Human Perception Capabilities for Socially Intelligent Domestic Service Robots

Alex Noyvirt

School of Engineering

Cardiff University

January 2016

The daily living activities for an increasing number of frail elderly people represent a continuous struggle both for them as well as for their extended families. These people have difficulties coping at home alone but are still sufficiently fit not to need the round-the-clock care provided by a nursing home. Their struggle can be alleviated by the deployment of a mechanical helper in their home, i.e. a service robot that can execute a range of simple object manipulation tasks. Such a robotic application promises to extend the period of independent home living for elderly people, while providing them with a better quality of life. However, despite the recent technological advances in robotics, there are still some remaining challenges, mainly related to the human factors. Arguably, the lack of consistently dependable human detection, localisation, position and pose tracking information and insufficiently refined processing of sensor information makes the close range physical interaction between a robot and a human a high-risk task.

The work described in this thesis addresses the deficiencies in the processing of the human information of today's service robots. This is achieved through proposing a new paradigm for the robot's situational awareness in regard to people as well as a collection of methods and techniques, operating at the lower levels of the paradigm, i.e. perception of new human information. The collection includes methods for obtaining and processing of information about the presence, location and body pose of the people. In addition to the availability of reliable human perception information, the integration between the separate levels of paradigm is considered to be a critically important factor for achieving the human-aware control of the robot. Improving the cognition, judgment and decision making action links between the paradigm's layers leads to enhanced capability of the robot to engage in a natural and more meaningful interaction with people and, therefore, to a more enjoyable user experience. Therefore, the proposed paradigm and methodology are envisioned to contribute to making the prolonged assisted living of elderly people at home a more feasible and realistic task.

In particular, this thesis proposes a set of methods for human presence detection, localisation and body pose tracking that are operating on the perception level of the paradigm. Also, the problem of having only limited visibility of a person from the on-board sensors of the robot is addressed by the proposed classifier fusion method that combines information from several types of sensors. A method for improved real-time human body pose tracking is also investigated. Additionally, a method for estimation of the multiple human tracks from noisy detections, as well as analysis of the computed human tracks for cognition about the social interactions within the social group, operating at the comprehension level of the robot's situational awareness paradigm, is proposed. Finally, at the human-aware planning layer, a method that utilises the human related information, generated by the perception and comprehension layers to compute a minimally intrusive navigation path to a target person within a human group, is proposed. This method demonstrates how the improved human perception capabilities of the robot, through its judgement activity, can be utilised by the highest level of the paradigm, i.e. the decision making layer, to achieve user friendly human-robot interactions.

Overall, the research presented in this work, drawing on recent innovation in statistical learning, data fusion and optimisation methods, improves the overall situational awareness of the robot in regard to people with the main focus placed on human sensing capabilities of service robots. The improved overall situational awareness of the robot regarding people, as defined by the proposed paradigm, enables more meaningful human-robot interactions.

Acknowledgement

First and foremost, I would like to thank my supervisors for the friendly guidance and encouragement. Also, I would like to thank the SRS Project’s team for the inspiration that they gave me during my early steps into the exiting world of robotics.

Dedication

To my family for putting up with me during the long hours of my research, especially when I was writing up.

Acronym	Description
pdf	Probability density function
GPU	Graphical processing unit
IK	Inverse Kinematics
HISP	Human Interpretable Signature of Points
CHISP	Conformal HISP
SHISP	Simplified HISP
DOF	Degrees of Freedom
HRI	Human-Robot Interaction
EDR	Elderly Dependency Ratio
MOCAP	Motion Capture (system)
SVM	Support Vector Machines
MHT	Multihypothesis Tracking
GLAICP	Global Local Articulated Interactive Closest Point
AAI	Area of Active Interactions
DBN	Dynamic Bayesian Network
MCMC-PF	Markov Chain Monte Carlo based Particle Filter
PAM	Pairwise Affinity Measure
HPAM	Historical Pairwise Affinity Measure
MOTA	Multiple Object Tracking Accuracy and
MOTP	Multiple Object Tracking Position
MSMRPC	Method for Selection of the Most Realistic Pose Configuration
MPA	Map of Prohibited Areas
HPDL	Human Presence Detection and Localisation
HPDT	Human Pose Estimation and Tracking in 3D
HTENP	The Human Track Estimation and Navigation Planning
RF	Random Forest (classifier)
TOF	Time-of-Flight
FP	False positive
FN	False negative
HMM	Hidden Markov Model
AAI	Area of Active Interactions
ICP	Interactive Closest Point
AICP	Articulated Interactive Closest Point
DST	Dempster-Shaffer Theory
OWA	Ordered Weighted Averaging
JPDAF	Joint Probabilistic Data Association Filter
IMKF	Identity Management Kalman Filter
MHT	Multihypothesis Tracking
HOG	Histogram of Oriented Gradients
SVM	Support Vector Machine
HAMP	Human–Aware Motion Planner
FOV	Field Of View
SLAM	Simultaneous Localisation and Mapping
LRF	Laser Range Finder
PCA	Principle Component Analysis

1 Introduction

"The First Law of Robotics: A robot must not injure a human being or, through inaction, allow a human being to come to harm." — Isaac Asimov

1.1 Motivation

Humans are social beings by nature. It is only natural for a person to want to interact with other people, to depend on them and to live with the hope that one could depend on loved ones in times of difficulty. Unfortunately, due to modern lifestyles and demographic trends, there are cases when people cannot always be surrounded by their family and extended family members. At times like these they need the support and care of someone else to cope with the challenges of everyday life. Due to the 'demographic transition', the elderly have become the fastest growing group in many developed countries (Bloom et al. 2011). In particular, the Elderly Dependency Ratio (EDR), which measures the ratio of the number of elderly dependants, aged 65 or over, to people of working age, aged 15 to 64, is predicted to double in Europe between now and 2050. The worst affected countries worldwide are expected to be Japan, Spain and Italy with EDR predicted to reach up to 60%-70% and Germany, France and the UK following closely with EDR between 45%-60%, as depicted in Figure 1-1. This trend, together with the increase in life expectancy, allowing people to live longer on their own with disease, disability and frailties, leads to significant policy challenges including: a) the need to reconfigure support provisions to recognise these new, considerably extended lives and b) the need to finance the cost of maintaining quality of life, particularly at the end of life (Harper & Hamblin 2014). How can modern societies successfully adjust to population ageing? Many people believe that prolonging elderly independent living at home, through additional support from the community and extended family members, will allow them to live with dignity, and as much independence as possible, while giving them the preferred choice to remain in their familiar home environment. Indeed, independent living in one's home, for as long as possible is a crucial factor in maintaining a high quality of life and is the preferred option for many elderly people. A growing number of social policies across the world have embraced this understanding, and, as a result, the concept of person-centred care at home has been embodied in social policies in some countries. The emphasis of such policies is on the individualisation of the care provided at home with the ultimate goal to provide a sufficiently high quality of life for elderly people. Key players in this process are the social care organisations that are relying on limited social budget funds. At the same time, there are growing concerns over the ability of nations to finance the social security and the long-term social care required to support the growing number and percentage of older dependents, at a time when the number and proportion of those who are economically active is declining (Harper & Hamblin 2014).

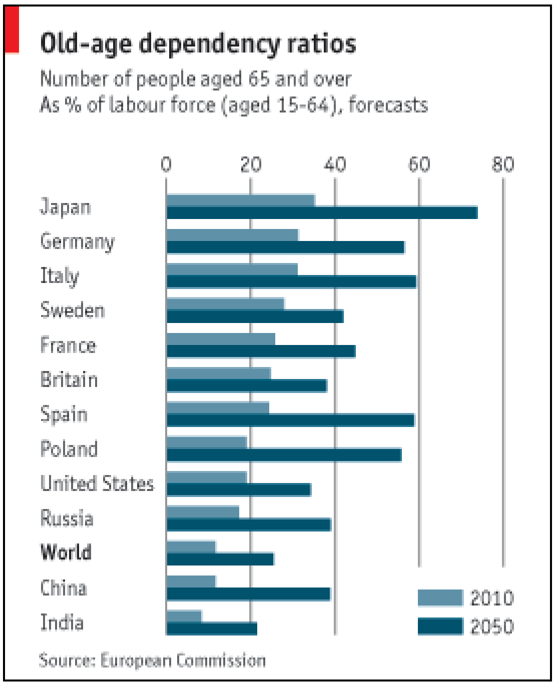


Figure 1-1: Forecast for Elderly Dependency Ratios in Different Countries arround the World (EUROSTAT, 2012)

One simple approach to the above problem could be the involvement, to a higher degree, of the extended family members to support and care for their elderly relatives. However, as is often the case in modern industrialised societies, the family members cannot provide support. They either live too far away or cannot stay at the home of the elderly person for extended periods of time due to work or other family commitments. Therefore, another more radical approach is needed.

With the advancement of robotics technology, the idea that the above gap in the care provision for elderly could be filled, at least to an extent, by autonomous or semi-autonomous mechanical helpers, acting as a remote proxy of the relative or the caregiver, at the point of delivering the service, is becoming an increasingly attractive one. Indeed, in a survey of 2000 people, 83% indicated that they were willing to accept a robot helper to retain independence. Also, 71% reported that they expect to be relieved from laborious tasks (Arras & Cerqui 2005). This trend has also been recognised by the increase of research funding related to service robotics in recent years. This increase is evidential of the increasing importance given to robotics technology to counteract the ageing demographic trends in many developed countries.

However, as mobile robots move to replace human caregivers in the home and nursing home institutions, they face a broad range of challenges that are not typical for the structured industrial environments where they typically operate. The unsuitable design of the human homes for a robot operation, e.g. multi-floor homes with staircases, represents a major obstacle to the successful application of service robots at home. Moreover, natural interactions with people also prove to be a very challenging area for a robot. A personal service robot should be able to engage with people using social interactions that are acceptable from a human point of view. In particular, a robotic system should follow closely the social interaction patterns that are typical for the particular context in which the interaction takes place. As it is not possible yet for a machine to think in the same way as people think, the focus has to be put on simplified but adequate computational models that can interpret human social behaviour and give appropriate reactions that are acceptable by the users. At the same time, it is essential that the outcome of the robot's reasoning process be in line with the human expectations as this will increase acceptance and trust in the users of the system. Indeed, it is considered that a service robot will be able to replace a human caregiver successfully only when it is capable of providing a comparable experience . This can be achieved through following relevant and socially accepted rules of interaction. As a result, by operating in a manner sensitive to the human, a service robot will be able to provide an individualised care service. Moreover, such a service will be suited to the user's needs and win their trust by meeting their expectations. Building better trust between the robot and the user will ultimately lead to increasing acceptance or take-up rates and more success for the service robotics in meeting the needs of society.

In conclusion, the main problem that the service robotics research community needs to resolve in the near future, is finding appropriate computational models, algorithms and supporting mechanisms, aimed at enhancement of the assistive capability of personal service robots. This will consequently enable mobile robots to become efficient caregivers at home and meet the needs of the aging population in a number of countries, One of the main elements essential for achieving the above goal is the improvement of the human perception abilities of the robot to a level that enables reliable robot control for close Human-Robot Interactions (HRI). Monitoring of the location of the personand their pose, as well as having the ability to analyse the human motion patterns, are considered important parts of the perception abilities of a service robot and, therefore, are addressed in this thesis.

1.2 Aim and Objectives

The **aim of this work** is to **enable personal service robots to engage in more natural and socially acceptable interactions with people by improving their human perception capabilities.** The human perception capabilities of a robot are defined as capabilities related to presence detection, localisation, position tracking, pose discovery and tracking of the people in the environment.

The driving force behind the definition of the above aim has been the desire to make physically assistive multi-functional service robots for elderly care at home more feasible in the near future. In particular, it is believed that empowering the service robots with human perception capabilities enables closer, more meaningful and natural human-robot interactions through more optimal control of the robot's behaviour. Ultimately, better human perception contributes to a safer and more meaningful coexistence between people and service robots, in the context of assisted living at home. The aim of this work is approached through investigation of a number of novel methods for human detection and localisation, pose estimation and tracking with increased robustness. Additionally, estimation of interpersonal interactions, achieved through analysis of human tracks, enables a service robot to reach a target person within a group while avoiding interruption to the ongoing social interactions. Finally, the proposed human-aware navigation method, which uses information from all previously proposed methods, demonstrates how the interaction between a service robot and its user can be improved by enhancing the robot's human perception capabilities. The interface between the low-level methods for human sensing and the high-level planning and decision making activities of the robot is governed by theparadigm that I propose for situational awareness in regards to people.

A number of objectives have been set to guide the research effort in this project. Firstly, the exact position of the person has to be inferred from the data, generated by the on-board sensors, to guarantee a successful interaction between a human and a robot. The research challenge behind this objective is to find a way of reconciling the contradictory readings of the sensors and to compensate for the missing readings.

Although several commercial systems for human tracking and motion analysis are currently available, they are not considered to be well suited for a service robot application. Typically, these systems rely on a range of special tracking technologies that require special callibration procedures or installation of multiple motion sensors throughout the environment. These technologies include: wearable optical markers, wearable magnetic sensors and multiple tracking cameras. The above human tracking systems are optimised for accuracy but are deemed to be either too intrusive or too cumbersome for everyday use in a service robotics scenario. It is considered that in order for a service robot to be accepted by its users, the intrusion into the user's life should be kept to the minimum and limited to the intrusion level caused by a caregiver. This condition can be achieved only by eliminating the multiple camera system configurations: relying solely on sensors on board the robot and avoiding any wearable markers. It is considered that robots, like people, should use only their on board sensors for their orientation and navigation to avoid any intrusion of privacy of their users. Therefore, the mimicking of the human sensing abilities by a robot that is using only its on-board sensors is considered to be an important limitation, followed in this work, for avoiding privacy issues and for gaining the user's trust. Indeed, as reported by Arabo et al (2012), placing multiple cameras in a smart home environment is considered a considerable privacy issue by the user. In practice, this can be avoided by limiting the sensors used only to those available on board the mobile robot. However, this configuration represents a significant challenge, associated with the inherent properties of the sensors, i.e. limited Field Of View (FOV), high noise level, limited useful detection range and low accuracy. The ambiguity, caused by the reduction of the number and the spatial distribution of the sensors, is addressed in this work by developing stochastic sensor models and using probabilistic inference.

Additionally, since the mobile robots already have a range of sensors on board that are gathering information about the surrounding environment, adding more sensors designated specifically for human sensing is considered a suboptimal approach as it requires additional resources, e.g. power. Therefore, an additional requirement is placed for the human sensing activities of the robot to re-use information about the environment that is already collected by the existing sensors of a typical service robot.

Based on the above analysis, ***the first objective is to investigate methods for improving the reliability of information about the human presence and location from existing on-board data sources.***

Although having reliable information about the presence and location of a human is considered to be a good foundation for human-aware robot control, on its own it is insufficient for delivering an efficient care service through meaningful human-robot interactions. For example, a number of close physical interactions, like the passing of an object, e.g. a glass of water, between the person and the robot requires reliable human pose information to allow fine-tuned control of the robot's manipulator in the vicinity of a person. If there is no human pose information or the available information is not sufficiently accurate or reliable, then the risk of collision between the arm of the robot and the human is still considerable and severely limits the range of possible physical interactions between a human and the robot. The current state of service robots in this respect has prompted the definition of the second objective of this work.

The second objective is to define a method for human pose tracking that improves on the reliability of existing methods.

Finally, although in a typical scenario most of the time the elderly person is alone at home, it is considered that the provision of a successful elderly care service also requires the ability of the robot to approach the person while he or she is engaged in interaction with other people, e.g. visiting relatives or caregiver. In such cases, it is important that the robot is able to behave appropriately when dealing with groups of people, e.g. by not obstructing the social interactions of a group, as socially compatible conduct is one of the key factors in theacceptance of robots . Therefore, in addition to the current human position and the human pose information, the robot should be able to evaluate the level of human interactions between a group of people and adapt its behaviour dynamically to approach the target person without interruption of the ongoing and potential human interactions.

The third objective for this work is to investigate a method for a) analysis of human tracks and b) adaptation of the robot behaviour in accordance with the infereded by the analysis interactions(I think "results" works better), in order to minimise the disruption to the interactions within the group of people.

1.3 Approach Overview

This thesis, instead of focusing on a single narrow topic, covers a broader range of subjects that are relevant for the achievement of the set aims and objectives. These include investigation of sensor properties with regard to human sensing, signal processing for human detection from multiple modalities, fusion of the individual detector outputs, human pose discovery and tracking, human track identification from noisy detections of multiple people. Overall, the combination of the proposed methods together with the underlying logical interconnections between them represents a consolidated system, referred to as a human perception framework, aimed at enhancing the human perception capabilities of the service robots. The human perception framework is structured in hierarchical layers through the paradigm, proposed below, for the robot's situational awareness in regard to people. Managing the complexity of the overall system for human motion is a challenging task. In this work, the systems engineering design approach is adopted for this purpose. As a result of the analsysis of the requirements in domestic service robotics; the concept, architecture, components, modules, interfaces between modules, and data exchange within the human perception framework are defined. In the conceptual model the main emphasis is placed on the definition of the logical building blocks of the system and interconnecting interfaces between them. Later, appropriate methods are investigated in each block to fulfil its defined function. Algorithms are developed for each method for benchmarking. Full optimisation of the code is considered a more technical challenge which is not pursued in this work.

1.4 Scope Limitation

As there is a wide range of possible modality and sensor combinations in regard to human to human detection, a limitation of the scope is considered necessary to allow sufficient focus to be achieved on the sensor modality combinations that are most typical for the contemporary service robots. Therefore the following modality limitations are introduced:

- The methods rely only on the typical sensors located on-board a service robot
    - Colour camera
    - RGB-D camera
    - Laser range finder

Human detection by other modalities, i.e. those which the current service robots do not typically have pre-installed sensors, e.g. sound arrays and thermal infrared cameras, is considered outside the scope of this work as installation of any additional sensors will increase the cost of the robots and make them even less affordable.

- The methods use only the on-board computational resources of a robot. As the robot has only limited computational and power resources, therefore the range of investigated methods is limited to only those methods that can operate on a single microprocessor. Methods that require significant computational resources, e.g. a parallel processing are not considered in the scope of this work;

- As a service robot operates only in an indoor environment the methods are designed specifically for indoor operation. This limitation enables usage of low cost infrared based sensors, e.g. RGB-D cameras for depth sensing which will otherwise be impossible as outside the interference of the infrared component of sunlight prevents the infrared based sensors operating normally.

In a different perspective, the research scope of this work is limited to the first three layers, i.e. perception, comprehension and human-aware planning, of the proposed context awareness in regards to people paradigm, described in 1.8. For example, high-level task planning and decision making activities of the robot control reside on a higher levels and therefore are considered outside the scope of this work.

A typical service robot, the Care-O-Bot3 robot (IPA (Fraunhofer) 2014), is used for validation of the proposed methods. It features all typical sensors for a service robot as described in the section below.

1.5 Experimental Platform

The experiments in this work have been carried out using the Care-O-Bot3 (IPA (Fraunhofer) 2014) mobile robot platform. The sensors available on board Care-O-Bot3 represents a typical set of sensors available on the contemporary service robots. This allows validating of the proposed methods for a wide range of service robots on the market.

Designed as a research platform, Care-O-Bot3 measures 1.45 metres in height. It has a manipulator arm on its right side with seven degrees of freedom. There are three fingers on the hand of the robot with tactile sensors on the tips enabling it to grasp various household objects using force feedback. There are three laser range finders at the base of the platform enabling full 360 degree view of the environment. The movable tray has an embedded touch screen display which can be used either for placing objects when serving a user, or as a graphical interface with the user who can select from a menu of options displayed on the tray. The remaining sensors are mounted in the laser head providing them with an elevated position. The set of sensors consists of a 3D Time-of-flight camera, recently replaced by a RGB-D sensor, and two RGB cameras working together as a stereo sensor, shown in Figure 1-2 below.

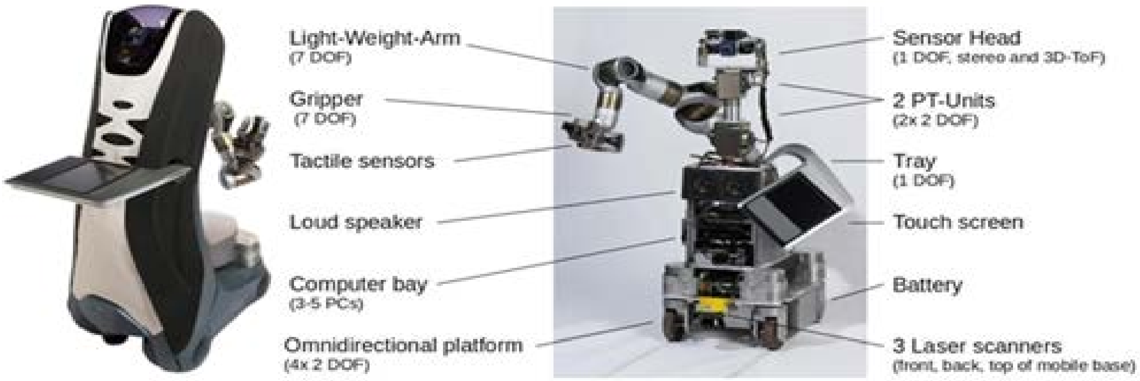


Figure 1-2: Care-O-Bot3 platform *(image courtesy of IPA, Fraunhofer)*

Most of the experiments aimed at evaluation of the proposed method have been carried out using Care-O-Bot3 in a simulated home environment. Some of the experiments, which do not require using the full robotic system, have been carried out in isolation on a module-by-module basis, i.e. by using the required sensor sub-system of the robotic system in an equivalent configuration, for example the RGB-D sensor or the laser range finder. In addition to recording the data from the robot's sensors, the position of the people in all experiments has also been recorded in parallel by a Motion Capture (MOCAP) System. The MOCAP data has been used as a ground truth for validating the results obtained by using the proposed methods.

1.6 List of Relevant Publications

The following publications have been made fully or partially based on concepts developed in this work:

- Noyvirt A., Setchi R., Albert B., GLAICP: a global–local optimization algorithm for robust human pose tracking from depth data, SMC, 2014, pages 2541-2546, DOI: 10.1109/SMC.2014.6974309;
- Noyvirt A., Qiu R., Human detection and tracking in assistive living service robot through multimodal data fusion, Industrial Informatics INDIN, 2012, pages 1176-1181, ISBN: 978-1-4673-0312-5;
- Qiu R, Noyvirt A, Ji Z, Soroka A, Li D, Liu B, Arbeiter G, Weißhardt F, Xu S, Integration of symbolic task planning into operations within an unstructured environment, International Journal of Intelligent Mechatronics and Robots, 2 (3) (2012) 38-57, DOI: 10.4018/ijimr.2012070104;
- Soroka, A.J.; Renxi Qiu; Noyvirt, A.; Ze Ji, Challenges for service robots operating in non-industrial environments, INDIN, 2012, pages 1152-1157, DOI: 10.1109/INDIN.2012.6301139;
- Ze Ji, Renxi Qiu, Noyvirt, A., Soroka, A., Packianather, M., Setchi, R., Dayou Li; Shuo Xu, Towards automated task planning for service robots using semantic

knowledge representation, INDIN, 2012, pages 1194-1201, DOI: 10.1109/INDIN.2012.6301131;
- Qiu, R.; Ji, Z.; Noyvirt, A.; et al., Towards robust personal assistant robots: Experience gained in the SRS project, IROS 2012, pages 1651-1657, DOI: 10.1109/IROS.2012.6385727;

1.7 Outline

This thesis is structured as follows:

- Chapter 1 introduces the focus of this thesis, describes the motivation behind this work, discusses key challenges of service robotics and lists the contributions and relevant publications resulting from this research;
- Chapter 2 gives background information about service robotics and identifies the problems related to human sensing.
- Chapter 3 reviews related work;
- Chapter 4, addresses the first objective: presenting a sub-system for detection of people in the environment using data input from typical sensors of a mobile robot;
- Chapter 5, addresses the second objective: describing human pose tracking-related problems and the proposed tracking method;
- Chapter 6, addresses the third objective: investigating the analysis of interactions between people in a group from the available detections and presenting the proposed framework for planning of the human-friendly navigation of the robot according to the social interaction rules;
- Chapter 7 discusses the results from the experiments with the methods proposed in Chapter 4, Chapter 5 and Chapter 6;
- Chapter 8 evaluates the achievement of the aims and the objectives of this work, discusses potential future directions for the research and makes final conclusions.

1.8 The Bigger Picture

Although the main focus of this work is defined as enhancing the human perception capabilities of service robots, addressed by Chapter 4 and 5, it is important that a method for integration of the perception information into the higher-level control of the robot is also considered. This allows integration of the proposed methods into the overall control of the robot. Therefore, a novel Robot's Situational Awareness Paradigm in regard to people is proposed, shown in Figure 1-1, to provide a structural view of the relation between human related activities of the robot at different level of abstraction. By positioning the methods proposed in this work into the overall control structure of the robot, the paradigm defines how the overall HRI can benefit from the improved human perception capabilities of the robots.

