-S. Choi, "Advanced indoor localization using ultrasonic sensor and digital compass," in *Control, Automation and Systems, 2008. ICCAS 2008. International Conference on*, pp. 223–226, 2008.

[13] H. Schweinzer and M. Syafrudin, "Losnus: An ultrasonic system enabling high accuracy and secure tdoa locating of numerous devices," in *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, pp. 1–8, Sept 2010.

[14] L. Ni, Y. Liu, Y. C. Lau, and A. Patil, "Landmarc: indoor location sensing using active rfid," in *Pervasive Computing and Communications, 2003. (PerCom 2003). Proceedings of the First IEEE International Conference on*, pp. 407–415, 2003.

[15] K. Chawla, G. Robins, and L. Zhang, "Object localization using rfid," in *Wireless Pervasive Computing (ISWPC), 2010 5th IEEE International Symposium on*, pp. 301–306, 2010.

[16] T. Sanpechuda and L. Kovavisaruch, "A review of rfid localization: Applications and techniques," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on*, vol. 2, pp. 769–772, 2008.

[17] F. Seco, C. Plagemann, A. Jimenez, and W. Burgard, "Improving rfid-based indoor positioning accuracy using gaussian processes," in *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, pp. 1–8, Sept 2010.

[18] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *INFOCOM*, pp. 775–784, 2000.

[19] P. Bahl and V. N. Padmanabhan, *Enhancements to the RADAR User Location and Tracking System*. TechReport MSR-TR-2000-12, Microsoft Research, 2000.

[20] B. Ferris, D. Haehnel, and D. Fox, "Gaussian processes for signal strength-based location estimation," in *in Proc. of Robotics Science and Systems, University of Pennsylvania*, 2006.

[21] A. Haeberlen, E. Flannery, A. M. Ladd, A. Rudys, D. S. Wallach, and L. E. Kavraki, "Practical robust localization over large-scale 802.11 wireless networks," in *Proceedings of the 10th annual international conference on Mobile computing and networking*, MobiCom '04, (New York, NY, USA), pp. 70–84, ACM, 2004.

[22] M. Youssef and A. Agrawala, "The horus wlan location determination system," in *Proceedings of the 3rd international conference on Mobile systems, applications, and services*, MobiSys '05, (New York, NY, USA), pp. 205–218, ACM, 2005.

[23] C. Figuera, I. Mora-Jimenez, A. Guerrero-Curieses, J. Rojo-Alvarez, E. Everss, M. Wilby, and J. Ramos-Lopez, "Nonparametric model comparison and uncertainty evaluation for signal strength indoor location," *Mobile Computing, IEEE Transactions on*, vol. 8, no. 9, pp. 1250–1264, 2009.

[24] G. Wassi, C. Despins, D. Grenier, and C. Nerguizian, "Indoor location using received signal strength of ieee 802.11b access point," in *Electrical and Computer Engineering, 2005. Canadian Conference on*, pp. 1367–1370, 2005.

[25] B. Li, J. Salter, A. G. Dempster, and C. Rizos, "Indoor positioning techniques based on wireless lan," in *First IEEE International Conference on Wireless Broadband and Ultra Wideband Communications*, pp. 13–16, 2006.

[26] O. Woodman and R. Harle, "Pedestrian localisation for indoor environments," in *Proceedings of the 10th international conference on Ubiquitous computing*, UbiComp '08, (New York, NY, USA), pp. 114–123, ACM, 2008.

[27] M. Susi, V. Renaudin, and G. Lachapelle, "Motion mode recognition and step detection algorithms for mobile phone users," *Sensors*, vol. 13, no. 2, pp. 1539–1562, 2013.

[28] A. Jimenez, F. Seco, C. Prieto, and J. Guevara, "A comparison of pedestrian dead-reckoning algorithms using a low-cost mems imu," in *Intelligent Signal Processing, 2009. WISP 2009. IEEE International Symposium on*, pp. 37–42, 2009.

[29] E. Foxlin, "Pedestrian tracking with shoe-mounted inertial sensors," *Computer Graphics and Applications, IEEE*, vol. 25, no. 6, pp. 38–46, 2005.

[30] F. Li, C. Zhao, G. Ding, J. Gong, C. Liu, and F. Zhao, "A reliable and accurate indoor localization method using phone inertial sensors," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, (New York, NY, USA), pp. 421–430, ACM, 2012.

[31] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easyliving," in *Visual Surveillance, 2000. Proceedings. Third IEEE International Workshop on*, pp. 3–10, 2000.