The proposed paradigm consists of a four layer stack which defines the control action links between each layer, from the low level perception tasks to the high level HRI. In particular, the control action links are Cognition, Judgement and Decision (making). The methods from each layer provide information to the higher level. For example, the tasks in the Perception layer provide human detection information, location coordinates, human pose configuration information to the Comprehension layer. The Comprehension layer, through cognition control action, analyses the available information in order to recognise the human actions, determine their intentions and estimate their interactions with objects or other people in the environment. At the next layer, i.e. Human-Aware Planning layer, appropriate navigation and action plans are generated through the judgement activity of the robot, using the available information from the comprehension layer. Finally, at the Decision Making layer, a decision is made regarding which of the available plans is to be enacted in order to enable the achievement of socially acceptable interaction with people.

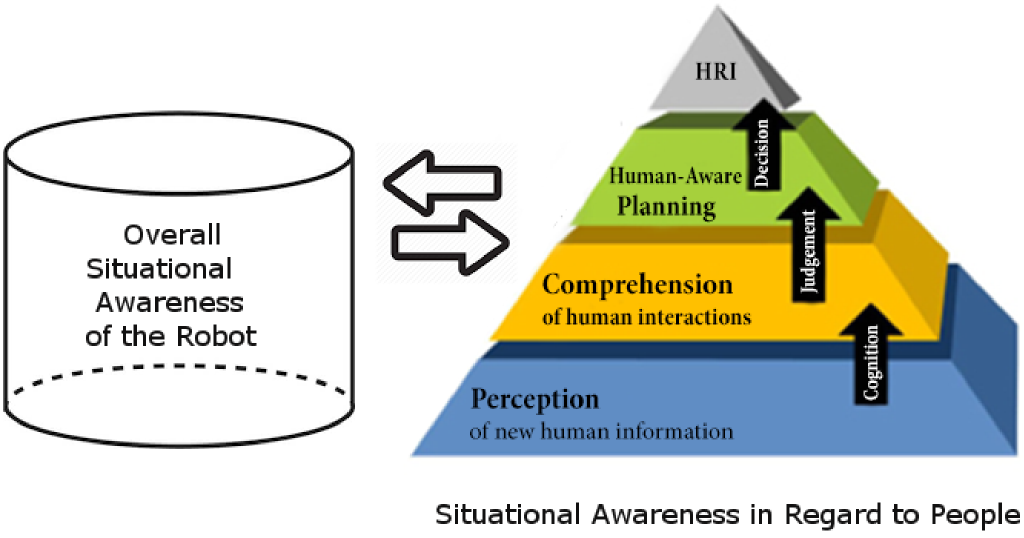


Figure 1-3: The proposed robot's situational awareness paradigm in regard to people

When mapped to the above paradigm, the methods proposed in Chapter 4 and 5 are associated with the first layer of the stack, i.e. "Perception of new human information" layer. The proposed method for analysis of the human interactions using human detections, described in the first part of Chapter 6, is considered to be part of the "Comprehension of human interaction" layer of the paradigm as it analyses human detections to generate human tracks and estimate the interactions between people. The method for generation of a minimally intrusive navigation path, forming the second part of Chapter 6, resides on the third layer of the stack as it generates a navigation path plan that minimises disruption of the human interactions. In this sense, Chapter 6 provides an example of as to how the results of the perception of new human information methods, the main focus of this work, can be integrated into the higher cognition and judgment activities of the robot. The methods described in Chapter 6 are considered to be examples of building blocks of the second and third layers of the stack.. As discussed in Chapter 8, future research work could address the full functionality of the higher levels of the paradigm in order to enable the insertion of newly available algorithms at various layers. Finally, the higher layer, HRI, which is a result of the Decision Making (DM) activity of the robot using the optional plans, is also considered to be outside of the scope of this work and therefore it is not addressed here.

2 Literature Review

This chapter provides a review of relevant work in human detection, localisation, pose tracking and human-aware robot navigation, with focus placed on research aspects that are related to the research presented in the later chapters.

Markerless human motion capture has been the focus of research for a considerable period of time. Such systems are still often challenged by problems such as self-occlusion and ambiguity. Several surveys provide an extensive overview of the sensors, models and human estimation methods for markerless human motion capture and analysis (Moeslund et al. 2006; Kolsch et al. 2007; Cedras & Shah 1995).

Camera-based methods have been the main focus of the human motion research so far (Agarwal & Triggs 2006; Lee & Nakamura 2007; Kehl & Van Gool 2006). Lately, methods associated with the recently developed depth sensors, e.g. time-of-flight cameras (Plagemann et al. 2010) and structured-light systems (Shotton et al. 2011), are gaining popularity as they can use the depth component of the image to greatly simplify the segmentation, detection and tracking.

2.1 Human Detection and Localisation

Automatic human detection has been used in a broad range of applications, such as safe robot navigation, visual surveillance, human-computer interface, performance measurement in sports, rehabilitation of patients with physical disabilities , behavioural studies of consumers in the retail industry, interactive gaming in the entertainment industry and autonomous vehicles within transport.

In service robotics, establishing whether or not there are people present in the environment and finding their precise location with a sufficiently high level of probability enables both safer co-existence of humans and robots and a higher level of cooperation between them. In particular, the availability of reliable and accurate information about the people around the robot enables an extended range of path planning and collision avoidance algorithms that make intelligent use of this information (Sisbot et al. 2007).

In transportation, where the ultimate goal is saving lives by preventing accidents due to human error and enabling autonomous self-driven cars, a related problem of automatic pedestrian detection has also been the focus of extensive research (Dalal & Triggs 2004; Felzenszwalb et al. 2008; Wu & Nevatia 2005). Although in the case of pedestrian detection, the detection range is typically much bigger in comparison with an indoor service robotics scenario, the underlying principles are very similar (Dollár et al. 2012). Many systems, in addition to the detection of the presence of pedestrians, also make use of human body silhouettes to estimate the posture configuration

(Gavrila & Munder 2006; Leibe, Leonardis, et al. 2008; Sharma & Davis 2007; Wu & Nevatia 2007). Still, due to the wide variability in appearance related to differences in clothing, -light conditions, scale changes, high level of articulation and frequent partial occlusion, pedestrian detection remains a challenging task that requires further reliability improvements and legislative changes before an introduction into road use.

Irrespective of the application domain, human detection is based either on individual sensor modalities like monocular cameras, e.g. (Enzweiler & Gavrila 2009), stereo or multiple cameras e.g. (Nedevschi et al. 2009), laser range finders, e.g. (Fod et al. 2002), depth cameras, e.g. (Luber et al. 2011), thermal imaging, e.g. (Wang et al. 2010), or on some combination of the modalities through data fusion techniques, e.g. (Walk et al. 2010; Bellotto & Hu 2009). Since the methods used for human detection are loosely specific to the particular modalities, they are reviewed by modality in the following sections.

2.1.1 Detection of People in Colour Images

Analysing imagery in order to determine if one or more people are present in the scene, a task sometimes referred to as figure ground segmentation, has been studied extensively (Moeslund et al. 2006; Piccardi 2004; Forsyth et al. 2005). However, as argued by Zhang (2010), despite the substantial research effort, an approach that is able to match the superior versatility of the human brain to processes visual information in semantic space for detection of people has not yet been found.

A typical initial step in the process is the background removal, which in the case of a mobile robot is a challenging task due to the non-static nature of the observation viewpoint and the resulting motion of the background image. A comprehensive review of the background removal techniques is given by Piccardi (2004). Two main approaches to figure-ground segmentation exist: pixel-based and object-based (Moeslung et al. 2011). In the former, each pixel in the image is compared to a model for that pixel in order to assess whether the pixel is foreground or background (Stauffer & Grimson 1999; Kim et al. 2004). The latter approach operates by translating and scaling a window and subsequently calculating the probability of the window containing a human image (Dalal & Triggs 2004; Leibe et al. 2005; Felzenszwalb et al. 2008; Dollar et al. 2009). In this process the human detection is reduced to a positive/negative classification decision for the particular window. For the classification phase a number of machine learning methods can be used. In particular, Random Forest (Yali & Donald 1997), Support Vector Machines (SVMs) (Vapnik 1995), AdaBoost (Freund & Schapire 1997) have been reported as achieving notable results in the past.

While the result from an object-based method is given by a bounding box, representing the positively classified window containing the human, the result from a pixel-based method is a blob or a silhouette, an output that is better suited for a direct human pose configuration fitting. At the same time, pixel-based methods are better suited for static backgrounds, due to the difficulties of modelling non-static scenes, such as the ones resulting from the non-trivial motion of the observing camera of a moving robot platform. Currently, the research in this area is focused mainly on how to update the pixel model during runtime to compensate for the changes in the scene. Approaches employing multiple models per pixel, for example Stauffer & Grimson (1999), and approaches based on stochastic approximation (Lopez-Rubio & Luque-Baena 2011), have demonstrated encouraging real-time performance results and could become particularly relevant for the needs of human detection in the context of service robotics in the future. In contrast, the part-based human detection methods, e.g. the Deformable Part Model (DPM), model people as a collection of human body parts, hypothesis about which is generated initially by using local features, e.g. edgelets (Wu & Nevatia 2005) or orientation features (Mikolajczyk et al. 2004). Subsequently, the separate part hypothesis is jointed to form a single hypothesis of a human presence in the bounding box. Although attractive, this approach is considered very heavy due to the number of steps involved, i.e. creating a densely sampled image pyramid, computing features at each scale, performing classification at all possible locations, and finally performing non-maximal suppression to generate the final set of bounding boxes (Cho et al. 2012).

As, in principle, the feature-based systems tend to operate much faster in comparison with the pixel-based systems, the majority of human detection research is currently focused on developing feature-based methods. The most notable developments in this area are reviewed below:

The SIFT theory (Lowe 1999), based on detection of local features, describes the appearance of the object in a particular spot. As the local features are invariant to image scale and rotation, they allow accurate detection of objects without the need foran extensive search in scale space. Subsequently, the SIFT descriptor has been further improved to provide robust matching in challenging conditions, e.g. across a substantial range of affine distortion, change in 3D viewpoint, addition of a noise component, and change in illumination (Lowe 2004).

Haar wavelet based features and Support Vector Machines (SVM) were proposed to be used for detection of objects and people (Papageorgiou & Poggio 2000). In related research, (Mikolajczyk et al. 2004) proposed human detection using human body parts detection and modelling the human body as a flexible assembly, or parts where the feature extraction and learning is based on Integral Images (Papageorgiou et al. 1998) and Adaboost (P. Viola & Jones 2001). The above approach is reported to work well even when the people in the images are partially occluded. Such a result is particularly relevant for observation from a single viewpoint, a restriction resulting from user privacy considerations in service robotics. Variations of the above method have been reported subsequently by Hou et al. (2007) and Luber et al. (2011). Overall, although the run-time performance of the above approach is satisfactory for real-time human detection, the required training phase is computationally demanding and necessitates the collection of extensive training image set.

Unlike most methods, relying on separate object segmentation and object class recognition, the Implicit Shape Model (ISM), a patch-based approach, combines both a single probabilistic framework that also generates a per-pixel confidence measure (Leibe et al. 2004). In this approach a codebook of local appearance is initially learned during the training process. Later, in detection, the extracted local features are matched against the existing codebook entries. When a match is detected it casts a vote for the hypotheses of presence of a pedestrian. Unlike the above Haar feature based approaches, the ISM approach requires a much smaller set of training images.

In another approach, a combination of local and global cues is used by Leibe et al. (2005) to do a probabilistic top-down segmentation in order to detect pedestrians in complex scenes. The simple set of features, namely the Histogram of Oriented Gradients (HOG), used to train a soft linear SVM classifier, is reported to perform extremely well in detection of humans in RGB images (Dalal & Triggs 2004). In fact, the human tracking framework reported by (Leibe, Leonardis, et al. 2008), based on the Implicit Shape Model (ISM) detector (Leibe, Leonardis, et al. 2008) and later also on Histogram of Oriented Gradients (HOG) detector (Dalal & Triggs 2004) or the Deformable Part-based Model (DPM) detector (Felzenszwalb et al. 2008) provides an opportunity to compare the three methods involved, i.e. ISM, HOG, DPM. In the comparision, although all three methods are reported to give comparatively satisfactory results, the HOG detector using frontal view of pedestrians has demonstrated a superior performance.

While the idea of using only monocular images is an attractive proposition, fitting in with the requirements of human detection in service robotics, it faces considerable challenges especially in the figure-ground segmentation phase. Efforts for increasing the resilience to disturbance using context information has been reported to use different approaches including: using motion information (Dalal et al. 2006; Viola et al. 2003), stereo depth (Ess et al. 2007; Gavrila & Munder 2006), scene geometry (Hoiem et al. 2006; Bastian Leibe et al. 2007), temporal continuity (B. Leibe et al. 2007; Wu & Nevatia 2005) or semantics (Murphy et al. 2003; Ommer & Buhmann 2005; Sudderth et al. 2005; Torralba 2003). The above methods have been applied as a complimentary context cue to achieve improved robustness of the human detection.

Human face detection, as a special case of human detection, when possible, e.g. in situations when the person is facing the camera, is a very powerful contributing factor to the overall human detection. In addition, it allows not only detection but also identification of the person. Then the identity information can be used for association of previous knowledge about the person with the current status informationof the person. As surveyed by Zhang & Zhang (2010) the most popular method for face detection, evolved as the de-facto standard of face detection in real-world applications, is the method reported by P Viola & Jones (2001), based on Integral Images and the standard AdaBoost algorithm (Freund & Schapire 1997). In later works, the use of RealBoost and GentleBoost has been proposed (Lienhart et al. 2003; Brubaker et al. 2008) with the aim of improving the performance of the face detector. Other works have proposed reusing previous classification results (Wu et al. 2004; Xiao et al. 2003), introducing asymmetry (Wu et al. 2008; Pham & Cham 2007; Paul a Viola & Jones 2001), setting intermediate thresholds during training (Šochman & Matas 2005; Xiao et al. 2007), or after training (Luo 2005; Bourdev & Brandt 2005), speeding up the training (Wu et al. 2003; Pham & Cham 2007) or the testing (Li et al. 2002) process, learning without subcategory labels (Seemann et al. 2006; Shan et al. 2006). Although face detector algorithms are used successfully in a number of real applications, with the recent state-of-the-art detectors reporting up to 70% detection rates and 0.5-3% false positive rates (Zhang & Zhang 2010), it is considered that face detection in completely unconstrained settings still remains a very challenging task, which necessitates further research (Jain & Learned-Miller 2010).

Overall, the majority of the methods for human detection need a uniform background for overcoming the challenges associated with the figure-ground segmentation phase. This restriction, in a mobile robot application scenario, represents a significant challenge due to the constant need for motion of the platform. Therefore, it is considered that in a service robot scenario, the information from a monocular camera is not sufficient for a reliable human detection. Thus, its augmentation by sensor fusion method with information from all other available sources will benefit the overall performance of the human detection. This approach is one of the key directions of this work.

2.1.2 Detection of People in Depth Images

Naturally, every human motion is governed by the laws of physics. These laws are part of the available background knowledge about human motion. The background knowledge, in combination with the current and past measurements from the sensors, enables estimation of the most probable human pose in any given time frame.

The first problem in the detection of people in depth images is to remove the redundant measurements, i.e. the clutter, resulting from sensor noise and other insignificant for human detection artefacts in the scene, e.g. objects like walls, floors and others. This task, known as background segmentation, enables only measurements from the person to be used in human motion estimation. Although the task of background segmentation has been simplified significantly by the introduction of depth cameras, a number of challenges still remain due to the high redundancy, uneven sampling density and the lack of explicit structure of the point cloud generated by the sensors (Nguyen & Le 2013).

Previously, the task of partitioning a set of measurements in the 3D object space into smaller, coherent and connected subsets, a process referred to as a point cloud segmentation, has been approached in different ways. Earlier examples are based on graph clustering methods (Douillard et al. 2011), e.g. Min Cuts (Wu & Leahy 1993), Normalised Cuts (Shi & Malik 2000) and Graph Cuts (Boykov & Funka-Lea 2006).

Later, (Golovinskiy & Funkhouser 2009) extended the graph cuts segmentation to point cloud data by using k-nearest neighbours (KNN) to build a 3D graph and assign edge weights according to an exponential decay in length. However, this method requires prior knowledge of the position of the object. Such a prior knowledge is not always easily available, which acts as a restricting factor in the application of the above methods.

The most efficient segmentation algorithms to date rely on reducing the dimensionality of the point cloud, which is achieved by projecting the 3D points to a 2.5D grid fixed to the ground plane (Himmelsbach 2010). For this reason the algorithm for segmentation of 2D images proposed by Felzenszwalb & Huttenlocher (2004), has proven to be very efficient and as a result has gained immense popularity in robotics, where the computation resources are typically severely limited (Schoenberg et al. 2010; Strom et al. 2010).

However, one problem encountered by the methods, based on dimensionality reduction, is under-segmentation. This problem manifests when points belonging to different objects are given as belonging to a single segment. The reason for this is the lack of modelling of the free vertical space between any pair of points in the 2.5D representation. Therefore, to counteract the above problem, methods addressing the segmentation in full 3D have been proposed (Moosmann et al. 2009; Klasing et al. 2009; Anguelov et al. 2005; Steinhauser et al. 2008). The first two methods use the characteristics of the scanning to establish neighbourhood relationships between points. Then, they use the neighbourhood relationship between points to extract local point features from estimated point normals, and finally to compute smoothness features that are used to segment the objects in the scene. The method of Moosmann et al. (2009) achieves nearly real-time performance while the method of Klasing et al. (2009) is reported to work at full real-time performance. However, both methods utilise the sensor properties and are not readily applicable to any generic point cloud. In contrast, the method of Anguelov et al. (2005) approaches the segmentation by applying machine learning techniques to train a Markov Random Field so it can classify points from the dataset using simple features. Although good results are reported by both methods, a real-time segmentation process is still not possible for big data sets (Himmelsbach 2010).

Even after the successful background segmentation, the other significant challenge for human detection, i.e. the non-rigid human body shape, which cannot be fitted into a single template, still remains. Solving this problem has been attempted by a number of approaches. These have been based either on a matching of edge features (Mori & Malik 2006; Toyama & Blake 2002) or silhouette features (Agarwal & Triggs 2004; Shen et al. 2009) to achieve human body pose estimation. Although many silhouette-based methods are reported, they tend to have an unacceptably high error rate, due to their tendency to get confused by scene clutter. Moreover, they are unable to observe the human limbs in images when they are positioned in front of the human body and this leads to highly ambiguous silhouette data.

More recently, a promising new trend for body human detection, based on Random Forests (Breiman 2001) learning of features, has emerged. Random Forests is an ensemble learning method for classification that operates by constructing multiple decision trees at training time and outputting the class that is made of the output by the individual trees that match the criteria. As the rate of convergence of the procedure of prediction depends only on the number of strong features, and not on the number of noise variables present, (Biau 2012), the method is particularly well suited for consistent real-time human detection from point cloud data. In particular, a method that utilises a pixel level classification through Random Forests (Shotton et al. 2011) is reported to be able to recognise human parts from point cloud data in real-time. Due to the amount of training data required, the training of the classifier in this method is performed using synthetically generated images. The overall approach works well for the intended use conditions: a large space with little occlusion, a fixed camera and a human facing the camera. However, as the above method relies on background subtraction, which necessities a fixed camera position and cannot adapt on-line to a dynamically changing background, it is considered not suited for application in a mobile robot scenario.

In addition to camera based human detection, in robotics, the wide spread use of laser range finders presents additional opportunities for reliable human detection as explained below.

2.1.3 Human Detection and Localisation Using a Laser Range Finder

Due to a number of safety requirements, the mobile robots operating near people have on board a number of laser range finders (LRF), mounted at an appropriate height to enable reliable detection of human legs. Typically, LRF sensors are used both for safety, i.e. detecting people, as well as for map building and localisation, through Simultaneous Localisation and Mapping (SLAM) (Bailey & Durrant-Whyte 2006). In particular, the main safety feature of a LRF safety system is to block any motion of the robot when a detection of an unrecorded object is made within the predefined safety zone, centred at the position of the robot. However, in addition to this basic collision avoidance feature, intended to stop the motion of the robot, additional information in the form of the measured ranges in different directions, is also made available. This information has been utilised by a number of proposed laser based methods aimed at extracting information about the presence of people in the scene from the sensor measurement. Typically, after a detection, the human is presented by a state that encodes the 2D position and the velocity vector (Kluge et al. 2001; Fod et al. 2002; Kleinehagenbrock et al. 2002; Schulz et al. 2003; Topp & Christensen 2005; Carballo et al. 2009). Additionally, leg tracking from laser range data has been investigated in the context of human tracking where a person is presented either as a single augmented state (Cui et al. 2006) or a high level track with two low level leg tracks associated to it (Tsokas & Kyriakopoulos 2010; Arras et al. 2008).

In early approaches (Kluge et al. 2001) the detection of people is achieved by classifiers searching for a moving local minima in the scan. However, such approaches fail to detect any stationary people or disambiguate between people and other moving targets. In later approaches, machine learning, relying on a set of geometric features, has been widely applied to distinguish the people from other moving objects in the scene. Methods for detection of humans in LRF data, based on Geometric rules (Xavier et al. 2005) or a maximum-likelihood estimation to detect moving objects (D Hahnel et al. 2003) have been reported to achieve satisfactory detection rates.

2.1.4 Multimodal methods for human detection

The idea that higher level information, generated by a certain detection module, can be used to provide cues to influence the human detection by other modules and thus improve the overall performance of the system, has been proposed in several works. These include, approaches using information cues from recognition into segmentation (Borenstein & Ullman 2002; Leibe et al. 2005), cues from geometry estimation into object detection (Hoiem et al. 2006; Bastian Leibe et al. 2007), from tracking into detection (Andriluka et al. 2008; Gavrila & Munder 2006; Okuma et al. 2004; Wu & Nevatia 2005) and cues from object semantics into visual odometry (Ess et al. 2008).

Based on the level of abstraction, fusion of information or data can be carried out at different levels of abstracion including: a) data level fusion, b) feature level fusion and c) classifier fusion (Bezdek et al. 1999).

Data level fusion is defined by the Joint Directors of Laboratories (JDL) (White 1991) as a multilevel, multifaceted process dealing with the automatic detection, association, correlation, estimation, and combination of data and information from single and multiple sources. Data fusion offers a number of advantages that involve data authenticity, e.g. detection confidence, reliability, reduction of ambiguity or availability, e.g. extension of spatial and temporal coverage (Hall et al. 1997). Several methods have been proposed to fuse imperfect, correlated, inconsistent data or data in disparate forms as reviewed in Smith & Singh (2006) and more recently in Khaleghi et al. (2013).

Feature level fusion is performed on a higher level than data fusion. In this process the feature sets, extracted from multiple heterogeneous or homogeneous data sources, are combined to create a new feature set that represents the object or the individual. The new fused feature set has higher dimensionality which increases the discriminating power in feature space. This enables feature selection schemes to be employed, as proposed in Ross & Govindarajan (2005) and Raghavendra et al. (2011), to extract a smaller subset of significant features from the larger set of features and improve as a result the fusion process performance. Additionally, some techniques have been proposed to make the features more informative (Gokberk et al. 2005; Jain & Ross 2002).

Classifier fusion is considered the optimal approach for improvement of the classification rates of the individual classifiers. Reported methods include majority voting, Borda count method (van Erp & Schomaker 2000), Dempster-Shaffer Theory (DST) of evidence (Dempster 2008), Bayesian rules and all fusion operator, such as the Ordered Weighted Averaging (OWA) operator (Yager 1988), fuzzy integral (Ralescu & Adams 1980) and so on. Also, the Choquet integral (Choquet 1954), has been successfully applied in multi-criteria decision making (Grabisch 1996) and in a similar way for classifier fusion, where the importance level of every classifier and the interactive effects between different classifiers can be described by a fuzzy measure (Li et al. 2012).

The reported laser and image based human detection methods are either hard constrained or they are using hand tuned thresholds. The work of Zivkovic & Krose (2007) merged the information from a learned leg detector and boosted Haar features, extracted from colour images, to detect human body parts. The approach reported in Schulz (2006) is based on exemplary probabilistic models that are learned from training data of both sensors. Then, the method applies a Rao-Blackwellized particle filter (Doucet, Freitas, et al. 2000) to track contours in the image based on Chamfer matching and infer the position of the person. The main drawbacks of the above method are that it is very sensitive to changes in lighting conditions, and its low scalability due to the substantial computational resources required for the Rao-Blackwellized particle filter.

2.1.5 Tracking of human location

Human position tracking, also referred to as tracking-by-detection (Andriluka et al. 2008) can be considered as equivalent to the more generic problem of the object tracking, surveyed in Yilmaz et al. (2006). However, in human tracking, additional constraints exist which are imposed by the motion model, specific to human motion.

The main idea in tracking-by-detection is based on connecting individual detections, made in sequential time-frames, into trajectories. There are many challenges that still remain in the tracking-by-detection of people, especially when a mobile camera is used. Firstly, a reliable detector is required that can find the target in each time frame. In the general case of tracking by detection, when there is a possibility of errors in detection, e.g. false positive or false negative detections, the tracking should be considered as a stochastic process and approached as a filtering problem. During filtering, the most likely estimate for the true state of the system is formed from a set of noisy observations. While various methods have been proposed for filtering, the Kalman filter (Kalman 1960) and its later variants suitable for non-linear systems (Julier & Uhlmann 2004) remain some of the most popular methods due to their simplicity,low overheads and high efficiency. Secondly, when the number of targets is unknown in advance or variable, it is a challenging task to determine whether any new detections should be associated with an existing target trajectory or a new trajectory should be initiated instead. Finally, when more than one target is tracked, e.g. a group of people scenario, background subtraction is no longer a viable option, due to the increased complexity, leaving tracking-by-detection as the only alternative. The biggest challenge in this case is how to determine the detections-to-targets mapping, a problem known as the data association problem. In such cases, after the detections are associated with a target, standard tracking approaches like the Extended Kalman Filter (EKF) (Gelb 1996), particle filters (Isard & Blake 1998) or a Mean-Shift tracking (Comaniciu et al. 2003) are usually applied. However, since the above tracking methods are based on a first order Markov assumption, i.e. the observation only depends on the immediately previous one, they cannot utilise a longer track history and therefore are prone to errors of tracking. In particular, after a single wrong data association, the Markov assumption based filters have a high risk of drifting away from the correct target. The Multihypothesis Tracking (MHT) (Cox 1993; Reid 1979), a general data association technique that produces compatible joint assignments integrating them over time has been used (Mucientes & Burgard 2006; Taylor & Kleeman 2004; Arras et al. 2008). Similarly, the Joint Probabilistic Data Association Filter (JPDAF) (Fortmann et al. 1983) based approaches reduce the risk of tracking failure by taking into account a longer track history over several time frames. Other multi-target data association techniques such as the global nearest neighbour filter or the track splitting filter, have also been proposed. However, these are considered suboptimal in nature to MHT as they over simplify the multiple tracking problem (Blackman 2004).

Overall, due to the combinatory nature and exponential complexity of the multi-target data association techniques, the multi-target filters are limited to evaluation of only a few timeframes history or a single track over longer time windows (Berclaz et al. 2006; Yan et al. 2006). A method for reduction of the High complexity of MHT has been proposed by Schumitsch et al. (2006), based on Identity Management Kalman Filter (IMKF). However, the overall performance of MHT still suffers from overly simplified statistical assumptions, like uniform distribution of the new tracks in space which can lead to false alarms.

Examples of typical human tracking systems include: 1) the system reported in Zajdel et al. (2005), which is based on a dynamic Bayesian network that can handle multiple targets. A down side of above system is that the targets can only be detected when they are in motion; 2) Luo et al. (2007) uses a tilting laser platform to extract body features that are later fused with the face detection from a camera. The approach is very useful in single person tracking, however, due to the increased complexity of the feature extraction from multiple targets, it has limited applicability for a group of people tracking; 3) Leibe, Schindler, et al. (2008), later extended by Ess et al. (2009), is based on a multi-hypothesis tracking scheme with adaptive model selection. The system can track multiple interacting humans by combining tracking-by-detection with automatic egomoion and scene geometry estimation. However, it still exhibits scalability problems and as a result has not found wider application.

2.2 Human Pose Estimation and 3D Tracking

Similarly to human detection, human pose motion analysis has been the focus of considerable research. This focus has resulted in a vast amount of literature, surveyed analytically in a number of review papers (Aggarwal & Cai 1999; Moeslund & Granum 2001; Poppe 2010; Kolsch et al. 2007; Yilmaz et al. 2006; Sminchisescu 2006).

Human motion analysis can be classified into three main classes: human pose estimation, human pose tracking and activity recognition. As the first two are applicable to the second objective of this work, they are reviewed in more detail below.

The pose estimation targets estimation of the skeletal position given the observations for a single frame (Mikić et al. 2003; Hofmann & Gavrila 2011; Wang et al. 2006; Ioffe & Forsyth 2001; Mori & Malik 2002; Mikić et al. 2003; Hofmann & Gavrila 2011; Wang et al. 2006; Ioffe & Forsyth 2001;Mori & Malik 2002). The estimated human pose is typically used in initialisation of human pose tracking process.

Human pose tracking performs continuous estimation of the human pose parameters over time, i.e. the state variables of the kinematic skeleton model, exploiting the temporal progression of the pose over time (Agarwal & Triggs 2004; Ankur & Bill 2004; H. Sidenbladh, J. Blanck 2002). Although possible, tracking by detection without exploiting the above temporal progression of the pose, is a challenging task that requires a full localisation of each body part in every frame to be able to allow a body pose reconstruction. For example, in Mori et al. (2004), segmentation is based on contour, shape and appearance cues and features based on colour, corners and edges. These are used to detect all body parts, which enables subsequent reconstruction of the body pose configuration. However, relying on the above process for each time frame is very inefficient from a computational resource point of view, as it requires a full search in the space of all human poses and body parts. The segmentation is also much less reliable as the available information from previous time frames is not used in the search but is discarded instead. Therefore, a tracking by detection approach is considered a feasible option only for exceptional circumstances such as the process of initialisation or in recovery after a tracking failure.

Overall, the pose tracking methods are classified into two main groups, i.e. generative and discriminative methods. The generative methods, (e.g. Pavlovic et al. 1999; Urtasun et al. 2006), explicitly model the relationship between pose and observations. This is typically achieved by learning the probability distribution in the space of poses and motion sequences using data from a motion capture system. Then, the tracking consists of generating a set of human pose hypothesis, through the model, which later are compared with the observations and computing a measure representing the degree to which they match. The discriminative methods, e.g. Elgammal & Lee (2004); Howe (2007); Sminchisescu et al. (2005), on the the other hand, do not use an a priori explicit model, instead, they rely on mapping, learned from training data, that directly links the observation space with the pose space. The discriminative methods can deal in a better way with incomplete data, e.g. limb occlusions, in comparison with the generative approaches. Moreover, they can reinitialise in the case of a tracking failure (Forsyth et al. 2005).

The discriminative methods can be divided into learning-based and example-based classes (Kolsch et al. 2007). The example-based human pose discriminative tracking methods, e.g. (Ong et al. 2006; Rogez et al. 2008; Poppe 2007), utilise a database to link the poses to observations, through feature descriptors, with pose configurations in pose space. The descriptors could be based on edges (Sminchisescu et al. 2006; Ong et al. 2006; Ankur & Bill 2006), histogram of oriented gradients (HOG) (Poppe & Poel 2006; Poppe 2007) or silhouettes (Elgammal & Lee 2007; Howe 2007). A drawback of all example-based human pose discriminative tracking methods is that they require large training datasets to achieve an adequate accuracy level (Mori & Malik 2006).

The models deployed in the generative methods vary considerably from author to author. They can be based on connected primitive geometric shapes such as cylinders (Deutscher & Reid 2005) and blobs (Caillette et al. 2008; Caillette & Howard 2006) or on the kinematic tree of the human skeleton (Marr & Nishihara 1978; Balan & Black 2006). Also, depending on the starting point of the generative methods, they can be divided into two main groups: top-down and bottom-up approaches.

In learning-based discriminative methods (Agarwal & Triggs 2004; Agarwal & Triggs 2006; Sminchisescu et al. 2006) a mapping is learned from the observation space to the pose space from the available training data. For example, Rosales & Sclaroff (2000) use a non-linear learning method in which the training data is clustered using forward functions to map an image to pose space as well as inverse mapping functions for pose space to image mapping (Rosales & Sclaroff 2002). Agarwal & Triggs (2004) report a method based on regularised least squares and a relevance vector machine (Tipping 2000) to generate a direct mapping between the calculated histograms of shape contexts (HSC) and pose.

Unlike the example-based discriminative approaches, the learning-based approaches do not need to store and search any large training datasets because of the direct mapping between image space and poses. However, they require extensive training before they can produce meaningful mapping of poses.

The accuracy of the discriminative approaches depends on the similarity between the human poses, present in the training dataset, and the pose configurations that are evaluated at run time (Wang & Yagi 2009). Therefore, the training dataset must be selected specifically to suit the testing scenario. Finally, in comparison with the generative methods, the discriminative approaches are less suited for multi-camera scenarios, especially when there is significant image difference between the training and testing data sets (Balan & Black 2006; Bandouch et al. 2008).

The biggest downside of the learning-based approaches, both generative and discriminative, is the need for a large amount of training data, which is a very resource intensive manual process as the training data should be sufficient to cover the enormous space of all possible body configurations. Recently, this problem has been addressed by using synthetically generated human appearance data, using human body models guided by human motion data captured by MOCAP systems. This training approach has proven to be much more time efficient in comparison with the conventional training data collection process, which relies on sensors collecting real-world data. For example, Shotton et al. (2011) reports using a computer cluster environment and Buys et al. (2013) uses GPU parallel computing to generate human appearance images that are subsequently used for training.

In analysis, there are certain problems that can be identified in the generative or discriminative approach. These include: a) difficulties associated with collecting sufficient amounts of data covering different anthropometric variations such as differences in body height, body mass, muscle performance, etc ; b) resource intensive computation requirements when using a large number of parameters in learning algorithms, and c) difficulties in estimating human motion when external factors not covered in the data collection process are involved.

The image based methods for human pose estimation apply to the problems where only a single monocular image source is available. One of the most important methods for image based human pose estimation relies on a histogram of oriented gradient responses (Dalal & Triggs 2004), which is a variant of Lowe's SIFT feature (Lowe 2004). These methods have formed a foundation on which a broad variety of human pose identification and face recognition methods have been developed.

Chamfer matching and coarse-to-fine searching in a 2D image grid have been applied successfully in the detection of human shapes in images (Gavrila & Philomin 1999). In the above work, a database of shape templates is partitioned into a set of separate clusters based on their dissimilarity to allow detection of multiple body poses. Recursive clustering creates a tree of shape templates. Matching of the observed image to previously stored templates is carried out by traversing the tree structure, following the path of most similarity in chambers. The approach is reported to achieve a near real-time pedestrian detection rate. However, in similarity to all other silhouette-based methods, it has a tendency to suffer from pose configuration ambiguities resulting in a low reported detection rate of only 80%.

As reported, some off-line multi-camera approaches have achieved high accuracy in pose reconstruction (Aguiar et al. 2007). However, the above accuracy is achieved at the expense of substantial computational resources. Such a trade-off is not feasible in a mobile robot scenario due to the limited computation resources.

Depth sensors offer a number of advantages over other more traditional modalities for the purpose of human body pose discovery in service robotics. The advantages, in addition to the depth component, include a relatively high refresh rate, e.g. 60Hz and ability to overlay a colour image with the depth image producing RGB-D data. The related human perception approaches are reviewed below.

Many authors have proposed detection of human body parts directly from depth images or RGB-D data as a pre-processing step to human pose identification (Plagemann et al. 2010; Shotton et al. 2011; Girshick et al. 2011). This approach reduces the negative effects caused by occlusion of human body parts on the human pose identification. Furthermore, as detection is carried out on a body part basis, failure to detect a single body part does not disrupt the continuous operation of the algorithm or cause a tracking failure. After the individual human body parts are identified, the human pose configuration is computed based on the spatial position of the body parts.

However, despite the improved resilience to occlusions, this class of methods is considered to suffer more from false detections. Therefore, it is considered that the body part based approaches could benefit from integration with a filter type pose tracker as this combination introduces temporal and spatial constraints, which minimise the effect of the inconsistent states on the body pose. For example, Buys et al. (2013) combines the ability to remove dynamic backgrounds by the way of the body part pre-processing with enforcement of temporal and skeletal constraints. The above ability to remove the background dynamically makes the algorithm suitable for a mobile robot scenario. A downside of the above approach is that it uses a per-pixel classification algorithm which requires substantial parallel computation, currently achievable only trhough parallel computation.

Another class of generative algorithms is based on the interactive closest point (ICP) algorithm (Besl & McKay 1992). Notable examples of this group include: Articulated ICP (Grest et al. 2005; Plankers & Fua 2003; Demirdjian et al. 2003), Nonrigid ICP (Dirk Hahnel et al. 2003), or the algorithm proposed in Knoop et al. (2006). In general, the ICP based algorithms offer an efficient alternative to the part-based tracking algorithms as they utilise the temporal restraint of the state of human body model over time without the need for additional computation resources. Moreover, kinematic and dynamic pose constraints can easily be enforced in this class of algorithms as a natural part of the model. The main downside of the ICP-based tracking algorithms is the tendency of the kinematic model to converge to the local minima instead to the real pose of the observed person. The above problem impedes the full convergence of the pose model to data, especially when there is external disturbance of the measurements, and eventually results in an unrecoverable human pose tracking failure.

2.3 Human-aware Robot Navigation

In addition to demonstrating an excellent efficiency in carrying out their tasks, the service robots are expected to engage with people and behave in a socially acceptable manner appropriate to the context (Dautenhahn et al. 2005). Human-Robot Interaction (HRI) is a younger sub-domain of robotics, addressing the above aspects. HRI has been a focus of intensive research attention over recent years (Bekey 2008; Feil-Seifer & Mataric 2009; Fong et al. 2003; Goodrich & Schultz 2007; Jensen et al. 2005). The HRI research domain can be divided into three main sub-categories: human-robot proximetrics, human-aware path planning, and robot to human behaviour. The first category studies the reaction of people when near robots, the second focuses on high level socially acceptable navigation planning and the third - on appropriate reactions of a robot in response to human actions (Gómez & Garrido 2013).

In general, because of the high intrinsic complexity, interactions between people are difficult to model directly. In such situations, machine learning methods present an attractive approach to their approximation. If the robot could learn by observing how people interact with one another and apply this knowledge to its interactions with people then it can eventually offer socially compatible HRI. In such a case, the required behavioural models could be updated automatically by the robot without the need of human intervention. However, most machine learning methods, researched to date, do not address the question: “How can learning be achieved for tightly coupled, physical interactions between a learning agent and a human partner as the human counterpart is also part of the learning system and overall dynamics ?" (Shuhei et al. 2012). Learning also requires large amounts of data and substantial computational resources. Most importantly, unless the interaction between a robot and a person is such that it already involves a person taking on a teaching role, having a passive learning agent, instead of active interaction partner may fundamentally change the way people behave in their interactions with the robot. Ultimately, such unnatural human behaviour could result in undesired results like learning of wrong interaction patterns. Therefore, it is considered a better suited approach that the robots should be able to interpret the interaction between two people by observing their interactions and learn from the extracted interaction patterns. With the improved human perception capabilities of the robot, the movements of the people could be identified and recorded by the robot to extract suitable information cues about the human actions thorough analysis of the human tracks. However, the robot will still face challenges in putting the human motion in the correct context of the interaction, e.g. understanding why the particular movement has been made at a particular time.

Overall, despite the significant recent progress, there are some challenges still requiring further research effort. These are considered to be mainly associated with the vulnerability of the human, especially when near a robot. Another challenge is the possibility for the interactions to evolve in an unanticipated manner and the inability of the robot to anticipate the further direction of this progression in order to make appropriate decisions (Goodrich & Schultz 2007).

Human-aware navigation can be considered as the intersection between research on HRI and navigation path planning (Kruse et al. 2013). Navigation path planning (LaValle 2006; Bekey 2008; Elbanhawi & Simic 2014) is a more mature discipline of robotics that investigates the computation of optimal navigation paths given a number of optimisation goals. A basic example of human-aware navigation is the generic framework for human-aware navigation, proposed in Lam et al. (2011), in which the rules for harmonious human-robot coexistence are predefined. However, in this framework only individual people and objects are taken into account but not groups as a whole as very little or no adaptive behaviour of the robot is possible.

A number of methods have been proposed recently to improve the way in which robots navigate around people in specific scenarios as surveyed in Kruse et al. (2013). Some of these include robots approaching people to join conversation groups (Althaus et al. 2004); robots avoiding persons who are blocking the way (Burgard et al. 1999; Thrun et al. 1999), robots adapting their speed of travel when travelling next to a person (Sviestins et al. 2007) or robots changing their velocity near people (Shi et al. 2008).The Human–Aware Motion Planner (HAMP) (Sisbot et al. 2007) goes a step further by considering the safety and reliability of the robot's movement as well as the human's comfort through attempting to keep the robot in front of people and visible to people at all times.

The navigation approach of the robot has to be appropriate to the context of the interactions with people. As reported by Walters et al. (2007), people have different reactions to an approaching robot depending on the context, e.g. whether they are sitting, facing a wall or standing in an open environment. Therefore, sensing the human position and state in advance helps to achieve adherence to the contextually correct set of social rules that are expected from an approaching robot. Actuation scaling, i.e. modification of different parameters of the robots actions based on human reactions, has been reported in a number of works, including: automatic identification of interaction acceptance by the human (Feil-Seifer & Mataric 2011), robot position in the personal space of a human (Laga 2009) and robot behaviour adaptation based on human pose estimation (Svenstrup et al. 2009).

A summary of the main features of human-aware navigation planning includes (Kruse et al. 2013):

- respect personal zones
- respect affordance spaces
- avoid culturally unacceptable behaviour
- avoid erratic motion or noises that cause a distraction
- reduce velocity when approaching a person
- approach from the front for explicit interaction
- modulate the gaze direction of the cameras

Several approaches have been proposed that use off-line learning techniques to build a map of typical destination points for people in the environment (Kanda et al. 2009; Ziebart et al. 2009). These points can be considered later as potential navigation goal destinations and therefore provide support in decision-making for planning of paths optimised to meet the person at the predicted destination point. Another learning-based approach, in which the segments of a roadmap receive different weightings based on the observed human behaviour, has been proposed by (Sehestedt et al. 2010).

Group-aware navigation is considered to be a separate branch of human-aware navigation. Social psychologists have long been interested in studying human behaviour in small groups. Many have emphasised the important characteristics of a group, like interdependence, communication, structure and shared identity of the participants (Levine & Moreland 1998). For example, when people walk together, they subconsciously coordinate their movements with each other, such as what distance to keep from each other and how to turn simultaneously without colliding. Unfortunately, very little research has been carried out regarding this complex interpersonal coordination or the interaction conventions that people follow when moving in groups (Ducourant et al. 2005; Marsh et al. 2006).

A number of works have explored the application of a cost function for human-aware path planning. The frameworks reported by Kirby et al. (2009) and Svenstrup et al. (2010) increase the cost in the regions around objects to allow bigger distances in comparison with what is necessary for collision avoidance. The bigger distances give the people in the scene a sense of security when they observe how the robot moves around objects. Similarly, the methods based on the comfort distance cost (Kirby et al. 2009; Luber et al. 2012; Scandolo & Fraichard 2011; Sisbot et al. 2007), define buffer zones around people. These zones aim to prevent the robot from moving too close to people. The work of Chung et al. (2009) also investigates an increase of the cost function of some regions, especially those that are not directly visible to the sensors of the robot, e.g. around corners. The design rationale is that if the robot cannot make sure that there are no people present in a certain area, it should be avoided altogether, as potentially, it is unsafe to navigate through the unchecked area. After the robot moves initially at a wider trajectory around the obscured area, it has an opportunity to observe the previously occluded space and then to re-plan a navigation path through the area according to the anticipated level of danger to people.

The navigation planner, proposed by Sisbot et al. (2007), prevents the robot from appearing from behind an object when a person is nearby. This approach is similar to the one above avoiding the occluded zone, but it also requires human detection, which is not always feasible in the occluded areas. In the human-inspired reactive and proactive motion planner, proposed in Guzzi et al. (2013), the actions of the robot are planned based on the monitoring of the location of people. Kessler et al. (2004); Scandolo & Fraichard (2011); Sisbot et al. (2007) propose that the robot can approach a person only using zones visible to the human, thus avoiding surprising them. However, the paths that are generated strictly to the above restriction may become very unnatural. This is due to the attempts of the robot to stay visible to people at all the times, which is not typical human behaviour.

In dynamic scenes, it has been proposed that when a person is in motion, provision should be made to guarantee a free space ahead of the person to prevent the robot from obstructing the free human movement.This is achieved by applying a higher cost to the area (Kruse et al. 2010; Scandolo & Fraichard 2011; Ziebart et al. 2009). Similarly, Rios-Martinez et al. (2012) and Scandolo & Fraichard (2011) apply a higher cost function ahead of the person to reserve the space for other purposes, e.g. watching television.

When following a person, the robot has to move in the same direction complying with the natural social motion patterns, which people follow especially in crowded situations and narrow spaces. The approach proposed by Ziebart et al. (2009) is based on iterative planning aimed at finding navigation paths that are free from obstruction in time and space. However, in crowded situations, as the number of iterations is not restricted, the method may run out of time and need to limit either its correctness or reactivity. In comparison, the method proposed by Henry et al. (2010) deals much better with crowded situations by modelling a preference of human paths depending on the density and motion characteristics of the crowd.

Temporal planning relies on a prediction of future human motion to generate appropriate navigation paths for the robot. Predicting human behaviour is not a trivial task for any algorithm. The temporal planner described by Ohki et al. (2010) is based on the wavefront type of algorithms (Mello et al. 2012) with the addition of the time dimension. However, the above planner has no means to deal with scalability issues that result from the added extra dimension in planning. Estimation of a future position of a walking person is proposed by Carton et al. (2013) to generate an optimal path towards the predicted point. The planner proposed by Kushleyev & Likhachev (2009) makes provisions for avoiding the scalability issue by planning only a small number of steps ahead, which, in essence, is only a local planning. Then, the global planning is undertaken by a standard static algorithm. In such a way, the scope of local planning allows temporal planning to remain tractable.

In conclusion, while the above approaches have addressed various aspects of HRI and human-aware navigation, it is considered that the problem of minimisation of the interruptions of social interaction in a group of people has not been addressed sufficiently so far and requires further investigation.

2.4 Discussion

Overall, the key advances related to human detection that enable tackling the above identified challenges in this work for the provision of reliable human sensing and motion analysis are: the HoG detector (Dalal & Triggs 2004), Viola and Jones' face detector (P Viola & Jones 2001) and the sequential Monte Carlo sampling methods for Bayesian filtering (Doucet, Godsill, et al. 2000). Despite the considerable progress in many relevant research areas, as reviewed above, the problem of reliable human detection, localisation, pose estimation and track analysis for improved HRI, remains largely unsolved in non-trivial situations and requires further investigation into the context of service robotics for elderly care. This is mainly due to the complicating factors faced by a service robot in a typical real home environment scenario: variability of human appearance due to clothing and illumination, partial occlusion due to self-articulation, complexity and high dimensionality of the human skeleton, ambiguities and rapidly changing dynamic environments.

The interpretation of human behaviour, which most of the time is generally unplanned and unstructured, is a challenging problem without a single outcome. Social rules depend on the context, which in many cases, because of its subtleties and nuances, is hard to grasp or interpret by an algorithm. Therefore, a strict adherence to the social rules achieved by simple repletion of recorded human actions does not necessarily guarantee that the robot will be perceived to be behaving in a human-friendly way. As reported by Laga (2009), humans when interacting with other humans try to maximise their individual comfort. If a robot mimics precisely the human behaviour it is unlikely that this action will maximise the comfort for its interaction partner which will definitely result in a suboptimal interaction.

Interaction between a robot and a person within a group is an even more challenging problem than interacting with a single individual. There are interactions between the people in the group that are considered important and of high priority. Therefore, it is considered that a service robot, in order to adhere to the established social norms, should also avoid disruption of the ongoing and potential future group interactions. It is necessary that the robot is able to analyse and understand the group member's behaviour to be in a position to decide when and how to intervene in the optimal way.

In conclusion, there is no single human detection and localisation approach in existence today that can produce sufficiently detailed and reliable human observation information that can be relied upon for achieving human friendly and socially acceptable robot behaviour in the context of service robotics for elderly care. Therefore, a clear need for more reliable human information can be identified as an enabling factor for improved HRI. In this work, the above need is addressed by combining all relevant existing sources of human information on-board a service robot. Then, interpretation of the above information, with consideration of prior knowledge of the human motion patterns in regard to group interactions is investigated to allow planning of a navigation path for a service robot that targets social behavioural rule compliance by minimisation of interruption of the interactions between the members of a human group.

3 Conceptual Model of the System

3.1 Acting in Social Domains - challenges associated with insufficient human perception capabilities of the robot

In service robotics, the robots are expected to engage in a variety of interactions with the people they serve. These interactions, to be natural and satisfactory to the users, require close adherence to the established social norms of behaviour. The interactions, an essential part of the daily assisted living, are aimed at the delivery of an efficient and user-friendly assistive service to people in a wide range of unstructured situations, typical for the home environments. Achieving a human-aware HRI requires high availability of precise and dependable information about the human position, motion pattern and pose configuration. Once available, this information can be used by the higher-level control mechanisms of the robot, i.e. task and navigation path planning, to control the robot appropriately in order to deliver the required level of service.