[32] I. Haritaoglu, D. Harwood, and L. Davis, "Hydra: multiple people detection and tracking using silhouettes," in *Image Analysis and Processing, 1999. Proceedings. International Conference on*, pp. 280–285, 1999.

[33] T. Darrell, G. Gordon, M. Harville, and J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pp. 601–608, 1998.

[34] J. Orwell, S. Massey, P. Remagnino, D. Greenhill, and G. Jones, "A multi-agent framework for visual surveillance," in *Image Analysis and Processing, 1999. Proceedings. International Conference on*, pp. 1104–1107, 1999.

[35] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1815–1821, 2012.

[36] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, 2008.

[37] A. R. Zamir, A. Dehghan, and M. Shah, "Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs," in *Proceedings of the 12th European Conference on Computer Vision - Volume Part II*, ECCV'12, (Berlin, Heidelberg), pp. 343–356, Springer-Verlag, 2012.

[38] Y. Chen, D. Lymberopoulos, J. Liu, and B. Priyantha, "Fm-based indoor localization," in *Proceedings of the 10th international conference on Mobile systems, applications, and services*, MobiSys '12, (New York, NY, USA), pp. 169–182, ACM, 2012.

[39] R. Wang, F. Zhao, H. Luo, B. Lu, and T. Lu, "Fusion of wi-fi and bluetooth for indoor localization," in *Proceedings of the 1st International Workshop on Mobile Location-based Service*, MLBS '11, (New York, NY, USA), pp. 63–66, ACM, 2011.

[40] C. E. G. T, I. Galvan-Tejada, E. I. Sandoval, and R. Brena, "Wifi bluetooth based combined positioning algorithm," *Procedia Engineering*, vol. 35, no. 0, pp. 101 – 108, 2012. International Meeting of Electrical Engineering Research 2012.

[41] L. Bruno and P. Robertson, "Wislam: Improving footslam with wifi," in *Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on*, pp. 1–10, 2011.

[42] Z. Yang, C. Wu, and Y. Liu, "Locating in fingerprint space: wireless indoor localization with little human intervention," in *Proceedings of the 18th annual international conference on Mobile computing and networking*, Mobicom '12, (New York, NY, USA), pp. 269–280, ACM, 2012.

[43] B. Zhang, J. Teng, J. Zhu, X. Li, D. Xuan, and Y. F. Zheng, "Ev-loc: integrating electronic and visual signals for accurate localization," in *Proceedings of the thirteenth ACM international symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '12, (New York, NY, USA), pp. 25–34, ACM, 2012.

[44] T. Teixeira, D. Jung, G. Dublon, and A. Savvides, "Identifying people in camera networks using wearable accelerometers," in *Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '09, (New York, NY, USA), pp. 20:1–20:8, ACM, 2009.

[45] T. Teixeira, D. Jung, and A. Savvides, "Tasking networked cctv cameras and mobile phones to identify and localize multiple people," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, Ubicomp '10, (New York, NY, USA), pp. 213–222, ACM, 2010.

[46] J. Hightower and G. Borriello, "Location systems for ubiquitous computing," *Computer*, vol. 34, no. 8, pp. 57–66, 2001.

[47] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 6, pp. 1067–1080, 2007.

[48] Y. Gu, A. Lo, and I. Niemegeers, "A survey of indoor positioning systems for wireless personal networks," *Communications Surveys Tutorials, IEEE*, vol. 11, no. 1, pp. 13–32, 2009.

[49] M. R., *Indoor positioning technologies*. Habilitation thesis, ETH Zurich, 2012.

[50] Y. Chen and H. Kobayashi, "Signal strength based indoor geolocation," in *Communications, 2002. ICC 2002. IEEE International Conference on*, vol. 1, pp. 436–439, 2002.

[51] R. Mautz and S. Tilch, "Survey of optical indoor positioning systems," in *2011 International Conference on Indoor Positioning and Indoor Navigation*, pp. 1–7, Sept 2011.

[52] R. Mandeljc, J. Pers, M. Kristan, and S. Kovacic, "Fusion of non-visual modalities into the probabilistic occupancy map framework for person localization," in *Distributed Smart Cameras (ICDSC), 2011 Fifth ACM/IEEE International Conference on*, pp. 1–6, Aug 2011.

[53] D. Simon, *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.

[54] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. D, pp. 35–45, 1960.

[55] Y. Bar-Shalom, *Tracking and Data Association*. San Diego, CA, USA: Academic Press Professional, Inc., 1987.

[56] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Artech House radar library, Artech House, 1999.

[57] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," in *AeroSense'97*, pp. 182–193, International Society for Optics and Photonics, 1997.