Despite the technological advances, the current service robots fail to engage successfully in intelligent and meaningful interaction with people. While the reasons for such a failure are multi-faceted, it is argued that one of the main reasons is the robot's inability to substitute the real world human-human interactions with compatible ones, from the range of possible contextual interactions. The targeted interactions have to be both acceptable from a human point of view and possibly feasible, cognitively and physically, for the robot. In particular, it is considered that a robot should not blindly repeat the actions of a human from the previous interaction. Firstly, the robot's physical characteristics, such as weight, speed and dexterity, differ substantially from those of a human and it may be inappropriate, difficult or even unsafe for its interaction partner if the robot merely repeats pre-recorded human action sequences. Secondly, as discussed earlier, as human actions in a typical interaction are intended to maximise their individual comfort (Laga 2009), people act in a certain way to achieve this goal while minimising effort. However, there is no clear evidence that their reactions to somebody else's actions are guided by the same principles. More likely, their reactions are guided by their past experiences, emotions, cognition and individual character. If a robot succeeds somehow in imitating by repeating the human actions, typical for the given context, e.g. by modelling somehow the human comfort or just by repeating pre-recorded human's actions, it would certainly fail to reason in a human-like way and react appropriately to the partner's action in the context of the particular interaction. Therefore, it is considered that in these cases a more suitable approach would be if an intelligent robot tries to find simplified substitutes for the human actions and reactions. Certainly the substitutions have to be from an appropriate set of actions that correspond to the context of the particular situation. The above simplification is needed both to make them physically possible for the robot to execute in a safe manner as well as to distance the robot behaviour from the notion of just playing recorded action sequences. At the same time, the set of actions has to be compatible with the relevant social norms of behaviour to be accepted by the people as a natural, sensible and efficient partner. An example of such a substitution of complex human interaction with a simplified one is proposed in Chapter 6: a method for a minimally intrusive navigation path approach to a person who is actively interacting within a group of people.

From a different perspective, it is widely regarded as normal that a typical untrained user of the robot would expect their mechanical helper to be able to interact with them as naturally as a human partner would do. In their view, such an interaction should follow the 'natural to them' social norms of behaviour. If a repository of human interaction patterns, containing a range of pre-recorded alternative interaction steps, is used in the above interaction, then the nonverbal human cues, computed from body language, gestures and interpreted human actions, are considered to be a valuable source of information for selecting the most appropriate option in each interaction step**. Therefore, in conclusion, the availability of timely and accurate information about the people in the scene is considered to be a fundamentally important factor that enables human motion analysis for extraction of relevant nonverbal interaction cues.** However, so far the service robots have not demonstrated sufficient and robust human perception and interaction capabilities. The main focus of this work is set as an investigation of methods that improve the current human perception capabilities of the service robots to enable nonverbal cue extraction and reasoning.

In an illustrative example, a hypothetical autonomous mobile robot is tasked to deliver an object to an individual. However, the person is actively interacting with the rest of the group members at that time. In such circumstances, the robot has to approach the person in the most efficient, natural and unobtrusive manner that is safe for the particular context without interrupting the current or potential future social interactions unnecessarily. However, there are many complicating factors that make the above task challenging. Initially, the robot has to identify the right person among the people present. This task would take little effort for a human. However, it is a much more challenging task for a computational algorithm, which is likely to struggle due to limited cognitive ability. The task is made even more challenging when there is a lack of reliable information about the human, e.g. due to occlusions, poor illumination or sensor limitations like sensor noise or low resolution. In particular, the robot may be unable to recognise people and their behavioural patterns in a complex scene due to the inability to deduct the human body pose configurations from a number of partially occluded human bodies. After a successful identification of the human position, the robot has to plan an appropriate navigation approach to the identified location. In an empty environment, this is a relatively straightforward task, e.g. a direct line of approach is acceptable. However, if the person is moving among other group members in a busy, dynamic environment with rapid human flows, e.g. a corridor of a crowded building, this becomes a very challenging task. It can be argued that the main difficulties in such a scenario stem from the need for the robot to predict the motion of people and to plan an appropriate navigation path in order to minimise disruption to the human interactions.

Moreover, the load carrying capacity of the robot arm is an important enabling factor in executing essential tasks associated with elderly care, e.g. moving objects that a frail person is unable to carry. Since a higher load carrying capacity of the manipulator typically relates to a sturdier and heavier mechanical structure powered by stronger servo drive motors, the increased mass and energy leads to an increased potential risk of serious injury to a human in the event of an accidental collision. Given the increased torque, forces and the longer stopping distances of a heavier arm, short distance proximity sensors alone, e.g. tactile skin or capacitive touch sensors that detect contact with a human at a very short range, are considered an inadequate measure to prevent a collision and potential injuries to a human partner in a close physical interaction. Therefore, the prevention of critical situations which could lead to a potential impact is regarded as an essential factor for inherent human safety characteristics of the robot that can enable closer physical infections. Therefore, in conclusion, accurate, highly available human pose information is necessary for such collision prevention to be possible.

Finally, if the person is interacting with other people, the robot has to take into account these interactions and try to avoid their disruption by crossing the space of active interaction. Current service robots, due to lack of sufficient human perception abilities and algorithms for interpretation of the human motion information, typically use only basic interaction patterns, typically deployed when dealing with objects, e.g. straight line navigation with obstacle avoidance. If a human is detected within the danger zone, it is typical practice that all robot motion is halted until the detection is cleared. The above simplistic approach, typical for industrial robots working in closely guarded cells, is considered to be largely unsuitable in service robotics where interacting closely with people is required. Therefore measures that minimise any unnecessary discomfort, inconvenience or danger to the people in the environment are required to improve the overall user experience. The main challenges identified as in need of addressing to assure more human-friendly service robots are considered below:

Insufficient human perception capabilities of the robots

The first challenge is linked with the lack of adequate information about the identity, the location and pose of the people in the environment.

As discussed earlier, due to variability of human shape, pose and appearance, detecting people reliably by using the robot’s sensors is not a trivial task. In particular, the imperfect characteristics of the sensors of the robot, e.g. limited field of view, limited resolution and noise, make the achieving of a high availability of the detection of people an even more difficult task.

Currently, a typical service robot, e.g. Care-O-Bot3 (IPA (Fraunhofer) 2014), uses a predefined map of the environment with the most likely human positions defined in advance. The final destination for a robot to approach a person is obtained by a selection from the list of possible human location by the robot's operator through the user interface device. The selected human location is then used by the control system to plan a navigation path for the robot to reach the location of the target person. As a last resort, if the location is incorrect or the person gets closer while the robot is moving, the system relies on an emergency stop safety mechanism triggered by an obstacle detection by the laser range finder. In particular, the above safety mechanism is based on a pre-set simple distance threshold between the robot and an obstacle. Such a rudimentary approach is effective in preventing injuries to people and damage in an industrial environment, but it is highly unsuitable for the normal operation of a service robot in a home environment. In particular, as this safety mechanism is triggered by any detection within the safety zone, e.g. a piece of furniture approached by the robot, it rapidly becomes a limiting factor for the normal robot operation. Since a typical service robot currently is unable to reliably distinguish people from objects, all detections within the safety zone are considered as equally dangerous and therefore in order to minimise danger to humans result in a complete and sudden halt to all motion of the robot. Arguably, the above rudimentary safety mechanism is an obstacle to the normal operation of the robot as it prevents it from getting close to people or objects to execute tasks that require a close distance, e.g.an arm manipulation. In the same way, the current safety considerations also prevent a service robot from using its robotic arm for manipulation of objects when there are people in the vicinity of the robot. Similar to the navigation problem above, when the robot control algorithms cannot be certain whether the detections are as a result of an object or a human, the arm manipulation has to be restricted or halted. As a result, the range of circumstances in which the robot can engage in HRI is severely limited. For example, a typical passive safety solution, currently employed by the robot designers in the situations when an object has to be passed between the human and the robot, is to decompose the action into two groups of actions. In the first group are the actions that are safer to execute. In the second group are actions that are deemed unsafe to be executed when a person is near the robot. Then, an attempt is made by the planning algorithms to substitute the actions from the second group with ones that are permitted. For example, instead of handing the object over directly between the robot's gripper and the user, the object is placed on an appropriate surface first, e.g. a tray or a table, to be taken by its intended recipient at a later stage. Using such a passive approach enables the robot to avoid direct physical interaction with the user in some assistive scenarios involving fetch and carry tasks but limits the ability of the robot to engage in a natural for the user HRI.

However, there are situations when it is impossible to avoid direct physical interaction between the robot and the user. These include cooperation tasks when the robot and the person work together to accomplish a single task. Similarly, there is need of close physical interaction in scenarios where the robot is providing active assistance to a frail individual, e.g. standing up assistance. In such situations, a precise control of the arm and the platform of the robot, based on reliable information about the person, is required to guarantee the safety of the close physical interaction and the successful accomplishment of the task.

In conclusion, improved control of the arm and base motion can be achieved with the constant availability of accurate and timely information about the position and pose of the person. Developing appropriate human perception algorithms allowing a mobile robot to detect human presence, location and pose configuration is considered one of the main goals of this work.

Inability of the robot's control algorithms to plan socially compliant navigation paths

A robot taking much longer to finish tasks, because it needlessly selects long distance paths to avoid getting close to people, could be seen as annoying by its users. On the contrary, a robot that always tries to take the shortest path possible to save time and energy regardless of the situation and the discomfort caused to the people could be perceived by its users as aggressive and hostile. Finding the right balance depending on the context of the task is an issue that has not been resolved successfully so far. However, it is believed that if the robot's motion is planned and executed in accordance with the socially accepted norms, an improved efficiency in task execution and increased user acceptance will follow (Koay et al. 2007). Interruption of the social interactions between the members of a human group by a moving robot is against the socially accepted norms and should be avoided in normal circumstances.

Although there are several methods, proposed for human-aware navigation of a mobile robot, as reviewed in section 2.3, little research has been focused on a human friendly navigation aimed at avoiding disruption of the social interactions within a group of people.

3.2 Practical Considerations

The challenges, described above, of achieving close human-robot interaction in service robotics are associated with the lack of sufficient human location and pose information. Most commercially available systems for human motion capture and analysis can provide sufficiently accurate and timely information that can be used for this purpose. However, as discussed earlier, such systems are expensive and require the use of passive visual markers or active devices. Therefore, they are considered inappropriate for application in service robots supporting the assistive daily living of an elderly user. Smart home environments are gaining popularity with the increased availability of lower-cost electronic sensors. However, relying on multiple cameras distributed throughout the house for controlling a robot, e.g. a smart home environment, has high risks of rejection by the end users, as such a data acquisition setup could be perceived as an intrusive and unacceptable invasion of the user's privacy. Therefore, a multiple-camera setup distributed throughout the home is considered to be an unsuitable approach and the focus of this work is placed on research of methods for markerless human detection, localisation and pose estimation that rely purely on sensors co-located on board the mobile robot platform. Indeed, locating the sensors solely on-board the robot allows less intrusive monitoring of the person and is considered to be an important contributing factor for the increased acceptance of the service robots. The rationale of such a design decision is supported by the observation that in real life people assume that they can be observed or monitored only when there is another human present in the room. For this reason, they are also likely to reject any monitoring system that invades their privacy. On the other hand, when only the robot's sensors are used for observation of a user people have the reassurance that their privacy is not invaded when the robot is outside the room. Relying on the above established privacy trust model enables easier user acceptance.

In the experimental platform, i.e. Care-O-Bot 3, which is a typical representative of a contemporary service robot, the sensor data relevant to human detection originates from a number of sources on the robot. These include two laser range finders, a RGB-D sensor, i.e. Microsoft Kinect, and two colour cameras. Their position and vertical field of view (FOV) is illustrated in Figure 3-1 below.

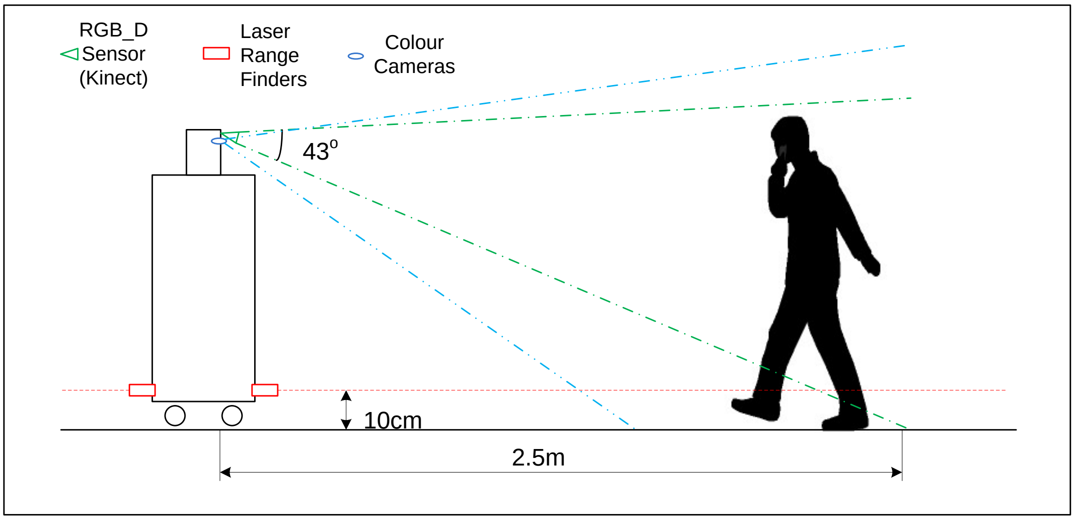


Figure 3-1: Experimental sensor spatial configuration – vertical field of view

The narrow vertical angular field of view of the Kinect sensor, 43°, as depicted in Figure 3-1, makes the detection of a whole human body of a standing average person possible only at distances of more than 2.5m. At shorter distances, the narrow FOV restricts the field of view only to upper human body detection. This limitation is taken into consideration in the design of the human detection methods as described later.

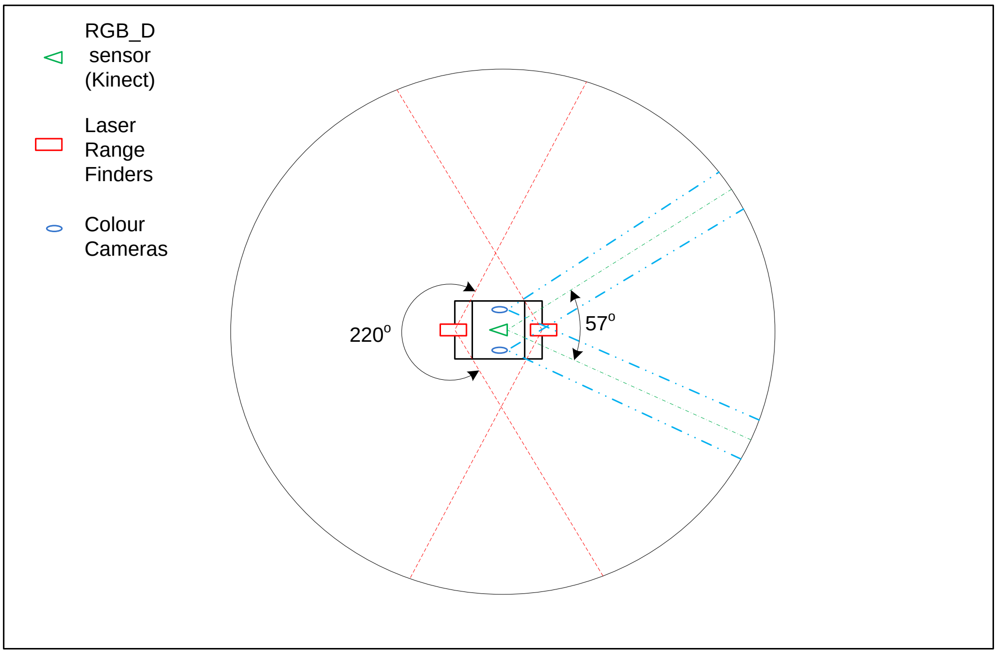


Figure 3-2: Experimental sensor spatial configuration - horizontal field of view

In the horizontal plane, the front and rear mounted laser range finders provide a 360° combined field of view. Such a wide field of view, which exceeds by far the FOV of the RGB-D sensor (57°), and the FOV of colour cameras (60°), guarantees uninterrupted observation of the suurounding space, i.e. no user can approach the robot runobserved . Although it is possible for the robot to turn its other sensors to observe a person, this is a slow and energy inefficient process possible only when the robot is not engaged in another task. Therefore, it is considered that the laser range finders, due to their wide field of view, are the best source of information for constructing an initial hypothesis about the probable presence and the location of people. Later, when appropriate in the context of the currently executed task, the controlling algorithms of the robot use the motion of the robot or the sensor head to confirm or reject the initial hypothesis about the human presence. Such a sequential multimodal detection procedure has the potential both to improve the detection reliability and allow efficient usage of the robot resources.

The proposed framework for human sensing is considered a generic one and universally suitable for a variety of current and future mobile robots. As such, after a short initial training period aimed at discovery of the parameters of the specific robotic platform, the framework for human sensing can be applied to any service robot configuration.

3.3 Research Questions

Based on analysis of the challenges that a service robots face when deployed as a mechanical helper in the unstructured environment of a typical home, the following research questions have been identified as main focus of this work:

- How to improve the perception capabilities of service robots in regard to people while using only information which is available from their existing sensors. Re-using the available sensor information in a better way to detect people is an important foundation which allows higher level algorithms to control the robot behaviour in HRI, as defined by the proposed context awareness in regard to people paradigm in 1.8. This research question addresses the challenges related to occlusion and variance in shape and colour of human clothing.

- How to analyse the perceived human information in order to extract cues regarding inter human interactions. This research question addresses the comprehension layer from the proposed paradigm where important information cues about the people and their actions are extracted from the human detections and made available to the next level for generation of human interaction aware navigation path for the robot. This research question addresses the challenge of correct interpretation of human motion patterns.

- How to use the human interaction cues to plan minimally intrusive navigation paths for a service robot. This research question addresses the operation of the next layer of the paradigm, i.e. the human-aware planning layer. At this layer the information cues about the human location are used together with pre-existing knowledge about the human interaction patterns to compute the robot's navigation path that minimises interruption to human interactions and therefore adheres to the established social norms for moving in crowded environments. This research question addresses the challenge of providing appropriate robot control strategies to enable socially compatible interaction.

The above research questions are addressed through the proposed system architecture that allows interconnecting methods developed for the different layers of the paradigm.

3.4 System Architecture

The system architecture proposed in this work is described in conformance with the ISO/IEC/IEEE 42010:2011 standard (BSI, 2012). The 42010:2011 standard is based upon a conceptual model, or a meta-model, of the terms and the concepts pertaining to Architecture Description. Also, the concepts defined by the above standard, e.g. Architecture Decision, Architecture Description (AD), Architecture Framework, Architecture model, Architecture Rationale, Architecture View, Architecture Viewpoint, Concern, Correspondence, Correspondence Rule, Model Kind, Stakeholder, Concern, System and System of Interest are used throughout this work to describe and refer to elements of the proposed system architecture.

There are several stakeholders of the human perception and minimally intrusive navigation system. These include the elderly end-user of the system, the members of the extended family, care givers, the remote operators, the system builders and the system designers. According to the standard terminology, each of these stakeholders has a number of concerns in regard to the system. For example, one of the elderly users' concerns is to know when he or she is monitored by the robot in order to preserve their privacy. The concern about privacy has led to an architecture decision to use only the onboard sensors of the robot in order to prevent any monitoring of the user when the robot is not in the room. A number of other concerns, related to functionality, usage, performance, resource utilisation, security, cost and feasibility, discussed in the previous sections, form the basis for architecture decisions, presented in Figure 3-3 below.

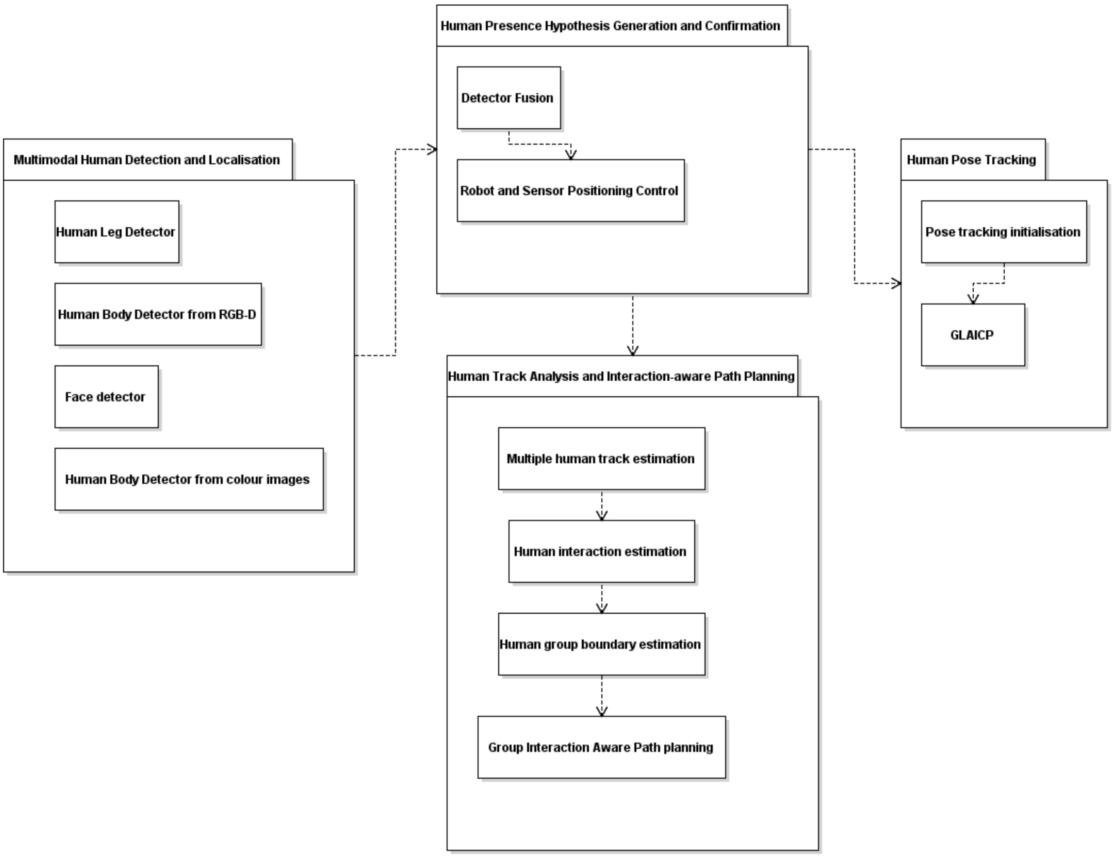

Human Presence Hypothesis Generation and Confirmation
Detector Fusion
Robot and Sensor Positioning Control
Multimodal Human Detection and Localisation
Human Leg Detector
Human Body Detector from RGB-D
Face detector
Human Body Detector from colour images
Human Pose Tracking
Pose tracking initialisation
GLAICP
Human Track Analysis and Interaction-aware Path Planning
Multiple human track estimation
Human interaction estimation
Human group boundary estimation
Group Interaction Aware Path planning

Figure 3-3: Architecture framework described in UML

The System for human perception and minimally intrusive navigation, consists of the following four modules:

- The Multimodal Human Detection and Localisation (MHDL) module, presented in Chapter 4 - this module uses sensor information to establish the number of people and their position on the 2D occupancy map;

- The Human Presence Hypothesis Generation and Confirmation (HPHG) module, described in Chapter 4, fuses information from the sensors to generate a human presence hypothesis;
- The Human Pose Tracking (HPT) module, presented in Chapter 5, uses information from the MHDL module together with the depth data stream to estimate the most likely human pose and to track it in 3D;
- The Human Track Analysis and Interaction-aware Path Planning (HTAIPP) module, presented in Chapter 6 – this module uses all available human detection to estimate the number and position of the most probable tracks.

The four modules continuously exchange information, as indicated by the dependency links in the UML diagram. In particular, once detection of one or more people is made by the MHDL module, the 2D location of the detection is used as the initialisation parameter of the human pose tracking algorithm in the HPT module. Additionally, all detections generated by the HPHG module, both true positive and false positive, are used for the multiple track estimation algorithm in the HTAIPP module. Finally, information from the estimated tracks and human poses, together with context information, e.g. the room layout and the current position of the robot, are used to generate a minimally intrusive navigation path for the robot.

The information generated from all four modules is provided to the high-level control modules of the robot to support the decision-making process of the robot, including task planning, manipulator arm path planning and HRI. The decision-making of the robot, which informed by the human information generated by the proposed system for human perception, is considered outside the scope of this work.

4 Multimodal Human Presence Detection and Localisation

4.1 Background

In order to fulfil its mission of helping people at home, a service robot for elderly care has to be able to gather accurate and reliable information about the presence, the precise location, pose and the identity of the people around it. Several high-level control modules of the robot, e.g. decision-making, task-planning and navigation, need this information in order to control the robot so that it is able to engage in interactions with people efficiently and safely.

Challenges addressed

As discussed, using multiple distributed cameras throughout the home is considered by some as an intrusion of privacy of the person receiving the care. However, relying solely on the robot sensors for human detection and motion analysis comes with a considerable set of challenges. These challenges are associated with the limited range, narrow field of view, low signal to noise ratio and low accuracy of the sensors. Moreover, the constant movement of the mobile robotic platform adds additional challenges as the position in space is not known but can only be estimated,i.e. through the SLAM algorithm, with a given amount of certainty. Additionally, as the sensors are used for other purposes as well, e.g. navigation and manipulation of the robotic platform, they cannot always be optimally directed in the correct spot for optimal human detection. Moreover, when the robot has to move, e.g. to execute a task, the observed field-of-view (FOV) is changing with the robot motion. Although it is possible that the sensors could be mounted on a motion compensating tilt-and-pan platform, which would provide a relatively constant FOV, in most cases this is an impractical solution.

Occlusion is another challenge that has to be tackled when detecting people from the robot position. The relatively low viewpoint of the robot's sensor is considered far inferior to the elevated viewpoint of the sensors used in a typical smart home environment. The varying amounts of missing data in one or more modalities represents a significant challenge for any sensor fusion algorithm as it needs to be able to provide a meaningful output regardless of the wide variations in data availability. Finally, real–world characteristics of the sensors, like the signal to noise ratio and limited resolution, result in a lower level of certainty about the detections.

An approach to human detection and localisation that can cope with the above challenges is proposed in this chapter. The approach is based on a mathematical model that combines several information sources to compute the most probable location of the people within the environment. The underlying key idea is that higher level information, obtained by processing the data originating from a sensor, can be fused dynamically with information originating from other sensor modalities to improve the accuracy and reliability of the combined detector. The improved perception abilities of the robot are accepted as ultimately contributing to more user friendly HRI.

In particular, utilising the above idea, with the increase in the number of sensors, the uncertainty in the human location model is reduced, due to improved data redundancy. The reduced uncertainty results in more precise and reliable human detection and localisation which provides a solid foundation for the high-level control algorithms to engage in closer physical interaction with people.

In an overview of the approach, firstly, a separate human detector extracts useful information cues about the presence of people in each available sensor modality. Then, the information cues are used in combination to build a sufficiently accurate probabilistic spatial dynamic model that provides improved human presence classification output in comparison with the individual classifiers. The above classifier fusion process is based on the method, proposed in this chapter, for adaptive weighing of the output of individual classifiers in the ensemble using historical and dynamic measures. In particular, the variation in the weighting depends on the learned performance of the classifier in a similar context in the past. It also depends on the variance of the tracker assocsiated with each sensor modality, as a measure of its temporal consistency.

In summary, the classifier fusion method, proposed in this chapter, addresses the above identified challenges in relation to the detection of people. It is designed to use only the on-board sensors of the robot and compensate for their deficiencies, i.e. limited FOV of the sensors, occlusion, substantial sensor noise and limited sensor resolution. The key idea applied for fusion of the classifier output is to give the highest importance to the classifiers with the highest probability of an accurate detection.

The main contributions of this chapter are the proposed method for classifier fusion, which is characterised both by a quick response reaction of the ensemble in dynamic environments as well as a high precision of detections when the measurements are consistent over time. Additional contributions are considered in the proposed method for leg detection from laser range finder data, the HISP feature descriptor and the proposed method for human body detection from depth data.

4.2 Problem Definition

The problem investigated in this chapter is the first problem that the robot has to tackle when dealing with people. Firstly, it has to establish the presence or absence of people in the room. Secondly, it has to localise them in 2D Euclidean space with sufficient precision to use this spatial information for higher level control to engage in more meaningful HRI with its users.

More formally the problem addressed in this chapter is stated as:

A human detection algorithm is needed to monitor the surrounding space around the robot by using the input from all available sensors on-board the robot. Upon detecting the presence of a human, the algorithm should make an initial hypothesis about the number of people in the room and their most probable locations. Then, the algorithm should generate commands to position the robot appropriately to observe the identified areas of interest with the maximum number of available sensors. Finally, the algorithm, using the complementary information from all sensors observing the person/persons, is required to make a decision about their precise number and location.

4.3 Subsystem for Human Detection and Localisation

Due to the particular characteristic of the robot's sensors, the best available option for building of an optimal initial hypothesis about the presence of people is to rely on the laser range finder measurements. The main reason for this is the wide field-of-view, typical for this class of sensors. In Care-O-Bot, the experimental platform used for validation, the combined FOV of the front and rear laser range finders covers essentially the whole 360° space around the robot as illustrated in Figure 3-2.

After the initial hypothesis about a human presence is made, the control algorithms of the robot have to make the necessary adjustments to its pose and position to confirm or reject the hypothesis. The adjustments include turning the platform base or panning or tilting the sensor head to allow the maximum number of all available sensors to observe each region of interest. The motion of the robot is an expensive process in terms of time and energy but it is necessitated by the relatively narrow field-of-view of RGB-D and other camera sensors. Finally, after an appropriate position is achieved, the robot, utilising information from all appropriate sensors to observe the particular region of interest, confirms or rejects the human presence hypothesis in each region. This is achieved by using the proposed classifier fusion method, described in detail in 4.7. The whole process of building an initial hypothesis about human presence and location, its confirmation and finally its elicitation, is illustrated conceptually in Figure 4-1.

Typically, in mobile robotics, a 2D grid, referred to as an occupancy map, is used for the purpose of recording the relative position of the robot against other objects in the environment. The occupancy map is created and continuously updated by the SLAM (Durrant-whyte & Bailey 2006) algorithm, running on the robot's computer. It is proposed that the occupancy map should also used for the recording of any occurences of a personbeing detected and to be made available to other modules of the robot, e.g. navigation path planning, decision making, as well as for visualisation of the human position to a remote operator.

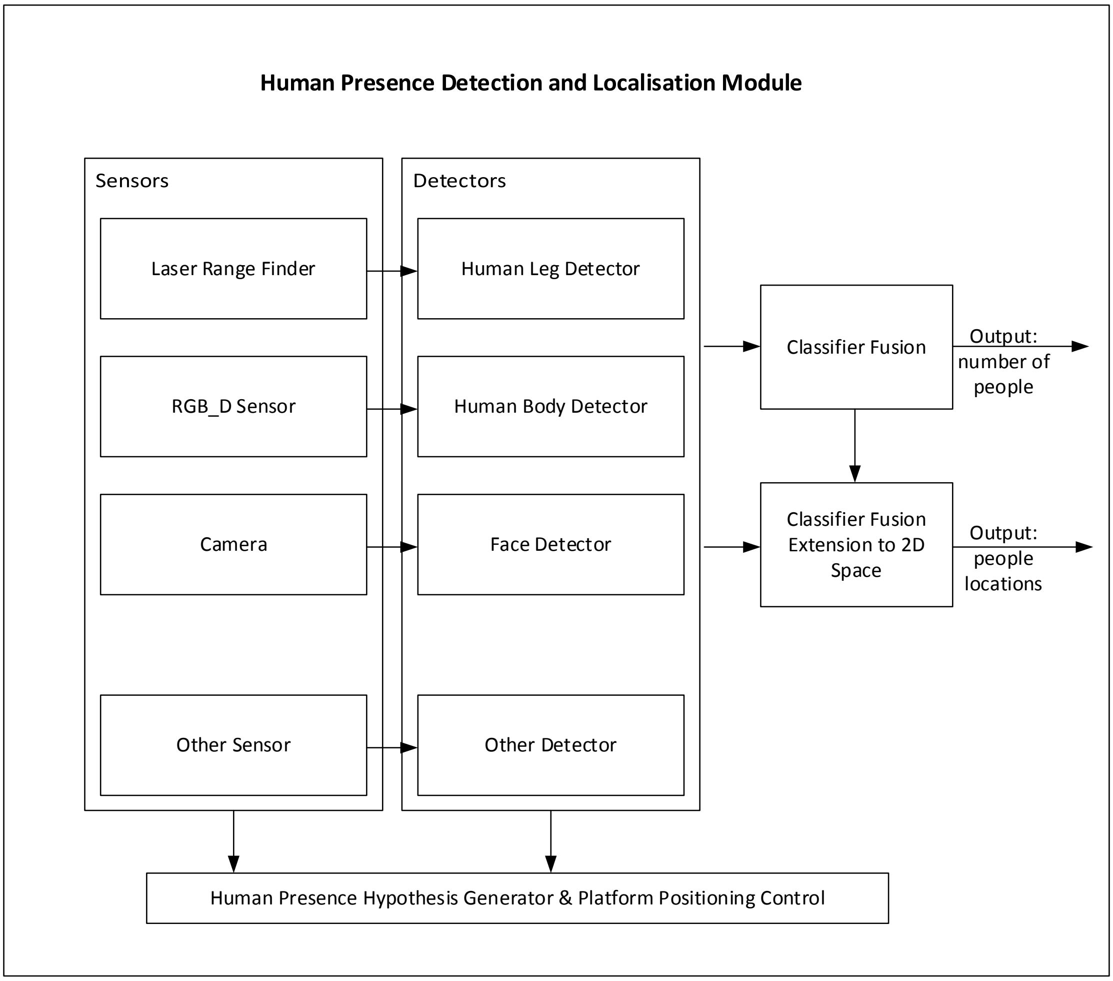


Figure 4-1: The conceptual model of the Human Presence and Localisation Subsystem

4.4 Human Detection and Localisation from a Laser Range Finder

Laser range finders (LRF) offer several attractive sensor properties that make them well suited for the purpose of human detection by a mobile robot. These properties include a very wide field of view (FOV), covering almost 360° of space, long detection range, high precision and relatively low noise. However, the down-side of this class of sensors is that they only allow a single scan plane. This together with the sequential nature of scanning results in relatively low scan rates, typically in the range of 1-2 Hz.

Since a single scan plane severely limits the amount of useful information, which can be gathered from the environment about the people, it might seem beneficial to extend the LRF detection capabilities to 3D space by tilting the sensor. Although this is possible, e.g. by mounting the sensor on a tilt unit, enabling scans in multiple planes, the time for a full 3D scan will increase proportionally to the number of scan planes. The above solution would make the sensor prohibitively slow for the detection of any moving object. For example, if an LRF with 2Hz scan rate is mounted on a tilt unit to cover a 90° vertical segment it would require 60 sec for a single scan of the whole space at a resolution of 0.5°. Such a slow scan rate is only suitable for static scenes, like those in architecture.

There are highly specialised LRF scanners developed especially for autonomous road vehicles that can tackle the above problem. They can do simultaneous scanning in multiple planes, e.g. 64 planes in Velodyne HDL-64 (Velodyne Lidar 2010), but their heavy weight and high cost makes them unsuitable for domestic service robot applications. **Therefore, as a conclusion, it is considered that the optimal approach for human detection in mobile robotics is to use the wide area scan capabilities of the LRF to make only an initial hypothesis about the presence and location of people. Subsequently the hypothesis can be confirmed and improved by using other sensors with more suitable characteristics, e.g. faster scanning speeds and a field of view covering the whole body.**

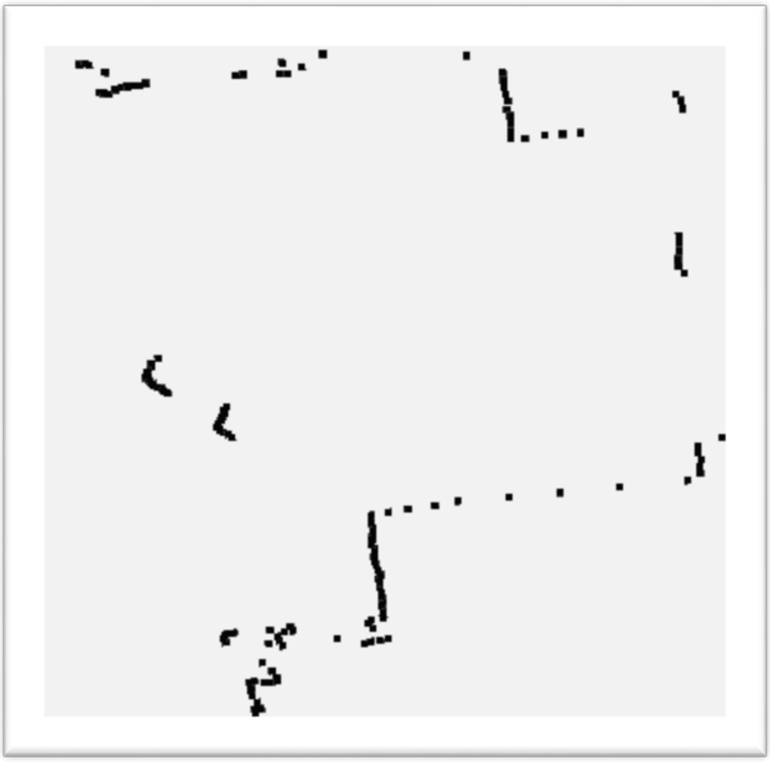


Figure 4-2: A typical scan image from a laser range finder from a service robot in a home environment. *Note: the chevrons in the middle represent detections of human legs*

In regard to generating the initial human presence hypothesis, a fixed position LRF can acquire only very limited information about the people in the scene due to its characteristics. In particular, the detections are the points where the laser range detector beams hit the surface of the detected human body. As an illustration, a typical LRF, fixed at a height of 10 cm above the ground for detection of obstacles, after observing a person, outputs two ellipsis, one for each leg, representing the cross section of the measurement plane of the sensor and the human legs, as shown in Figure 4-2 above. If the person is moving, the shape of human legs in the scan becomes distorted due to the sequential scanning of the LRFs. This distortion leads to additional difficulties in human detection. However, due to the relation between the distortion and the speed of motion of the detected object it is considered that it can be used to extract an advantage, i.e. extra information about the human motion parameters, as described below.

Overall, the proposed method consists of two main phases: pre-processing phase and human detection hypothesis generation phase. In the pre-processing phase the data resulting from the sensors 2D point cloud is parsed to remove the unwanted artefacts from the sensors and reverse the distortion effect. In the hypothesis generation phase, geometric features are extracted from the 2D point cloud and the human candidates are identified using classification based on a machine learning method.

4.4.1 Analysis of the Characteristics of a Laser Range Finder in regard to Detection of People

The aforementioned attractive properties of the laser range finders, i.e. relatively high accuracy, lower level of noise and improved robustness against illumination changes and vibration, make them a preferred choice for a number of sensing applications. When compared with other sensors it can be found that only triangulation based approaches offer higher accuracy. However, despite their relative superiority, the laser range finders still exhibit some deficiencies that make human detection challenging. These deficiencies include: missed detections, false positive detections, ability to scan typically only a single plane and perhaps most importantly very low scan rates.

It is considered that the increased level of uncertainty, resulting from the above sensor deficiencies, can be mitigated to a large extent by deployment of a probabilistic inference, e.g. a filter, that combines the observations from a number of consecutive frames to increase its certainty about the presence and the position of people. Filtering out of the outlier detections is possible by using the likelihood probabilistic sensor model. This model is represented by the probability distribution function of the predicted detection being given a number of known parameters, e.g. the shape of the leg, estimated speed and direction, and the distance from the sensor.

The likelihood model reflects the specifics of the operation of the sensor and underlying physical characteristics of the time of flight range measurements. There are cases, even within the working range of the sensor, when the reflectivity of the material is insufficient for a reliable detection, due to the infrared absorption properties of the object. The insufficient return signal may result in a failure of the sensor to measure the distance. The sensor interprets the lack of a return signal as an out-of-range measurement and erroneously returns the maximum possible reading. This problem manifests itself as a gap in the scan line and represents a challenge for the detection of a human presence, as it can result in a false negative detection of the person. Therefore, the proposed probabilistic model takes into account the probability of missed detections by the LRF and eliminates, by inference, the false maximum range readings. The result is then used in the next step by a trained binary classifier to decide if the observed scan image is a human leg or not.

In other cases, it is also possible that erroneous detections are made by the sensor when there is no artefact in that particular position. These, for example, could be a result of noise in the amplifying module of the photo receiver or another unknown reason. The above false positive detection could result in a wrong detection of a person.

Finally, it is also possible that a temporary occlusion in the way of the laser beam could cause a false detection. Although this is an external factor to the sensor, because of the risk of false human detection, it is still accounted for by the probabilistic measurement model.

The probabilistic method used in this work for human leg detection is based on the key principle that erroneous measurements, described above, can be filtered out by repeatedly observing the same area by using consecutive measurements. This temporal process is straightforward in a static environment, where the averaging of the image over several timeframes results in the same image with the required accuracy. However, in a dynamic environment, e.g. moving people, Bayesian filtering is required to mitigate the adverse effects of the false positive and the false negative readings while not interfering with the correct scan lines of the moving people.

4.4.2 Probabilistic Measurement Model

A probabilistic measurement model enables successful stochastic inference about the interpretation of the detections in a dynamic environment, which subsequently leads to an improved ability of the robot to make appropriate decisions using low confidence cues. This is especially critical for the dynamic parts of a scene, e.g. moving people or objects, as they introduce a substantial level of uncertainty that cannot be tackled directly through simple averaging of the measurements of multiple subsequent scans.

For example, knowing the position of a human presence even with only a small degree of certainty is much better, from the robot decision-making point of view, than not having any information about the location of the human. The reason for this is that even the slightest cue with probability above a random guess, about the presence of people, can be employed in combination with other similar cues, originating from different modalities. The combined result can have a higher probability of a correct detection than any of the modalities taken separately. Such a multiplier effect, achieved by stochastic inference, plays an important role in the proposed method for enabling the robot to make a better informed decision and contributes to increasing the passive safety of human-robot coexistence.

This section introduces a probabilistic sensor model for the Laser Range Finder (LRF) measurements. The probabilistic measurement model is deployed for removal of unwanted artefacts in the detection of humans by LRF. It is important that all uncertainties affecting its real world operation of the LRF are captured by the model, e.g. noise, missed detections and not-modelled object dynamics, as these will affect the inference process. The model consists of two parts: a) a part modelling of the distortion of the detected shape of moving objects and b) a part modelling of the characteristics of the sensor.

The dynamic environment nature in which the robot operates introduces an extra layer of complexity in comparison to the detection of static environments. Due to the limited scan rate of the sensor and the time needed for the scanner beam to sequentially scan different points on the surface of the object, the measurements are taken essentially at different times. As the moving object has changed its position between the two timeframes, the resulting detected shape of a moving rigid object becomes distorted. The distortion depends on the direction of the movement of the object, the motion of the robot and the scanning rotation motion of the LRF. This presents a challenge both in the correct identification of the object as well as in the precise estimation of its correct position. The basic underlying mechanism of the above distortion effect in sequential scanning is illustrated in Figure 4-3 below.

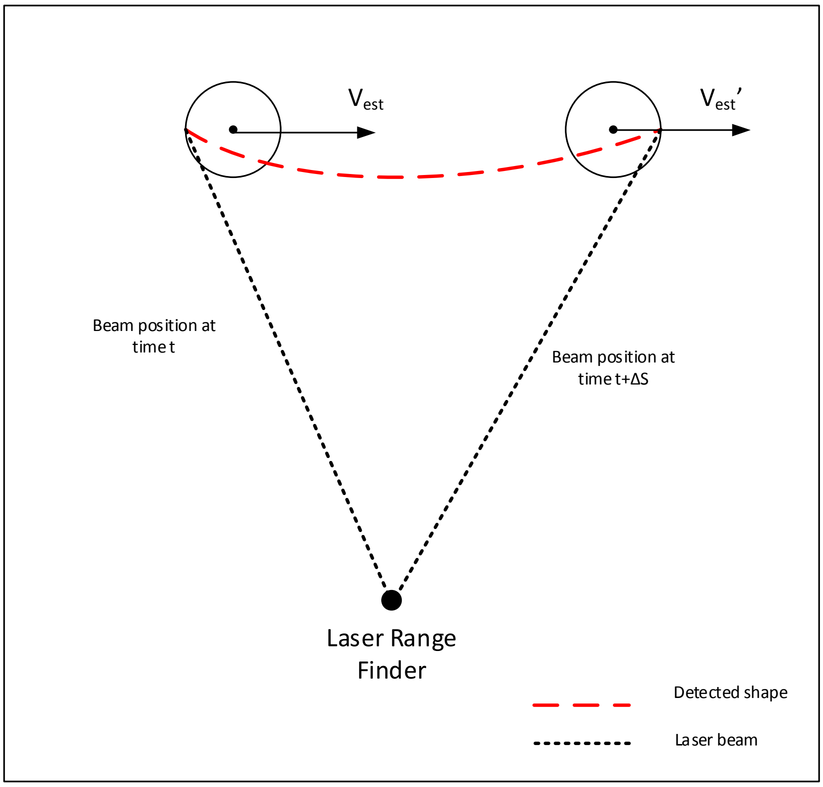


Figure 4-3: The distortion of the scan image caused by time delay between laser beams detecting points at different times

For example, when the detected human leg is moving in the same direction as the scanning rotation of the sensor, as is the case in the figure, the measured width of the detected leg segment increases due to the different position of the leg at the times when the first and last points of its surface are detected. In an extreme case, when the object is moving with the same speed or faster than the laser scanning beam, it becomes impossible to detect the whole object using the sensor as the laser beam never catches up with the object. Similarly, distortion occurs to the detected image when the detected human leg is moving in the opposite direction to the rotation of the laser beam. In such cases, the effect manifests as shortening of the measured object width. In the most extreme case, if the speed of the scanned object becomes very high and is moving in the opposite direction to the laser rotation, the scanned image collapses to less than the scan resolution of the LRF. In the above case, the detection will be either represented by a single detection point or will be missed altogether.

The distortion effect of a moving human leg, illustrated in Figure 4-4 below, is modelled mathematically by the leg distortion appearance model (DAM), which is presented in Appendix D.

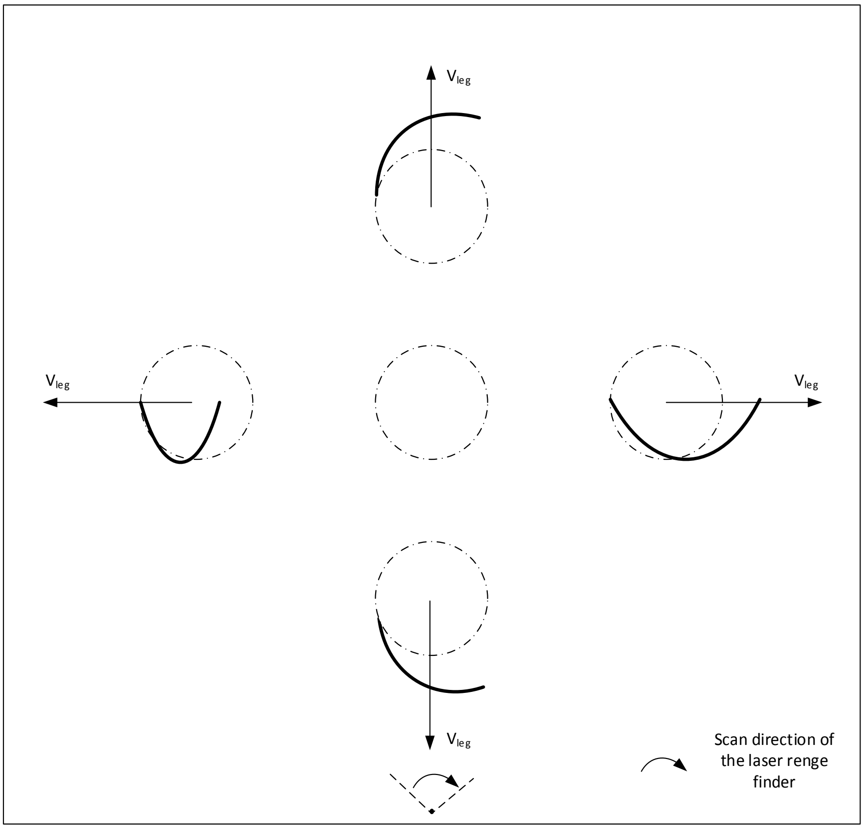


Figure 4-4: Illustration of the shape distortion of a moving leg in different directions

Discussion

The mathematical model of the distortion assumes that the speed of the leg is known precisely. However, in reality this is not the case. The speed of the leg is only estimated by filtering and as a result it is given as a mean and a variance. As expected, without the distortion effect the image of the human leg would exhibit the characteristics of a rigid body, i.e. all its points move in the same way while preserving the fixed distances between them. Additionally, the certainty in the position of all detected points is the same for all of them, and equal to the certainty for the body centre, which can be used as the reference point of the position. However, due to the above effect, the detected image of a human leg can be characterised as non-rigid body, i.e. it is distorted as the points detected later in the scan will have more time to travel further in comparison with those detected earlier. Due to the uncertainty in the estimated speed of the leg, described above, the distortion will also result in different uncertainty in the original position of various points from the image, which will depend on the time of their scanning. In particular, the above uncertainty is directly related to the time delay of the detection of subsequent points in comparison to the first detected point from the leg and the certainty to which the leg motion speed is known. The point from the leg that are detected first, i.e. when the laser beam first hits the surface of the human leg, will directly correspond to its real position. However, due to using only estimated speed, the uncertainty of later points will increase with the time delay measured from the first detected point.