[58] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[59] A. Doucet, N. De Freitas, and N. Gordon, *Sequential monte carlo methods in practice*. Springer-Verlag, 2001.

[60] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer Publishing Company, Incorporated, 2008.

[61] A. Doucet, N. d. Freitas, K. P. Murphy, and S. J. Russell, "Rao-blackwellised particle filtering for dynamic bayesian networks," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, UAI '00, (San Francisco, CA, USA), pp. 176–183, Morgan Kaufmann Publishers Inc., 2000.

[62] Y.-K. Ku, C.-S. Nam, and D.-R. Shin, "Efficient indoor localization and error correction algorithm," in *Advanced Communication Technology (ICACT), 2010 The 12th International Conference on*, vol. 1, pp. 453–457, 2010.

[63] W. Honcharenko, H. Bertoni, and J. Dailing, "Mechanisms governing propagation between different floors in buildings," *Antennas and Propagation, IEEE Transactions on*, vol. 41, no. 6, pp. 787–790, 1993.

[64] K.-W. Cheung, J.-M. Sau, and R. Murch, "A new empirical model for indoor propagation prediction," *Vehicular Technology, IEEE Transactions on*, vol. 47, no. 3, pp. 996–1001, 1998.

[65] S. Seidel and T. Rappaport, "914 mhz path loss prediction models for indoor wireless communications in multifloored buildings," *IEEE Tran. on Antennas and Propagation*, vol. 40, pp. 207–217, Feb 1992.

[66] S. L. Cebula, A. Ahmad, J. M. Graham, C. Hinds, L. A. Wahsheh, A. T. Williams, and S. J. DeLoatch, *Empirical Channel Model for 2.4GHz IEEE 802.11 WLAN*. Proceedings of the 2011 International Conference on Wireless Networks, 2011.

[67] S. Mazuelas, A. Bahillo, R. Lorenzo, P. Fernandez, F. Lago, E. Garcia, J. Blas, and E. Abril, "Robust indoor positioning provided by real-time rssi values in unmodified wlan networks," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, pp. 821–831, Oct 2009.

[68] V. Honkavirta, T. Perala, S. Ali-Loytty, and R. Piche, "A comparative survey of wlan location fingerprinting methods," in *Positioning, Navigation and Communication, 2009. WPNC 2009. 6th Workshop on*, pp. 243–251, 2009.

[69] B. Ferris, D. Fox, and N. Lawrence, "Wifi-slam using gaussian process latent variable models," in *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, (San Francisco, CA, USA), pp. 2480–2485, Morgan Kaufmann Publishers Inc., 2007.

[70] J. Seitz, T. Vaupel, S. Meyer, J. Gutiérrez Boronat, and J. Thielecke, "A hidden markov model for pedestrian navigation," in *Positioning Navigation and Communication (WPNC), 2010 7th Workshop on*, pp. 120–127, March 2010.

[71] V. Moghtadaiee, A. Dempster, and S. Lim, "Indoor localization using fm radio signals: A fingerprinting approach," in *Indoor Positioning and Indoor Navigation (IPIN), 2011 International Conference on*, pp. 1–7, 2011.

[72] J. L. Crowley, P. Stelmaszyk, and C. Discours, "Measuring image flow by tracking edge-lines," in *[1988 Proceedings] Second International Conference on Computer Vision*, pp. 658–664, Dec 1988.

[73] R. Deriche and O. Faugeras, "Tracking line segments," *Image Vision Comput.*, vol. 8, pp. 261–270, Nov. 1990.

[74] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society of Industrial and Applied Mathematics*, vol. 5, pp. 32–38, March 1957.

[75] H. Kuhn, "The hungarian method for the assignment problem," in *50 Years of Integer Programming 1958-2008* (M. Jünger, T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey, eds.), pp. 29–47, Springer Berlin Heidelberg, 2010.

[76] F. Bourgeois and J.-C. Lassalle, "An extension of the munkres algorithm for the assignment problem to rectangular matrices," *Commun. ACM*, vol. 14, pp. 802–804, Dec. 1971.

[77] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems*, vol. 29, pp. 82–100, Dec 2009.

[78] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *Oceanic Engineering, IEEE Journal of*, vol. 8, no. 3, pp. 173–184, 1983.

[79] D. Reid, "An algorithm for tracking multiple targets," *Automatic Control, IEEE Transactions on*, vol. 24, no. 6, pp. 843–854, 1979.

[80] S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *Aerospace and Electronic Systems Magazine, IEEE*, vol. 19, pp. 5–18, Jan 2004.

[81] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Comput. Surv.*, vol. 38, Dec. 2006.