Therefore, for tracking purposes, after the cluster of points has been identified as an image of a human leg, the coordinates of the first detected point are used as the most representative reference position for the leg. Intuitively, the first detected point could be used for determination of the position of the centre of the human based on the assumption that the radius of the leg is known and the direction of the laser beam on the first detection lies on the tangent of the circle representing the leg. However, in reality, because of the limited angular resolution of the sensor, it is unlikely that the very first point from the leg is detected. In fact, the lateral resolution of the sensor is proportional to the distance to the object and inversely proportional to the angular resolution, d, of the laser scanner:

$$d = \frac{2\pi R\theta}{360}, \tag{4.1}$$

where R is the distance to the detected object, θ is the angular resolution

For a typical scanner with 0.5° angular resolution the axial resolution at a distance of 10m is 8.7 cm. As the position of the first detection point is uniformly distributed within a single lateral resolution step, it is considered that the error in estimation of the centre of the human leg would be unacceptably high for the purpose of localising the person.

Therefore, a method is proposed below for reversing the distortion effect of the sequential scanning of a moving object. After the distortion is removed, a circle is fitted to the resulting image. Finally, the circle is used to compute the coordinates of the centre of the leg. The whole process is illustrated in Figure 4-5 below.

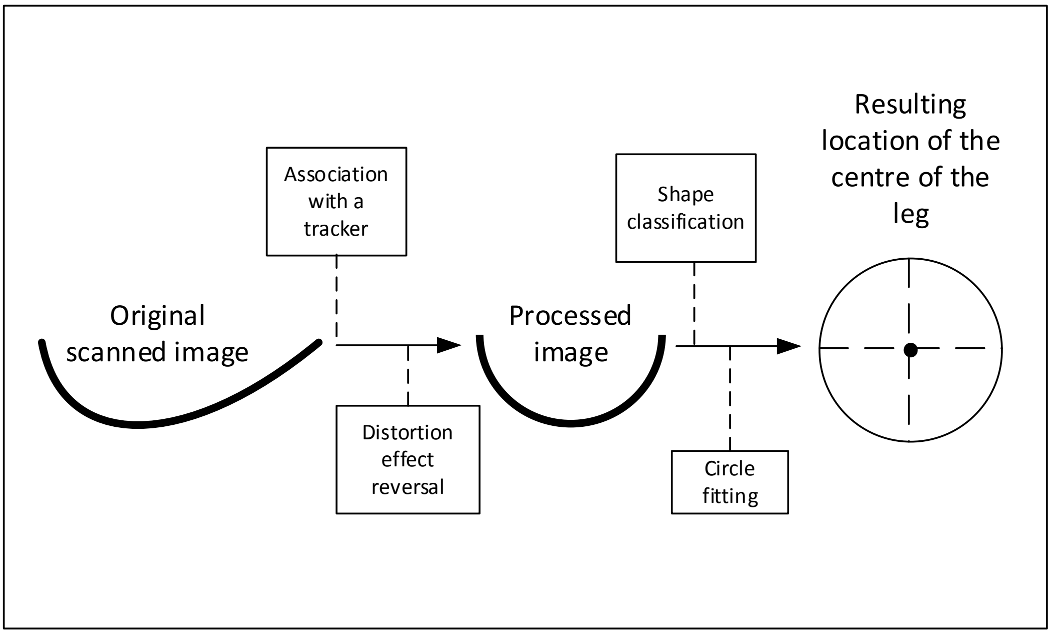

Original scanned image
Association with a tracker
Distortion effect reversal
Processed image
Shape classification
Circle fitting
Resulting location of the centre of the leg

Figure 4-5: Overview of the laser scanned images processing

In overview, after the successful acquisition of the point set from the image cluster, the method uses the estimated speed, as reported by the associated tracker to restore the original shape of the target as explained in 4.4.5. Then the resulting image is used in the classification step, explained in 4.4.7. Finally, if the object is positively classified as a human leg, a circle is fitted to the image, as explained in Appendix B. The centre coordinates of the circle are then returned by the method to be used as the reference position of the detected human leg by other algorithms of the robot.

An additional benefit from the distortion effect can be obtained from the dependence of the shape distortion effect on the object's speed. The benefits stem from solving the inverse problem – obtaining an estimation of the speed of the object with a previously known shape from the distortion of this object in the image scan. This information is particularly useful for new targets, as there is no tracker associated with them to provide a speed estimate. Initiation of the tracker with a speed estimate allows it to converge much faster in comparison with the case when there is no such estimate.

4.4.3 Clustering of the scan data

The points measured in a laser scan are assigned into clusters which, after reversing the distortion, allow computation of the geometric features of the clusters, as described in 4.4.6, to be carried out. The clustering is based on the jump distance between sequential points.

However, as described in section 4.4.2, it is probable that a false positive measurement might be reported due to a sensor error in the laser range finder. Therefore, a new clustering algorithm is proposed to prevent the above measurement errors from ending the current cluster and starting a new one. The method is based on the principle that the probability for a series of successive wrong measurements is very small, as described in detail in 6.7. In particular, when the distance between two sequential points is bigger than the predefined threshold distance, d_{tresh}, event referred to as a jump distance, a hypothesis for the start of a new segment is made. If most of the subsequent points after the jump are positioned away from the last cluster of points then the hypothesis for a new segment is confirmed and the algorithm initiates a new segment. Otherwise, if some of the points after the jump are relatively closer to the last points of the cluster the hypothesis of a new segment is rejected. In this case the detection that is positioned away from the cluster is ignored as an outlier and the points after the jump are added to the old segment.

4.4.4 Identification of new targets

After the leg segments are formed, using the algorithm presented in 4.4.3, they are mapped to their predecessors from the previous time frame using information from the associated tracker. Alternatively, if such a mapping cannot be established, they are finally labelled as new targets.

The mapping between a cluster and its predecessor is calculated through the predicted position of the tracked targets. For this purpose, the probability distributions given by the measurement model, explained in 4.4.2, are used. In particular, if the cluster fits into a previously predicted position sufficiently well then it is considered to be an ancestor to the previous cluster. Otherwise, if no such fit can be found the cluster is regarded as a newly detected target and a new tracking filter is initiated and associated with it.

4.4.5 Compensation for the distortion effect in sequential scanning

After the clusters are associated with existing targets, their estimated speed is inferred by updating the Kalman filter of the particular tracker. In the case of a new target, the speed is estimated using the distortion of the detected shape, i.e. by applying equations (9.45) and (9.53).

Next, using the estimated above speed of the leg, the clusters are processed to reverse the effect of the distortion effect. In particular, this is achieved by translating each node in the clusters by the reversed distortion distances, which are given by equations (9.44) and (9.53).

The resulting shapes are then further processed to classify them into two classes, i.e. legs and non-leg shapes using a machine learning method as described below.

4.4.6 Geometric Features Definition

The probabilistic measurement model, described in 4.4.2, is used for clustering the measured points. After a scan is complete, points with a maximum uncertainty are removed, as they will be either background or errors. The rest of the detections are clustered, using a jump distance partitioning with a pre-defined fixed threshold as described earlier in 4.4.3. After clustering, the smallest segments, i.e. those containing less than a minimal pre-defined number of points, are discarded, as they are considered unsuitable for human detection. The rest of the segments are processed to compensate for the distortion effect in scanning, as discussed in 4.4.2, and subsequently a number of geometric features are computed as discussed in detail in Appendix B. Finally, using the computed features, a classification is carried out as discussed in the next section 4.4.7.

4.4.7 Leg Detection through Binary Classification

In the second phase of human hypothesis generation, a Random Forest (RF) binary classifier (Breiman 2001) is used for the classification of clusters computed from the LRF data. The RF is selected because of its superior speed, both in training and evaluation, as well as its inherent multi-class nature. Moreover, RF has proven to be more noise tolerant in comparison with other existing classification methods.

At the same time, RF, similar to other discriminative learning methods, suffers from several disadvantages that have also been taken into consideration. The main disadvantage is the requirement of very large training datasets to achieve the full potential of the classifier (Caruana et al. 2008). The above disadvantage of RF has been tackled successfully by the proposed procedure for automatic collection of training data as discussed below.

Training of the classifier

A large training dataset containing both positively and negatively labelled data have been automatically generated using the fact that the negative training dataset originates from static objects that are present in the environment. In particular, the following data collection procedure is used for training:

Initially, the robot platform is driven around the environment while no people are present to collect data using its sensors. From the resulting LRF detections, the geometric features are computed and stored as a negative labelled data set. Also, the locations of the static objects in the environment are registered on the occupancy map as obstacles for later background subtraction. Detecting the static objects is simplified by the fact that due to the random nature of the sensor's noise it is extremely unlikely that a false positive or a false negative reading will be made continuously in the same cell over a number of consecutive frames, as described in 6.7. In practice, after a sufficient number of measurements are accumulated over time, the cells with a true positive detection score higher than a predefined threshold rate, R_{tr}, are marked by the algorithm as an obstacle and avoided by the navigation planning algorithm proposed in Chapter 6.

In the second step, by driving the mobile platform in the environment when people are present, a large dataset of scans is collected. Then, using background subtraction with information from the occupancy map, the scans are processed to remove the obstacles. The output of the above procedure is a dataset of human detections, which is used in the positive training of the classifier.

Finally, using the stored positive and negative datasets, clustering of the measurement data is performed to filter out any unsuitable data segments. The geometric features of each segment from the datasets are computed, as described in 4.4.6, and fed subsequently into the RF classifier for training.

Run-time classification

In run time, the process for each new scan image starts with the pre-processing phase. As described, it includes clustering, rejection of any unsuitable clusters, reversing the effects of scan image distortion and generation of the geometric features. After the pre-processing phase, the features are computed and the resulting data vector is fed into the pre-trained RF classifier for classification. It returns a positive or a negative classification class label, used to identify the leg artefacts in the image. The centre of the circle around each positively classified leg cluster, computed as described in 4.4.2, represents the reference position of the detected leg which is used to update the associated leg tracker.

After the classification of the clusters, a proximity search is performed over space for the position of legs using the Euclidean distance criteria, described in 4.4.3. The result of the search is a set of pair of legs identifying a human candidate. The two centres of each pair of legs are connected with a line and the middle point of the line segment between the centres is computed and used as the reference point for the human candidate.

In the next step of human detection and localisation process, i.e. classifier fusion, each of the above human candidates is confirmed or rejected using additional human information cues, resulting from detections from different sensor modalities, described in the following sections.

4.5 Human Detection and Localisation from an RGB-D Sensor

RGB-D sensors are becoming common in mobile robots where they are used for the purpose of object discovery and shape recognition. This section investigates methods for re-using the signal generated by an RGB-D sensor for human body detection. The general approach for body part identification is based on a machine learning classification approach that relies on novel local depth feature signatures to identify human body parts. The result from the classification is then used in 4.7 as one of the complimentary cues, contributing to the overall human presence and localisation decision.

4.5.1 Analysis of the RGB-D sensor properties with regard to human sensing from a mobile robot

The low cost RGB-D sensors aimed at the home gaming market, e.g. Microsoft Kinect based on the Primesence technology, have found an alternative application in mobile robotic applications due to their relatively high performance to price ratio. Typically in these applications, an RGB-D sensor replaces a much higher cost Time-of-Flight (TOF) depth sensor. In some cases, when appropriate algorithms are used, it can even replace a lower-end laser range finder. Such replacement is made possible through the development of intelligent algorithms compensating for the severe deficiencies of RGB-D sensors, as analysed below.

Modern RGB-D sensors are typically based on the structured light approach. According to the inventors of the sensor (Freeman et al. 2013), the operation principle is based on a projection of an infrared speckle pattern, produced by splitting of a single infrared beam by a diffraction grating. The reflection of the pattern is then captured by an on-board infrared camera and correlated on a part-by-part basis to a series of reference patterns, pre-stored in the sensor. In particular, when the pattern is projected onto an object that is at a different distance than that in the stored reference plane, the position of the region in the pattern corresponding to the object is shifted. The extent of the shift is proportional to the depth disparity that is in the direction of the line between the IR light source and the centre of the IR camera. After the above shift is measured for each speckle pair, the disparity image is generated for the whole image. From the disparity image, the distance to the object is computed and stored in the resulting depth image with VGA resolution (640 x 480 pixels). Subsequently, the raw distances are normalised and quantised between the values of zero and 2047 for each pixel of the image. Finally, the depth values are returned to the user as a stream of 11-bit integers at a frame rate of 30Hz.

Due to the simplicity of the above design and the economy of scale, it has been possible to achieve competitive technical characteristics at a very low cost to the end-user. Indeed, the sensor offers competitive refresh rate and x-y resolution properties, in comparison with other much more expensive depth sensors on the market. The competitive characteristics and the low cost of the sensor have resulted in considerable research interst and a number of successful applications of the sensor in robotic applications, e.g. 3D indoor mapping, Simultaneous Localisation and Mapping (SLAM) and object shape recognition. The ubiquity of the sensor has inspired the idea of re-using its information for the purpose of human detection and human pose recognition as described below.

In parallel with the above beneficial characteristics, the sensor exhibits significant deficiencies, which originate from its design. The deficiencies consist mainly of the significant non-linear random error in the depth component and the deterioration of the depth resolution at distances beyond the designed primary use zone, i.e. the 2.5-3m required for the home gaming market. In addition to the above deficiencies, the resulting point cloud has highly inhomogeneous density, which prevents most of the standard algorithms for point cloud processing from normal operation. The above factors contribute to the challenges of human detection based on a low cost RGB-D sensor. Therefore, it is considered that new methods, compensating for the sensor deficiencies are required to make the sensor feasible for human detection, especially at longer distances. Therefore, as this observation is in line with the objective and the research questions of this work, a further investigation is carried out as described below.

RGB-D Sensor Deficiency Analysis in relation to Human Sensing

Firstly, the accuracy of the depth data deteriorates rapidly as the objects get further away from the sensor. This can be demonstrated by a simple experiment, aimed to evaluate the distribution of the measurement error as a function of the measured distance. In the experiment, several depth images are acquired by the sensor facing a flat surface, i.e. a wall, at a range of different angles and distances. These have been compared with a parametric model of the wall position. Then, the depth errors of the measurements in the point cloud are calculated as the difference in measured distance and the computed distance from the model along the same beam. Finally, the data is aggregated per distance ranges to compute the error histograms, shown in Figure 4-6. The histograms illustrate the depth error distribution at different distances. The findings confirm the highly nonlinear nature of the sensor.

In particular, it is useful for the development of human sensing methods to take into consideration the spreading of the measurements with the increase of the distance. For estimation of the error in run-time, a model of the standard deviation error of the sensor has been built from the experimental data. The model is visualised in the Figure 4-7 below. In particular, an exponential function, i.e. $F(z) = ae^{\beta z}$ is used to model the increase of the error of the measurement error with the increase of distance. The best fit, resulting in a quality of fit defined by: R-square=0.99, is achieved at values of the coefficients $a = 1.98$ and $= 0.85$.

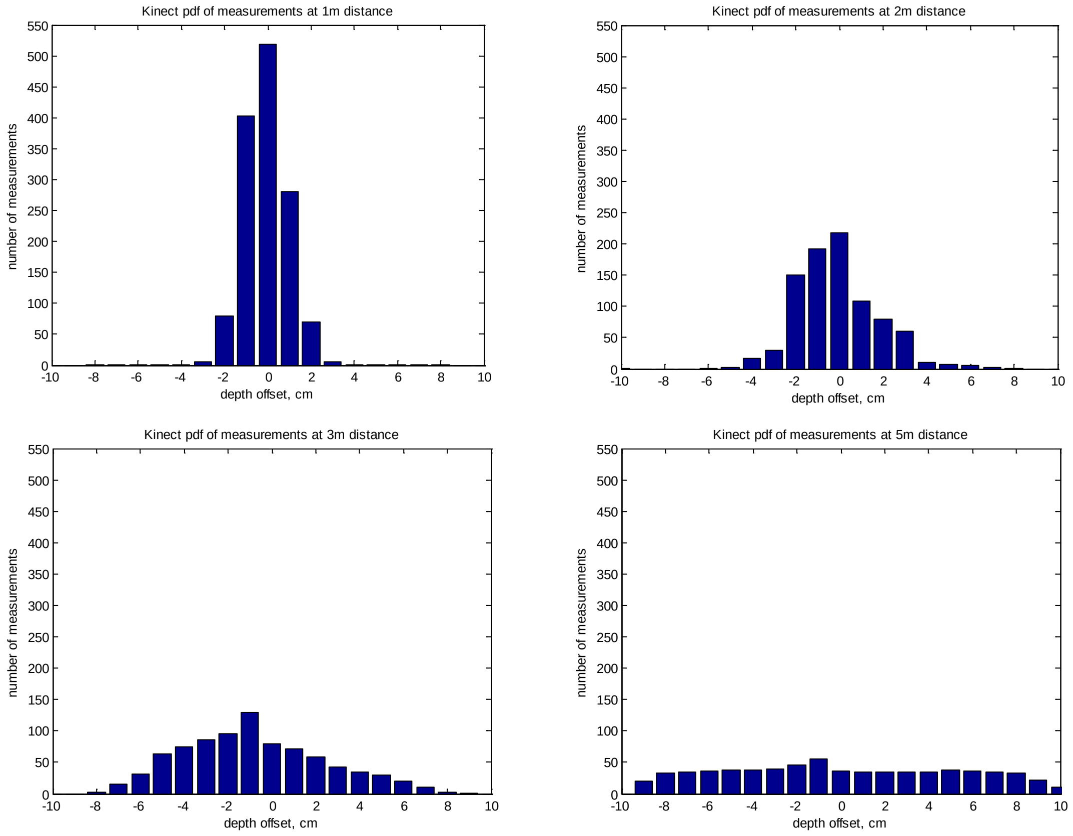


Figure 4-6: Distribution of measurement at various distances

The above sensor model is used in generation of synthetic data with a similar noise profile for initial evaluation of the proposed algorithms, for estimation of the noise pattern of the generated point cloud by the sensor in run-time as well as training of classifiers as it avoids the need for manual labelling of the data. In particular, the random error noise of the sensor is approximated by a Gaussian with a standard deviation that corresponds to modelled standard deviation of the error at the particular measurement distance.

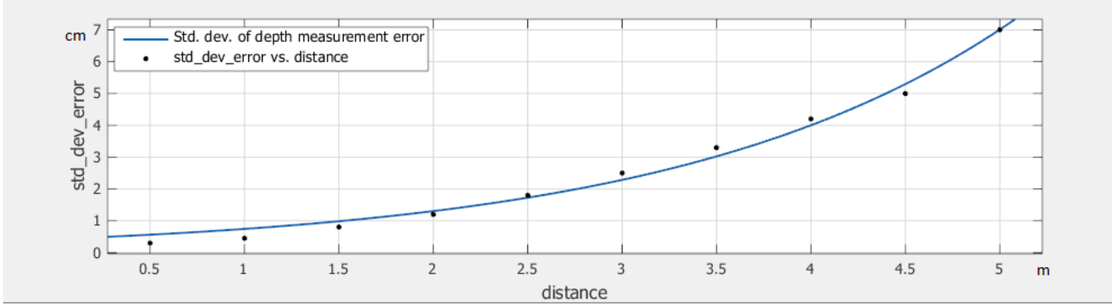


Figure 4-7: Model of the standard deviation of the Kinect's depth error

Related to the increasing random error of the sensor is the effect of rapid loss of depth resolution at longer distances. Due to the nonlinear internal quantisation, selected at design time by the manufacturer, approximately 90% of all possible depth values of the output signal are used to encode the distances up to 3m, leaving the remaining 10% of the values to cover the remaining 62% detection range, up to 8m. Such a nonlinear depth resolution function, resulting in a single step increment of about 15 cm at longer distances, causes the observed earlier exponential increase of the sensor error and represents a major challenge for the use of the sensor for human detection in the range beyond the specified adequate use area, i.e. 3.5 m.

The above challenge of the depth quantisation is shown in Figure 4-8, where the contour of a human body, displayed in turquoise , is detected by the sensor as several quantised depth layers, displayed in purple.

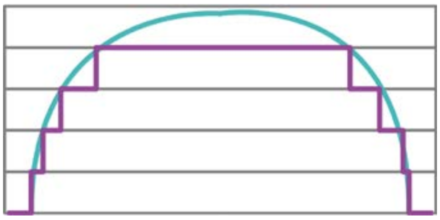


Figure 4-8: Illustration of the quantization effect of the sensor

Due to the above quantization effect, any smooth surface of the human body or an object is transformed into a number of stacked layers of 2D images. With the increase of distance, this effect results in progressive deterioration of the image to the point when the human body becomes unrecognisable by the detectors, as demonstrated by observations of the human body at different distances in Figure 4-9 below.

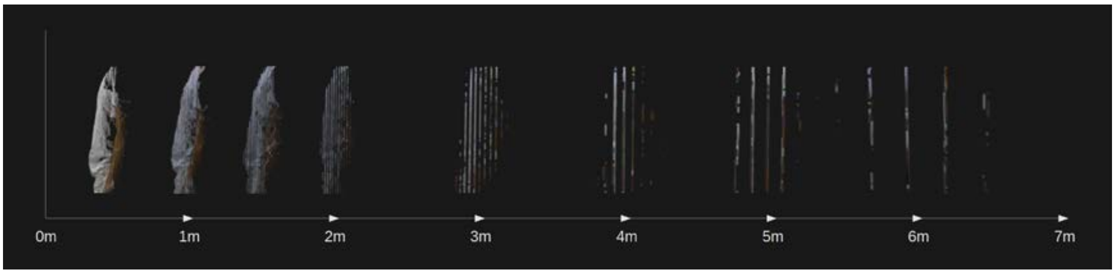


Figure 4-9: Effect of progressive loss of depth resolution of the Kinect sensor at different distances.

The above illustration demonstrates, the rapid degradation of the signal beyond that specified by the vendor “adequate play” distance of 2.5m.

Due to the above layering phenomenon, the typical local surface descriptors, designed primarily for smooth surfaces, cannot be applied efficiently for detection of people by RGB-D sensors.

Another problem with the sensor, associated with the underlying physical principle of measuring the depth from the structured light originating from a single source, is the inhomogeneous density of the resulting point cloud. Indeed, as observed in the experiments carried out, the density of the point cloud is strongly dependent on the distance between the sensor and the object. The above issue represents an additional problem for most of the local surface descriptors that typically operate on a homogeneous point cloud.

In conclusion, to combat the above deficiencies of the sensor and improve its usability for human body detection, a new descriptor with improved resilience to the layering effect is proposed below.

4.5.2 Human Interpretable Signature of Points (HISP)

In this section, a new local feature descriptor, i.e. Human Interpretable Signature of Points (HISP), is proposed. It is designed to offer increased resistance to the layering effect of the RGB-D sensor and to be invariant to distance. In particular, it represents the local variations of the surface topography of a depth image by an approximation aimed at reducing the layering effect and the variable with distance point cloud density, both representing typical properties of the low-cost RGB-D sensors.

For the purpose of human detection, the HISP descriptor is used to learn the typical signatures of various human body parts and later is applied again in the run-time classification from point cloud data, originating from RGB-D sensor. After the body parts in the depth image are identified, the segmented regions of the image are classified to determine the presence and the location of people in the image. Additionally, the position of the human body parts is used for initialisation of the human pose tracking algorithm, described in Chapter 5.

Overall, the HISP descriptor is based on the computation of a quantitative measure of the spatial distribution of the points that surround a selected key point. This is achieved by dividing the surrounding space into several 3D geometric shapes and counting the number of points belonging to the volume defined by each of the shapes. The 3D shapes are selected specifically to guarantee the invariance of the descriptor to rotation.

Initially, the space division process starts with the building of a reference coordinate frame, which is based on establishing the surface normal at the location specified by the given key point. It is important that the above coordinate is stable because it determines the distribution of points and hence the invariance of the signature. From the normal and the key point, the local coordinate system is constructed, forming a base for construction of the rest of the geometric shapes. The shapes are then used for splitting the space and evaluating the number of points from the point cloud in each shape as described below. The overall method for computing a HISP is summarised in the Figure 4-10 below.