[82] E. Trucco and K. Plakas, "Video tracking: A concise survey," *Oceanic Engineering, IEEE Journal of*, vol. 31, no. 2, pp. 520–529, 2006.

[83] D. Beymer and K. Konolige, "Real-time tracking of multiple people using continuous detection.," in *IIEEE International Conference on Computer Vision (ICCV)*, 1999.

[84] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, pp. 4–16, Jan 1986.

[85] Z. Khan, T. Balch, and F. Dellaert, "Mcmc-based particle filtering for tracking a variable number of interacting targets," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, 2005.

[86] A. G. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, "Multi-object tracking through simultaneous long occlusions and split-merge conditions," *Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 666–673, 01 2006.

[87] R. Collins, "Multitarget data association with higher-order motion models," *Conference on Computer Vision and Pattern Recognition*, 2012.

[88] D. Lee, I. Hwang, and S. Oh, "Optimus:online persistent tracking and identification of many users for smart spaces," *Machine Vision and Applications*, vol. 25, no. 4, pp. 901–917, 2014.

[89] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, pp. 267–282, Feb 2008.

[90] R. Mandeljc, S. Kovačič, M. Kristan, and J. Perš, "Tracking by identification using computer vision and radio," *Sensors*, vol. 13, no. 1, pp. 241–273, 2012.

[91] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE CS Conference on Computer Vision and Pattern Recognition*, CVPR '99, 1999.

[92] Y. Shen, W. Hu, J. Liu, M. Yang, B. Wei, and C. T. Chou, "Efficient background subtraction for real-time tracking in embedded camera networks," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, SenSys '12, (New York, NY, USA), pp. 295–308, ACM, 2012.

[93] R. Chandra, J. Padhye, L. Ravindranath, and A. Wolman, "Beacon-stuffing: Wi-fi without associations," in *Mobile Computing Systems and Applications, 2007. HotMobile 2007. Eighth IEEE Workshop on*, pp. 53–57, March 2007.

[94] I. Sethi and R. Jain, "Finding trajectories of feature points in a monocular image sequence," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-9, pp. 56–73, Jan 1987.

[95] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[96] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "On the levy-walk nature of human mobility," in *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications*, April 2008.

[97] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Communications and Mobile Computing*, vol. 2, no. 5, pp. 483–502, 2002.

[98] J. Ning, L. Zhang, D. Zhang, and C. Wu, "Robust mean-shift tracking with corrected background-weighted histogram," *Computer Vision, IET*, vol. 6, pp. 62–69, January 2012.

[99] I. Cox, "A review of statistical data association techniques for motion correspondence," *Int'l J. Computer Vision*, vol. 10, no. 1, 1993.

[100] P. Kaewtrakulpong and R. Bowden, "An improved adaptive background mixture model for realtime tracking with shadow detection," in *An improved adaptive background mixture model for realtime tracking with shadow detection*, 2001.

[101] T. S. Rappaport and S. Sandhu, "Radio-wave propagation for emerging wireless personal-communication systems," *IEEE Antennas and Propagation Magazine*, vol. 36, pp. 14–24, Oct 1994.

[102] N. Gordon, "A hybrid bootstrap filter for target tracking in clutter," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 33, pp. 353–358, Jan 1997.

[103] S. Oh, S. Russell, and S. Sastry, "Markov chain monte carlo data association for multi-target tracking," *IEEE Transactions on Automatic Control*, vol. 54, pp. 481–497, March 2009.

[104] C. Hue, J.-P. Le Cadre, and P. Perez, "Tracking multiple objects with particle filtering," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 38, pp. 791–812, Jul 2002.

[105] C. Hue, J. pierre Le Cadre, and P. Perez, "The (mr)mtpf: particle filters to track multiple targets using multiple receivers," in *In 4th International Conference on Information Fusion*, pp. 3–33, 2001.

[106] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized monte carlo data association for multiple target tracking," in *Proceedings of the seventh international conference on information fusion*, vol. 1, pp. 583–590, I, 2004.

[107] S. Särkkä, A. Vehtari, and J. Lampinen, "Rao-blackwellized particle filter for multiple target tracking," *Information Fusion*, vol. 8, no. 1, pp. 2 – 15, 2007. Special Issue on the Seventh International Conference on Information Fusion-Part {II} Seventh International Conference on Information Fusion.

[108] S. Papaioannou, H. Wen, A. Markham, and N. Trigoni, "Fusion of radio and camera sensor data for accurate indoor positioning," in *Mobile Ad Hoc and Sensor Systems (MASS), 2014 IEEE 11th International Conference on*, pp. 109–117, Oct 2014.