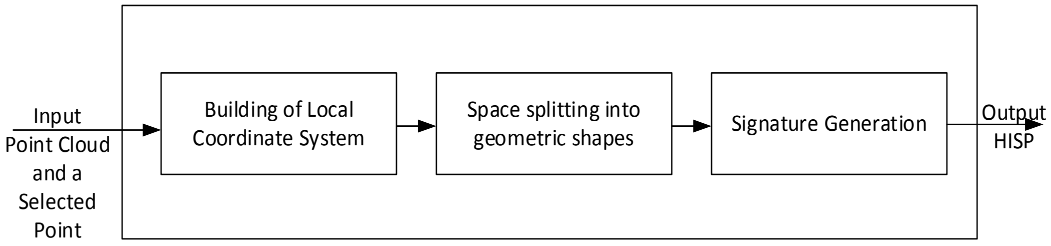


Figure 4-10: Overall method for computation of a HISP

4.5.2.1 *Local Coordinate System*

The HISP signature calculation starts with building a local coordinate system. For any given point from the point cloud, the surface normal, passing through this point, can be computed from the position of the surrounding points. This operation is equivalent to solving the problem of a plane fitting to the point and its surrounding neighbours. Then, as the normal of the fitted plane is in the same direction as the normal to the point, the only required additional step is to determine the orientation, i.e. the sign, of the normal through the key point. This step is necessitated by the fact that the normal of the fitted plane does not have orientation.

Given the acquired point cloud, there are two possibilities for obtaining the surface normal at any point of interest:

- Obtaining the underlying surface from the point cloud, using surface meshing techniques and then computing the surface normal at the point of interest. The problem with this approach is the additional computational overhead, incurred during the creation of the mesh;
- Using an approximation to infer the surface normal at the key point directly. This technique is the preferred choice in HISP due to its computational inexpensiveness. Also, achieving an absolute precision in the finding the surface normal is not essential for the computation of HISPs because the embedded mechanism in HISP for dealing with fluctuations in the orientation of the normal reduces the effect of small fluctuation of the orientation of the normal.

There are several different surface normal estimation methods in existence, as surveyed in Klasing et al. (2009). The simplest, fastest, and, therefore, the most suitable method for use in HISP is based on the first order plane fitting (Berkmann & Caelli 1994). In this method, the problem is reduced to a least square plane estimation problem for the neighbouring points (Shakarji 1998) by using the approximation that the normal to a point on the surface is sufficiently close to the normal of a plane tangent to that surface.

In particular, if the plane is represented by a point x and a normal $\vec{n}$; and the distance from any of the neighbouring points p_i to the plane is defined as d_i, then d_i can be found as the projection of $\vec{n}$ onto the vector from x to p_i. Subsequently, a right angle triangle can be constructed, by drawing a line from the “tip” of the vector $(p_i - x)$ perpendicular with $\vec{n}$, which has an angle ϑ between the angles and the hypotenuse of length $|(p_i - x)|$. The length of the projection, or the “near side”, is then given by: $|\vec{n}|\cos(\vartheta)$. Since the dot product can be defined as $(p_i - x) \bullet \vec{n} = |\vec{n}||(p_i - x)|\cos(\vartheta)$, to get the length of the projection the dot product has to be divided by $|(p_i - x)|$ which gives the distance from the point p_i to the plane the following equation:

$$d_i = \frac{(p_i - x) \bullet \vec{n}}{|(p_i - x)|} \tag{4.2}$$

The values of the normal $\vec{n}$ and point x are computed, using the least- square minimisation, so that the distances of the plane to the neighbouring points, are minimised. Then, as described in (Shakarji 1998), by taking the point x as centroid of the k neighbouring points:

$$x = \overline{p} = \frac{1}{k}\sum_{i=1}^{k} p_i \tag{4.3}$$

the solution for $\vec{n}$ can be found by analysing the eigenvalues and eigenvectors of the covariance matrix $C \in \mathbb{R}^{3x3}$ of the all neighbouring points and applying singular value decomposition (Golub & Kahan 1965), given as:

$$C = \frac{1}{k}\sum_{i=1}^{k}\xi_i\ (p_i - \overline{p})(p_i - \overline{p})^T, C\vec{v}_j = \lambda_j\,\vec{v}_j, \mathrm{j} \in \{0,1,2\} \tag{4.4}$$

In the above equation, the term ξ_i, representing the weight for the p_i, is most often equal to 1 as the points are given equal weights, C is the covariance matrix which is symmetric and positive semi-definite, and its eigenvalues are real numbers $\lambda_j \in \mathbb{R}$. The eigenvectors $\vec{v}_j$ form an orthogonal frame that corresponds of the principle components (I T 2002) of the dataset of neighbouring points. If $0 \le \lambda_0 \le \lambda_1 \le \lambda_2$, the eigenvector $\vec{v}_0$ corresponding to the smallest eigenvalue λ_0 is an approximation of the normal vector $\vec{n}$ or the opposite normal vector, i.e. with a minus sign, i.e. $-\vec{n}$.

The only problem with the above approach is that there is no method in existence to establish analytically the sign of the $\vec{n}$. This results in an ambiguity of the solution in cases when the orientation of the normal is estimated via the Principle Component Analysis (PCA) (Pearson 1901) method for a synthetic point cloud without defined position for the observer. However, in a real world robotics scenario, the observer position is typically well known in space, i.e. the position of the robot. This position can be used to assign in a trivial manner a consistent sign to all surface normal vectors, computed from the point cloud. In particular, this is achieved by orientating them towards the position of the sensor, as proposed by Rusu (2009a). The computation of the sign of the normal is achieved by adding the following condition, based on a positive dot product between the normal and the vector formed between the viewpoint and a point p_j for which the normal is computed as:

$$(p_k - p_j) \bullet \vec{n} > 0 \tag{4.5}$$

If for some reason the viewpoint is unknown, then a method, proposed by (R. Bro 2007), based on a singular value decomposition is also applicable to the problem of resolving the sign ambiguity of the surface normals, albeit at the expense of extra computational resource.

In conclusion, the PCA algorithm is selected for use in the HISP descriptor for computation of the normal at the key point as it is currently the fastest algorithm in existence for plane fitting. Overall, the speed and recall are the more important requirements for plane fitting part of HISP than achieving an absolute precision. The comparatively lower accuracy of PCA does not represent a significant issue for the HISP due to the approximation procedures employed at latter stages of the HISP computational chain. In particular, these approximation procedures, described in 4.5.2.3, mitigate to a large extent the lower precision in finding the surface normal. The above trade-off of speed against absolute precision of the descriptor is considered a justified design decision that enables significant acceleration of the feature computation from a typical RGB-D point cloud stream.

4.5.2.2 Space Division

The main idea behind the HISP descriptor is that it extracts and stores only high-level spatial information about the distribution of the surrounding points around any given key point. The distribution of the points is measured by dividing the space around the key point into several pre-defined volumetric 3D shapes, i.e. spheres, cones and slices, using the proposed method below, and counting the number of points in each of the shape. The resulting volumes representing a cross section of the 3D shapes are further referred to as bins.

Spheres

Firstly, the space around the key point is divided into a number of embedded spheres, centred at the key point. The process is illustrated in Figure 4-11. The rationale for using spheres is that this approach provides information about the radial distribution density of the surrounding points while it guarantees rotational invariance.

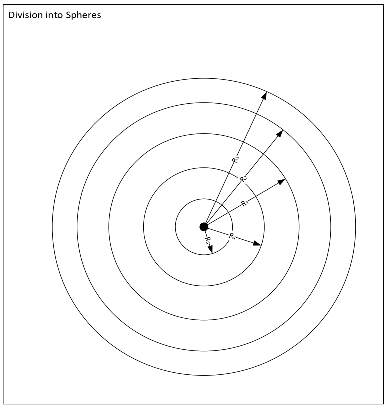


Figure 4-11: Division of space into spheres for HISP computation

The number of spheres and the radius of the biggest one are provided as input parameters to the HISP algorithm. The radius of each smaller sphere is then computed to ensure that the volume increase is preserved with the addition of each subsequent layer. This dependence guarantees constant weights of points at different distances from the key point. Overall, this feature gives distance invariance of the HISP signature.

Cones

Secondly, the space is further divided into solid three-dimensional figures, as illustrated in Figure 4-12. These shapes are further referred to as cones because of the visual resemblance with the cone shape. The cones are defined by: a) the key point at their apex; b) the angle at the apex α; and c) the base that is bound by the surface of the external sphere until it reaches the locus, made of all straight line segments joining the apex to the perimeter of the base. All cones in the descriptor share the same axis direction, defined by the direction of the normal $\vec{n}$. Finally, the desired 3D shape is achieved by removing the volume of each embedded cone inside, as illustrated in Figure 4-12. Similarly to the radius of the spheres earlier, the angle at the apex, α, for each cone is computed in such a way to ensure that the volume increase between cones is kept constant for all cones. This gives the same weight representation in the signature to points belonging to the different cones.

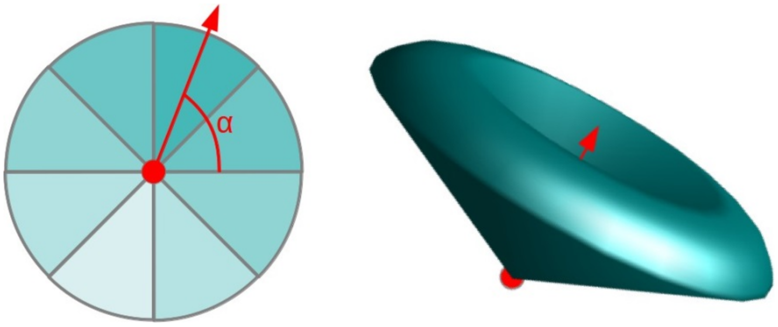


Figure 4-12: Division of space into cones for HISP computation

In similarity to the number of spheres, the number of cones is an input parameter, which is provided to the HISP algorithm.

Slices

The space around the key-point is further divided into 3D solid figures, referred to as slices. The slices are defined by a central point, i.e. the key point, and the radius of the external sphere, R_1, as shown in Figure 4-13 below.

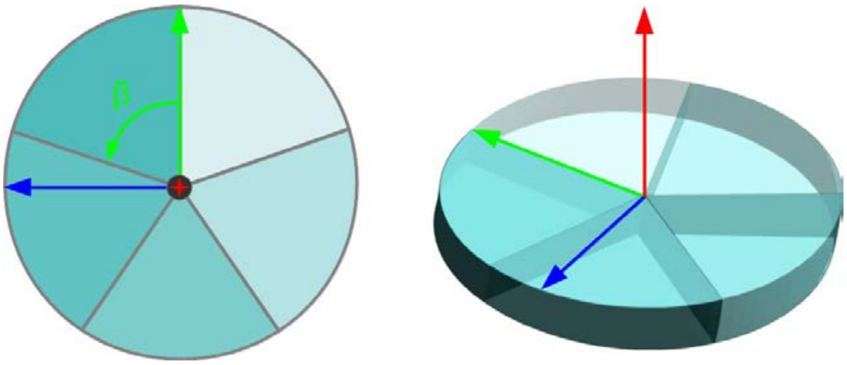


Figure 4-13: Division of space into slices for HISP computation

The angle β is calculated from the number of slices, which is an input parameter to the HISP algorithm.

Finally, after the spheres, cones and slices are defined in the space, based on the position of the key point and its normal $\vec{n}$, their cross section represents the volumetric bins, used for generation of the HISP signature.

4.5.2.3 Signature Generation

The HISP signature is a high-level descriptor of the distribution of the surrounding points in the space around the key point. The HISP captures not only the information about the surface points but also points that are close to the surface. The design objectives for the signature are: computational efficiency, ease of comparison between signatures, storage space efficiency and invariance to distance and point cloud density. Additionally, as the signature uses much less information to represent the local features than the original point cloud dataset, it is important that it has good precision and recall characteristics in order to be able to provide a reliable classification of human body parts.

After the surrounding space is divided into bins, as described in 4.5.2.2, the points surrounding the key point are assigned to their corresponding bin. However, the point to bin assignment is not done directly as this would cause dramatic changes in the signature resulting from a small fluctuation in the orientation of the normal. Instead, the assignment of points to a bin is computed using the mechanism proposed below to reduce the impact of noise and imprecision in the orientation of the normal on the resulting signature. The mechanism assigns partially points to adjacent bins, based on the distance from the point to the centre of mass of the bins, as shown in Figure 4-14. The reasoning behind the above approach is to minimise the effect of the variations in the spatial position of the normal on the HISP signature. Without the above partial point assignment mechanism, small fluctuations in the normal would assign points to different bins, resulting in points jumping between neighbouring bins. In turn, this would result in generation of different incomparable signatures from the same region of the point cloud.

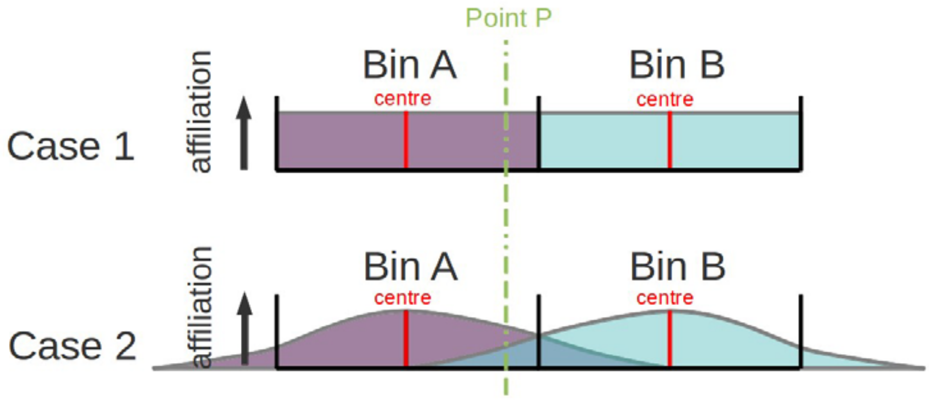


Figure 4-14: The proposed flexible bin assignment scheme for points

In particular, the assignment is done according to the complex affiliation function. The function is designed in similarity with the potential field function in physics, e.g. the gravitational force between two bodies with mass, where each body exerts a gravitational pull on the rest of the bodies. Similarly to the gravitational force, the affiliation function $A(d)$ is equivalent to the inverse of a quadratic function of the distance d between the point and the centre of the j^{th} bin and is given by the function below:

$$A_j(d) = \frac{k}{{d_j}^2} \quad , \tag{4.6}$$

where k is a normalisation coefficient specific for each point and j is an index of the bin. The reason for selecting an inversely quadratic function is to give rapidly diminishing affiliation weights to the points that are not immediately adjacent to the border of the bin. This reduces the noise effect on the signature while still preserving the spatial specificity properties of the descriptor.

As one point is partially affiliated with many surrounding bins, the coefficient k is normalising constant that is computed on the basis that the total sum of all affiliation functions is equal to one:

$$\sum_{j=1}^{M} A_j(d) = 1 \tag{4.7}$$

The sum in the equation symbolises the fact that all of the point's partial affiliations accumulatively result in the weight of one whole point assignment. It should be noted that as the affiliation function is inversely quadratic its effect is diminishing rapidly over distance. Therefore, in practise, for computational efficiency purposes, only the impact of the neighbouring bins are considered and the parameter M in (4.7) is the number of the neighbouring bins.

Overall, as a result of the variable affiliation assignment of a point, the jumping effect is mainly avoided as the assignment of points to bins gradually shifts with any fluctuation in the orientation of the normal. As a result, the signature becomes more robust to random noise and quantisation errors, as described in 4.5.1. The proposed mechanism also improves the recall characteristic of the HISP descriptor which is considered an important design factor for detection of human body parts.

After the affiliation is completed, the number of points assigned to each bin are counted. Then, the HISP signature is generated as described below. For each cone, contained in each sphere, the number of points is computed for each slice. Then, the separate assignment scores are concatenated creating the raw signature, represented by the bar graph, shown in Figure 4-15.

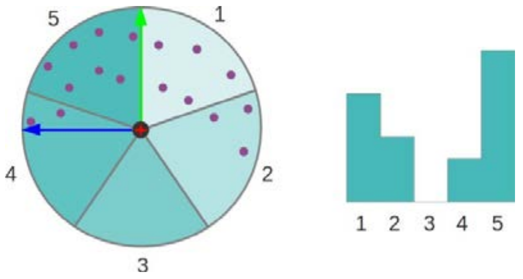


Figure 4-15: Generation of raw signature per cone.

The PCA algorithm, used in 4.5.2.1, provides the direction of the normal $\vec{n}$, shown by the red cross in the centre of the circle on the left side of Figure 4-15. However, the position of the other two vectors, used for building the local coordinate system and shown by the blue and green arrows respectively, cannot be determined mathematically. All that is known for them is that they lie within the plane defined by the normal. Such an uncertainty means that the whole local coordinate system can freely rotate around the normal at the key point. Therefore, to eliminate the above uncertainty in the descriptor, the raw signature is ordered by the number of points in each slice. By doing this, the angular information about the distribution of the points is lost. However, this masks the mathematical indeterminism resulting from the use of the PCA. It also introduces rotational invariance to the descriptor by preventing the comparison between two identical but rotated by PCA signatures of the same region from generating a negative result.

The above procedure for building the ordered signatures of a cone is repeated for all cones and all spheres. Finally, the individual signatures are concatenated based on the order of the cones and spheres to generate the final HISP descriptor, as shown in Figure 4-16 below.

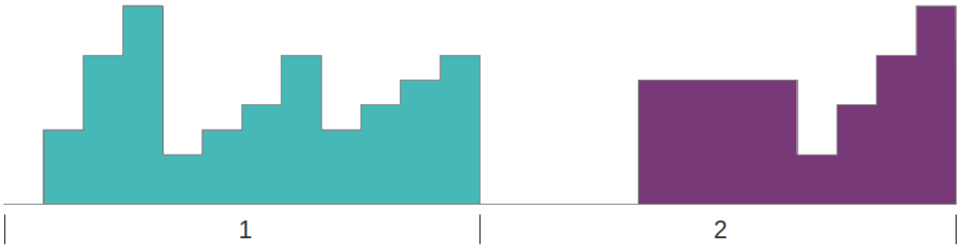


Figure 4-16: Combined signature for all cones and spheres for two spheres each containing three cones.

When homogenous point clouds are used, the descriptor shows invariance to the detection distance. However, in real applications, the point cloud generated by the Kinect depth sensor, or other 3D sensors, as analysed in 4.5.1, is of variable density that depends on the detection distance. The above density variation creates a problem for the comparison of the HISP signatures. In particular, with increase of the distance, the lack of sufficient information, i.e. number of measured points per voxel, results in distortion of the signature at a longer distance, which typically leads to an increased number of false positive and false negative matches in local feature searches for detection of people in the scene. The above distortion effect together with the resulting signatures is shown in Figure 4-17 below.

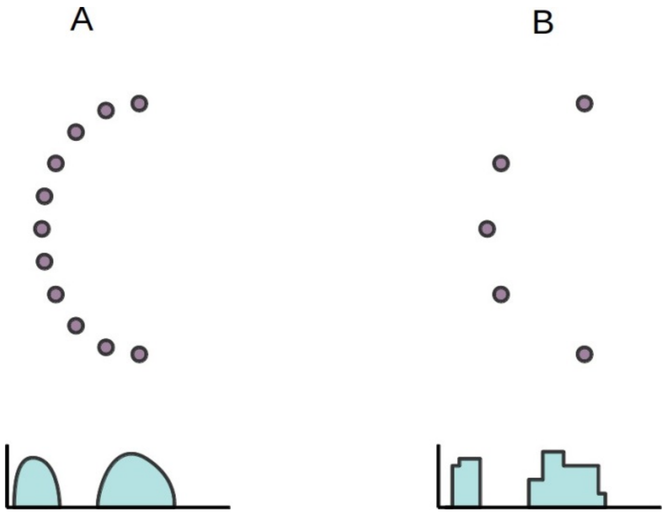


Figure 4-17 The effect of the detection distance on the signature. Displayed are the point cloud and the resulting signature: case A) at short distance; case B) at long distance

Although the missing points cannot be predicted, due to lack of sufficient information, it is possible to compensate for the effect by modifying the structure of the HISP signature. For this purpose, the density of the point cloud is homogenised by applying the Voxel Grid algorithm (Rusu & Cousins 2011). In effect, this is equivalent to selectively reducing the amount of the available information at shorter distances to improve the distance invariance of the descriptor. The application of the Voxel Grid algorithm gives the additional benefit of reducing the number of points in the point cloud, which subsequently reduces the computational load and the memory requirements of HISP descriptor. In particular, the operational principle of the Voxel Grid algorithm is based on a two step procedure: firstly, it splits the space into small boxes, i.e. voxels; secondly it looks into each of them and if there are any points it replaces them with a single point, positioned into their centroid. By doing these two steps, the Voxel Grid algorithm homogenises the point cloud, making it suitable for input into HIPS.

Experimentally, it has been confirmed that the dependency on the detection distance of HISP is reduced significantly when the resulting homogenised point cloud is used, as illustrated in Figure 4-18 below.

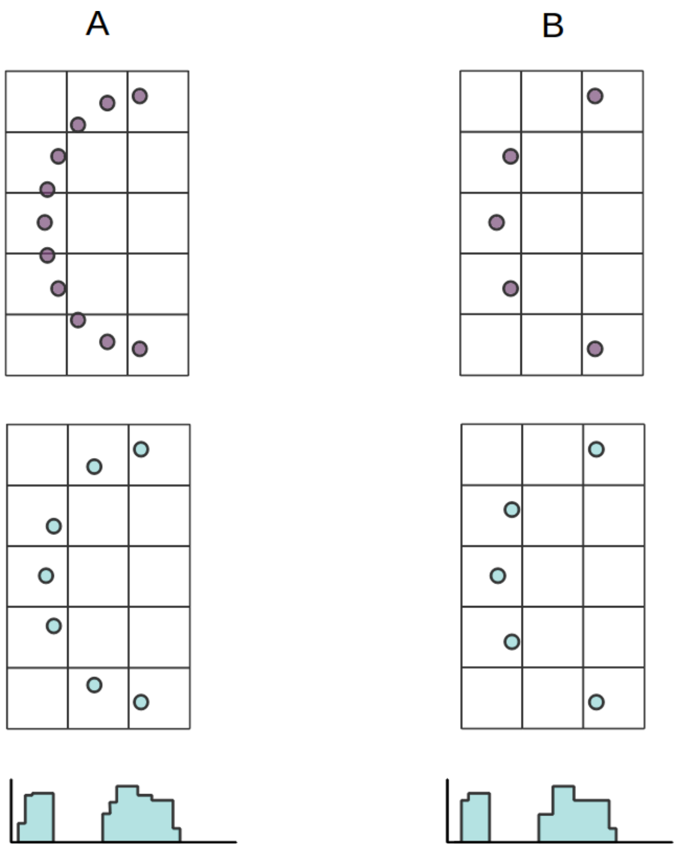


Figure 4-18 Reducing the effect that the distance has on the signature with the Voxel Grid algorithm

Efficiency of Computation

Because the generation of a HISP signature requires execution of a significant number of operations, it is not possible to achieve real-time processing of live data stream from a typical RGB-D sensor. However, because the same procedure is repeated independently for each pixel from the depth image, the process is a suitable candidate for parallelisation on a GPU. Once the point cloud is homogenised and transferred to the internal memory of the GPU, each core can start computing separately the HISP signatures and evaluating the extent of the match with the provided reference signature. This approach is believed to be able to achieve real-time performance. However, as this development is deemed to be only a technical challenge, it is considered outside the scope of this work.

4.5.3 People Detection through Human Body Parts Discovery from Depth Images via Random Forest Classification using HISP-based Features

The human detection approach followed in this section partially resembles the one used by Shotton et al. (2011) where, after segmentation from the background, simple depth features are used to label each pixel with a body part label. However, unlike Shotton et al. (2011) which uses the depth difference between two points from the depth image as the only feature for classification, the proposed method is based on more complex features, i.e. the HISP signatures described in 4.5.2. Using more complex features, describing the surface in more detail, enables better specificity of human detection. This leads to the ability to detect individual human body parts from a moving camera that is not an option with the approach of (Shotton et al. 2011).

Initially, in the pre-processing step, the point cloud is filtered using the Bilateral Filter (Kopf et al. 2007) and then it is down-sampled using the PCL's Voxel Grid algorithm, which is based on point cloud decimation (Renze & Oliver 1996).

In the next step, a pixel level labelling is performed using a Random Forest (RF) classifier that is using the generated HISP features of the points that have been computed by down-sampling of the original point cloud.

Classifier Training Considerations

For the purposes of training of the classifier, a sequence of training depth images is collected from a depth camera by observing people performing random motions. The collected images are subsequently segmented and labelled manually with the names of the human body parts. Then, after down-sampling, aimed at reducing the number of pixels, the HIPS signatures have been calculated on the pixel basis and used, together with the body parts labels, for training of the classifier.