[109] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second ed., 2004.

[110] M. Luber, J. Stork, G. Tipaldi, and K. Arras, "People tracking with human motion predictions from social forces," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 464–469, May 2010.

[111] T. Bouwmans, F. El Baf, and B. Vachon, "Background modeling using mixture of gaussians for foreground detection-a survey," *Recent Patents on Computer Science*, vol. 1, no. 3, pp. 219–237, 2008.

[112] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding*, vol. 122, no. 0, pp. 4 – 21, 2014.

[113] M. Shah, J. D. Deng, and B. J. Woodford, "Video background modeling: recent approaches, issues and our proposed techniques," *Machine Vision and Applications*, vol. 25, no. 5, pp. 1105–1119, 2014.

[114] V. Renaudin, M. Susi, and G. Lachapelle, "Step length estimation using handheld inertial sensors," *Sensors*, vol. 12, no. 7, pp. 8507–8525, 2012.

[115] Z. Xiao, H. Wen, A. Markham, and N. Trigoni, "Robust pedestrian dead reckoning (r-pdr) for arbitrary mobile device placement," in *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 187–196, Oct 2014.

[116] P. Ondruska and I. Posner, "Deep tracking: Seeing beyond seeing using recurrent neural networks," in *The Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, (Phoenix, Arizona USA), February 2016.

[117] R. Harle, "A survey of indoor inertial positioning systems for pedestrians," *IEEE Communications Surveys Tutorials*, vol. 15, pp. 1281–1293, Third 2013.

[118] kim, Jeong W., "A Step, Stride and Heading Determination for the Pedestrian Navigation System," in *Presented at GNSS 2004, Sydney, Australia*, Dec. 2004.

[119] L. Fang, P. J. Antsaklis, L. A. Montestruque, M. B. McMickell, M. Lemmon, Y. Sun, H. Fang, I. Koutroulis, M. Haenggi, M. Xie, and X. Xie, "Design of a wireless assisted pedestrian dead reckoning system - the navmote experience," *IEEE Transactions on Instrumentation and Measurement*, vol. 54, pp. 2342–2358, Dec 2005.

[120] I. Bylemans, M. Weyn, and M. Klepal, "Mobile phone-based displacement estimation for opportunistic localisation systems," in *2009 Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, pp. 113–118, Oct 2009.

[121] C. Randell, C. Djiallis, and H. Muller, "Personal position measurement using dead reckoning," in *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, ISWC '03, (Washington, DC, USA), pp. 166–, IEEE Computer Society, 2003.

[122] S. H. Shin, M. S. Lee, C. G. Park, and H. S. Hong, "Pedestrian dead reckoning system with phone location awareness algorithm," in *IEEE/ION Position, Location and Navigation Symposium*, pp. 97–101, May 2010.

[123] P. Goyal, V. J. Ribeiro, H. Saran, and A. Kumar, "Strap-down pedestrian dead-reckoning system," in *2011 International Conference on Indoor Positioning and Indoor Navigation*, pp. 1–7, Sept 2011.

[124] S. Beauregard, "A helmet-mounted pedestrian dead reckoning system," in *3rd International Forum on Applied Wearable Computing 2006*, pp. 1–11, March 2006.

[125] Y. Makihara, N. T. Trung, H. Nagahara, R. Sagawa, Y. Mukaigawa, and Y. Yagi, *Phase Registration of a Single Quasi-Periodic Signal Using Self Dynamic Time Warping*, pp. 667–678. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[126] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: Zero-effort crowdsourcing for indoor localization," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, Mobicom '12, (New York, NY, USA), pp. 293–304, ACM, 2012.

[127] H. Ying, C. Silex, A. Schnitzer, S. Leonhardt, and M. Schiek, *Automatic Step Detection in the Accelerometer Signal*, pp. 80–85. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[128] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3*, IAAI'05, pp. 1541–1546, AAAI Press, 2005.

[129] J.-g. Park, A. Patel, D. Curtis, S. Teller, and J. Ledlie, "Online pose classification and walking speed estimation using handheld devices," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, (New York, NY, USA), pp. 113–122, ACM, 2012.

[130] A. Brajdic and R. Harle, "Walk detection and step counting on unconstrained smartphones," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, (New York, NY, USA), pp. 225–234, ACM, 2013.

[131] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, DMKD '03, (New York, NY, USA), pp. 2–11, ACM, 2003.

[132] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 379–385, Jun 1992.

[133] H. A. Bourlard and N. Morgan, *Hidden Markov Models*, pp. 27–58. Boston, MA: Springer US, 1994.