Manually annotating large amount of data, needed for training of the classifier, is a time-consuming and laborious task. Insufficient training limits severely the performance of the classifier. Options for automated rendering of the training data that are able to generate vast quantities of synthetic depth data have been considered. One of them, proposed by Buys et al. (2013), is a pipeline for synthetic generation of human body data, using a virtual model of a human generated by an open source software, i.e. MakeHuman (MakeHuman.org 2014). Generated in this way, the model consists of a mesh structure with an underlying kinematic chain that can be additionally manipulated to simulate different human poses. As proposed by Buys et al. (2013), this manipulation can be achieved by mapping the virtual human model to human motion data, originating from a human motion capture system. Finally, by using the ray tracing method (Buys et al. 2013) the mesh representing the human in different poses is converted into a point cloud. Subsequently, the point cloud is used in training of the classifier, in the same way as the manually annotated depth data is used to train the human body part classifier. Since achieving of a high precision commercial grade detector is not targeted in this work, using a manual annotation of the data is considered sufficient for demonstration of the principal of operation of the method. Further improvement of the characteristics of the classifier, through automation of the training process, is feasible and will be considered as future work.

Run-time

In run-time, a feature vector $\tilde{f}_{HISP}(\tilde{I},\tilde{x})$ is computed for each point $\tilde{x}$ from the current downsampled image $\tilde{I}$. Then $\tilde{f}_{HISP}(\tilde{I},\tilde{x})$ is passed to the pre-trained RDF classifier, where, a tree n from the forest trees produces a posterior distribution over the point label $P_n(c|\tilde{I},\tilde{x})$. Combining the individual probabilities produces the forest estimate of the class:

$$P(c|I,x) = \frac{1}{N_{tr}} \sum_{n=1}^{N_{tr}} P_n(c|I,x), \tag{4.8}$$

where $P(c|I,x)$ is the probability that the $\tilde{x}$ point has label class c as, N_{tr} – number of RF trees, $P_n(c|I,x)$ the probability given by the individual trees.
If the probability is higher that a certain predefined threshold for a class then the point is labelled as belonging to the class, representing one of the main body parts, i.e. shoulder, arm or shank, chest, thigh, hand. Additionally, a class representing a non-human surface is introduced.

After every point in the down-sampled image is processed, the result is stored together with the point for further use. For example, Figure 4-19 below visualises a colour coded map of the labels computed for a typical depth image containing human.

It can be observed from the image that the human body parts are, on the whole, classified correctly. The flat vertical walls of the room are also classified correctly. Only the floor, which is removed at the next step of the method, is classified incorrectly (in green) as human chest due to its flat appearance. It is considered that with more extensive training that is including points of the floor point cloud part in the negative training set this deficiency can be largely avoided.

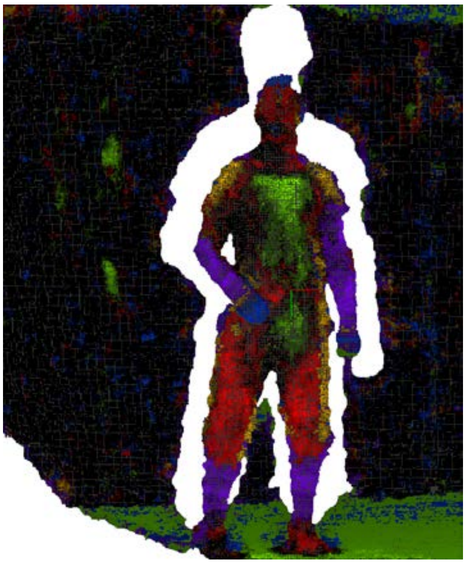


Figure 4-19: The colour coded map with colour labels of human body parts. *Legend: orange – shoulder, purple – arm and shank, green – chest, red – thigh, blue – hand, black – non-human*

Finally, after background removal, using the Grabcut method (Rother et al. 2004); floor plane removal, using RANSAC plane fitting method (Fischler & Bolles 1981) and segmentation into different clusters, using Euclidian clustering algorithm (Oehler et al. 2011), a classification is performed on the different clusters to determine which of them corresponds to a human body. The classification is based on a two-thirds voting method, i.e. if the points labelled as belonging to a human body part represent at least two-thirds from the points in the point cloud segment and at least three different body parts are identified, the segment is considered a representation of a human body. Then a positive human detection with the position at the centre of mass of the point cloud segment is made. By varying the percentage of classified points in the voting criteria of the detector, the specificity and sensitivity of the detector can be changed. After processing of all point cloud segments and identification of all detected human bodies in the depth image, the position of the human candidates is recorded on the room occupancy map, to be used later by the classifier fusion module, as described in 4.7.

4.5.4 Human Detection from Colour Images

Colour component data is part of the RGB-D data stream that is captured by the sensors of the robot. It also represents one of the modalities in which people can be detected efficiently by the service robot.

The advantages of detecting people in colour images in comparison with detection in depth images include capability to do detection at very long distances, especially when high-resolution cameras with optical zoom capabilities are used. However, as a drawback, detection using only colour imagery is influenced strongly by illumination conditions and variance in clothing. Therefore, in the case of human detection by a robot the colour image modality can be best used in conjunction with other modalities, e.g. for building or confirming a hypothesis about the human presence.

There are several well-established algorithms for detecting people from colour images. As analysed in 2.1.1, the HOG detector offers superior performance among the methods for human body detection from colour images. Therefore, this work does not focus on developing new algorithms for this purpose. Instead, the original visual HOG detector (Dalal & Triggs 2004) is used and the output of the detector is fed into the classifier fusion module.

In addition to the HOG detector, a face detector, based on the original Viola-Jones face detection algorithm (P Viola & Jones 2001) is used in parallel to provide additional cues from the colour component about the presence and the location of people. Although the face detection is only applicable when people are facing the camera, when it is made, it serves as a very strong cue for the overall human detection classification output.

Although the distance to the face or person in both HOG and the face detector is not available directly from the colour image, it is possible that a rough distance estimate can be obtained by matching the scale of detection with pre-recorded detection scales.

The distance can also be estimated by matching the colour image to the depth image component in the RGB-D data on a pixel-by-pixel basis. However, for this approach to work, it is necessary that both sensors are calibrated in advance.

4.6 Tracking of the Human Location through Filtering

For the purpose of the sensor fusion algorithm, proposed below, it is necessary that the detected locations of targets are tracked. Tracking is achieved by applying a standard probabilistic filter as explained below.

Human motion in 2D can be approximated by a linear system. This approximation makes it possible that a Kalman filter is used for tracking of targets using detections sequentially from each modality. In comparison with other filtering methods, e.g. Particle filters, the Kalman filter offers very low computational overheads, which allows tracking of the detections of many sensors using only the limited computational resources on-board of a mobile robot.

In summary, the filtering is done recursively in a predict – update cycle according to the Kalman filter algorithm (Thrun 2005) as shown in Algorithm 4-1 below.

In the algorithm, $\boldsymbol{\mu_{t-1}}$ and $\boldsymbol{\Sigma_{t-1}}$ are the mean and covariance matrices at time $t-1$, $\boldsymbol{z_t}$ is the measurement vector, i.e. the detections from the particular sensor, $\boldsymbol{\mu_t}$ and $\boldsymbol{\Sigma_t}$ represent the belief, i.e. mean and covariance, at time t. The Kalman gain, K_t, determines to what extend the measurement is incorporated into the belief.

```
Input: Prior ($\boldsymbol{\mu_{t-1}}, \boldsymbol{\Sigma_{t-1}}$) , Measurement $\boldsymbol{z_t}$, Control $\boldsymbol{u_t}$
Output: Posterior ($\boldsymbol{\mu_t}, \boldsymbol{\Sigma_t}$)
1: prediction step     : $\bar{\mu}_t = \boldsymbol{A_t}\mu_{t-1} + \boldsymbol{B_t}u_t$
                       : $\bar{\Sigma}_t = \boldsymbol{A_t\Sigma_{t-1}A_t^T} + \boldsymbol{R_t}$
3: gain calculation    : $K_t = \bar{\Sigma}_t C_t^T (C_t\bar{\Sigma}_t C_t^T + Q_t)^{-1}$
4: update step         : $\mu_t = \bar{\mu}_t + K_t(z_t - C_t\bar{\mu}_t)$
                       : $\Sigma_t = (I - K_t C_t)\bar{\Sigma}_t$
6: return $\mu_t, \Sigma_t$
```

Algorithm 4-1: Kalman filter used for tracking of the individual targets

The detector fusion approach proposed in this work is based on incorporation of a measure of uncertainty from the trackers of the individual detectors, i.e. the covariance of the individual Kalman filters, Σ_t, into the weighing of the classifiers as described in section 4.7.

Finally, an individual filter is assigned to track the detections of each separate person in the scene. The assignment is based on the proximity of the target to the current tracker position, based on the pre-defined minimum and maximum thresholds, provided as parameters to the algorithm. This approach works well when there are a limited number of people in the observed space and they are well separated from each other. However, with the increase of the human density the probability of a wrong assignment of detection to a tracker increases. Therefore, a more sophisticated method for detector-tracker assignment is proposed later in Chapter 6 to deal with the more complex situations.

4.7 Detector Fusion through Adaptive Combination of Classifiers

When the output from several classifiers for human detection is available simultaneously, finding the optimal way of combining the results is essential for the quality of the overall detection decision. Making the correct decision, about the presence and location of the people in the environment, is critical for the processes of high-level path and task planning of the robot, which are a necessary foundation for achieving a meaningful HRI.

The overall aim in combining the classifier outputs is to benefit from the positive aspects of the classifiers while reducing the influence of their deficiencies. In practical terms, the goal of the classifier fusion is to represent lower false positive and false negative rates of the ensemble classifier in comparison with those of the individual detectors. Therefore, a specific design of the proposed ensemble classifier method is targeted that can achieve improved sensitivity and specificity. In particular, given a supervised binary classification problem K, i.e. the human detection from a number of human m detectors $D = \{a_1, \dots , a_m\}$, using different modalities, and a set of n training observations, $T = \{(x_1, a_1, y_1, z_1),\ (x_2, a_i, y_2, z_2), \dots , (x_n, a_m, y_n, z_n)\ \}$, consisting of a set of patterns generated by the human detectors and each instance x_i belongs to a domain X of observations; detector label y_k is a binary label from the set of detector labels $Y = \{human\ present,\ human\_not\_present\}$; and z_l is manually annotated label about the human presence at the particular location x_i , the task is to find a function $f: X \rightarrow Z$ that maps optimally the observation set X into an element of the ground through binary labels $Z = \{(z_1, \dots, z_n)\}$.

The proposed adaptive fusion method in this section is inspired by the mixture of experts approach (Jacobs et al. 1991). However, the proposed method, unlike the one in (Jacobs et al. 1991), employs a novel method for computing the weights of the classifiers in the ensemble. The proposed adaptive weighting is considered to be better suited to the human detections from a mobile robot due to the combination of specific dynamic and static factors that are taken into account in the fusion process. In particular these factors are related to the current and past performance characteristics of the individual classifiers.

In principle, if the classifiers are similar in terms of output, or in the extreme case, equal, combining them is not expected to bring any significant advantage. On the opposite, when the classifiers differ it is considered that the overall information gain in the ensemble could benefit from the diversity of the individual classifiers. In a typical service robotics case, the signals originate from different sensors and are further processed by diverse in nature algorithms. Such a diversity can be viewed, depending on the context, as a measure of complementarity, dependence or orthogonally between the individual classifiers as discussed in Kuncheva & Whitaker (2003). The above sensor divercity has motivated the proposed classifier fusion method.

The main challenge in building an accurate ensemble classifier is finding a suitable method for combination of the output of the individual classifiers. It is important not only that the ensemble classifier outperforms the individual classifiers but also that it can achieve a meaningful output even when there is missing or inconsistent data from the individual classifiers. Therefore, establishing a suitable way in which classifiers should be combined is not considered to be a trivial task. Standard strategies are based on learning classifier combination functions from data, by relying on static pre-recoded datasets. However, classifier selection based purely on a comparison of the performance of the available individual classifiers on a representative dataset is unsuitable in the context of human sensing in service robotics. The reason for this are the numerous unknown external factors influencing the performance of the classifiers. Building a dataset covering all the external factors is considered a complex and largely unfeasible task. In particular, the highly dynamic and unpredictable nature of the environment greatly complicates the task of assessing the run-time classifier performance by using a pre-build test dataset. The fluctuations in performance of the classifiers are explained by the influence of a range of external factors, some of which are not known. Since measuring or modelling of them is not a trivial task they cannot easily be included in the training of the classifiers. For example, a classifier can perform satisfactorily under certain conditions at one time frame, and in the next one, under apparently similar conditions, it may fail because of the effect of an unknown external factor that has not been included in the learning process. For example, in the context of service robotics, such an external factor could be the lighting conditions, e.g. the spectral distribution of the light which depends on the amount of direct sunlight entering into the environment. As the infrared component of the sunlight spectrum affects the performance of the infrared based sensors, and subsequently the performance of the accosted classifiers, it represents one of the above factors that cannot be learned. At the same time, it is unfeasible to model or learn the performance of the classifier under every possible variation of the lightning. The above inability to predict, measure and subsequently learn every possible variable, affecting the performance of the classifiers, has inspired the proposed below approach. The approach is based on a combination of an approximate estimation of the expected performance of the individual classifiers, learned from historic data, and the runtime estimation of the consistency of the classifiers over a number of consecutive time frames. The later fine-tune adjustment of the importance weights of classifiers acts as a compensation for the omission from the learning of the additional influencing factors. The combination of dynamic and learned performance measures results in the increased ability of the classifier fusion algorithm to adapt rapidly to changes in the environment, while still maintaining operational ability when by the tracking information becomes unavailable.

4.7.1 Clustering of the human detections

Before an ensemble classifier can be applied, it is necessary to determine which detections are considered related to a particular target. Such a clustering allows only relevant classifiers to be included in the ensemble classifier.

The computation of the clusters is performed according to the clustering algorithm proposed below. The main idea is to use the confidence of the trackers, i.e. the variance, as a measure of the temporal consistency of the detectors. Both the variance and the Euclidean distance between detections form the basis for a decision about cluster formation. In particular, for two detections to belong to the same cluster at least one of the conditions listed in Table 4-1 must be satisfied:

Label	*Condition*
Condition A	the tracker variance area of the evaluated detection overlaps with the variance area of a tracker of a detection and Euclidean distance between them is less than the maximum joining threshold distance, d_{max};
Condition B	the Euclidean distance between the two detections is less than the minimum threshold distance, d_{min}

Table 4-1: Conditions for forming a cluster

If one of the detections, meeting the above criteria, already belongs to a different cluster, the other detection also joins the cluster. If both detections belong to different clusters, then both clusters are merged into a single one. Finally, if none of the detections belongs to a cluster, then a new cluster is formed and it contains both of the detections.

The parameter d_{max} limits the cluster size in cases when the low confidence of a new tracker causes its variance to cover a large area, causing all detections within the area to be added to a single cluster. Overall, *Condition A* limits the tendency of the clustering algorithm when newly detected targets appear to include all trackers in a single cluster. The role of the parameter d_{min} is to limit the number of clusters when the confidence in tracking becomes very high, represented by very narrow peaks in the variance of the trackers. Since the detections, made by different sensors, can originate from different human body parts, and in addition it is also likely that there is random noise in the measurements, it is not unexpected to observe certain offset between the readings of the different sensors detecting the same body part. This offset, in case of high confidence of the trackers, will cause overfitting resulting in many small clusters per target instead of a single cluster. Therefore, the role of *Condition B* is to limit the tendency of the algorithm in certain conditions to overfitting.

4.7.2 Classifier Combination Algorithm

Instead of modelling or inclusion in the learning model of an every single factor influencing the performance of the classifiers, which is considered an unfeasible approach due to the high number of unknown factors, it is proposed that a novel classifier weighting method is used. The method, representing the main contribution of this chapter, is used to control the weights of the individual classifier outputs within the ensemble classifier. The method utilises both learned static classifier characteristics and dynamic sensor performance indicators to adapt the importance of the individual classifiers and respond to changes in the environment. In particular, the weighting function uses as an input, several data parameters that are known or possible to measure. It also relies on statistical learning of the performance of the classifiers under similar circumstances in the past. The output of the function is a vector consisting of the weights which are assigned to the output of each classifier to control its importance in the ensemble as described below.

Many factors affect the ability of the classifiers to make a successful detection, e.g. the luminosity and the contrast of the image as well as the sensor noise. However, it is considered that the most important and universally applicable factor influencing the human detection reliability in the context of service robotics is the detection distance. This observation can be explained by the degradation of the signal of the majority of the sensors with the increase of the distance. For example, depth sensors, as analysed in 4.4.1 and 4.5.1, are characterised with a depth component standard deviation error that increases proportionally to the square of the distance. Similarly, in image cameras with fixed focal lenses, the pixel image resolution is inversely proportionate to the detection distance. This effect results in an insufficient number of pixels representing the person in the image, e.g. a blurry image, making it impossible for a successful detection to be made beyond a certain distance.

Due to the high number of additional factors influencing the performance of the classifiers, it is difficult for the performance of the classifier to be learned. For example, such factors include the amount of ambient illumination or the reflectivity of the surface material determining the intensity of the reflected light of a laser range finder. Therefore, to compensate for the lack of ability to measure and model the full range of the factors influencing the classifier performance, the following two-tier weighing procedure is proposed. It compensates for the effect that the difficult to learn factors have by introducing a dynamic classifier performance measure. This measure also influences the weighting of the particular classifier in the ensemble. In particular, the measure is based on estimation of the temporal consistency of the output of the individual classifiers by using the variance measure of the associated filter. If the variance is low, the classifier performance is considered to be consistent with its previous detections and the weighting is therefore gradually increased. On the opposite, if the classifier produces inconsistent results at different times, its variance will gradually increase, which would indicate less certainty of its output. Consequently, the classifier will be given a proportionally smaller weight within the ensemble classifier.

The above procedure relies on the availability of the filter associated with each detector for the estimation of the temporal confidence of the classifier. When a new target appears, initially, it does not have a filter associated with it. In these cases, the weighing procedure defaults initially only to the learned historical performance of the classifier. After a tracking filter is associated with the target, its variance is initially high resulting in a small weight in the ensemble. Gradually as the variance is reduced, the two-tier weighting mechanism increases the importance of the classifier. In particular, the temporal confidence is measured by the variance-covariance matrix of the associated filter and the learned historical characteristic is based on statistical learning of performance of the classifier at different detection distances.

In particular, the weighing of the classifiers in the ensemble is done in a two-tier gating procedure described below. The first tier gating represents the historical performance of the classifiers while the second level represents their temporal consistency.

The rationale for the above approach is three-fold and includes:

- **dynamic protection from a failure of a single classifier**. If a classifier is underperforming under the current dynamic conditions of the environment, the uncertainty of its tracker will gradually increasing. The increase will continue until detections become more consistent. In case of inconsistent or missing detections, the increased uncertainty will result, under the proposed gating weighting procedure for classifier combination, to assigning of a proportionally smaller weight to the classifier in the ensemble. After normalisation of the weights within the ensemble, the smaller weight of the underperforming classifier will lead to increased impact of the rest of the classifiers;

- **both fast reaction to a rapidly changing environment and good accuracy in normal situations.** The proposed weighting procedure reacts quickly to changes in the environment due to the learned classifier performance from the measurement distance. At the same time when there isn't much rapid change in the environment, the trackers will be able to gradually build up sufficient confidence over time which will start contributing to the fine adjustment in the classifier combination;

- **a level of protection against a general inconsistency of all classifiers.** If all trackers of classifiers report low confidence through their variance-covariance matrices, e.g. as in case of newly tracked targets, then, because of the second tier gating and the normalisation of the weight parameters, the ensemble combination will be dependent mainly on the learned performance data for each classifier, which will produce less accurate but still sufficiently usable results.

The proposed two-tier gating procedure, illustrated in the Figure 4-20, offers a balance between the ability for reaction to rapid changes in the environment and robustness of ensemble performance by combining dynamic tracker-based adaptation and learned performance characteristics. Moreover, it offers a certain level of protection against total failure of one or more classifiers.

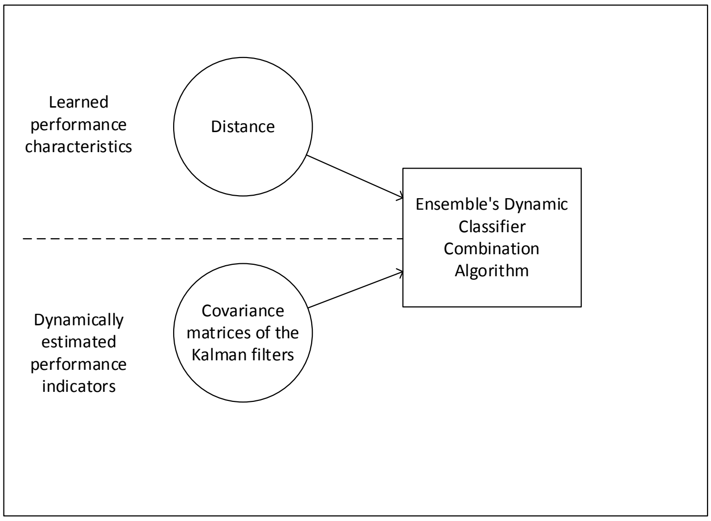

Learned performance characteristics
Distance
Ensemble's Dynamic Classifier Combination Algorithm
Dynamically estimated performance indicators
Covariance matrices of the Kalman filters

Figure 4-20: The concept of the two tier gating procedure for weighting of the classifiers in the ensemble

Using simultaneously both dynamic and static sources of performance information about the detector, i.e. the covariance of the tracker and the learned error rate at that particular distance, allows the proposed ensemble weighting algorithm to benefit both from the accumulated knowledge about a classifier and its current dynamic performance. In addition, this increases the robustness of the classifier weighing mechanism, e.g. when one of these sources of information fails to provide information the others takes over due to the dynamic weighing, used in the ensemble.

In particular, in circumstances when there is no sufficiently accurate historical model of the classifier performance the weighing mechanism relies mainly on the dynamic measure, inversely linked to the tracker covariance. Similarly, when a new object appears in the environment, the ensemble is still operational as it relies solely on the learned performance data. This heavy reliance on historical data continues until the confidence of the filter increases sufficiently to allow redistribution of the weighing of the classifier importance in the ensemble classifier. In principle, the confidence of the tracker increases when the latest detections are consistent with the predicted ones.

The increased confidence results in a reduction of the variance of the tracker and higher weight of the classifier in the ensemble.

Since the output of every classifier can be modelled as a probability (Platt 2000), (Cover & Hart 1967), given as:

$$p(\Omega|\boldsymbol{X},\vartheta_i) \simeq f_i\big(c_i(X)\big), \quad i = \{1\,..n_c\}, \tag{4.9}$$

$p(\Omega|\boldsymbol{X},\vartheta_i)$ is the output of the ensemble, where $\boldsymbol{X}$ is the input data, c_i is the i^{th} out of n_c classifiers, ϑ_i represents the parameters of the i^{th} classifier, f_i is the function that maps the classifier output to probability and Ω is the event that indicates the presence of a person at the particular position. Then, following (Jordan & Jacobs 1993), (4.9) can be transformed into:

$$p(\Omega|\boldsymbol{X},\boldsymbol{\vartheta}) = \sum_{i=1}^{n_c} g_d^i(X).\; g_h^i(X).\,p(y|\boldsymbol{X},\vartheta_i)\,, \tag{4.10}$$

where $g_h^i(X)$ is the historical and $g_d^i(X)$ is the dynamic gating functions of the i^{th} classifier, $p(y|\boldsymbol{X},\vartheta_i)$ is the probability of detection of the i^{th} classifier. The factor $p(y|\boldsymbol{X},\vartheta_i)$ is determined from the sequence of resent detections, positive or negative, of the classifier within the cluster using Bayesian interference as explained in the illustrative example given in 6.7.

The second level gating function, i.e. $g_d^i(X)$, is acting as a coefficient to the lower-layer output $g_h^i(X)$ instead of to already summed up lower-lever. This direct control of the historical importance allows a better response to changes by applying individual weighting, i.e. the function $g_d^i(X)$, to the classifiers outputs based on their runtime dynamic performance.

Dynamic gating function

In the proposed algorithm, the dynamic gating function $g_d^i(X)$, representing the dynamic temporal consistency of the classifier, is determined by the variance-covariance matrix of the associated Kalman filter, Σ_t as defined in Algorithm 4-1. In particular, the elements along the main diagonal of the matrix Σ_t determine the window of uncertainty of the elements of the state X - when elements in the diagonal are bigger the uncertainty is higher. Therefore, the dynamic gating function $g_d^i(X)$ is given as:

$$g_d^i(X) = \frac{A_d}{\sqrt{\sigma_{1,1}^2 + \sigma_{2,2}^2}}\,, \tag{4.11}$$

where $\mathrm{A_d}$ is a coefficient, given as an input parameter to the parameter, $\sigma_{1,1}^2$ and $\sigma_{2,2}^2$ are the diagonal elements of the covariance matrix Σ_t.

Historical gating function

The statistically learned classifier performance characteristics are represented in (4.10) by the historical gating function, $\mathrm{g}_{\mathrm{h}}^{i}$, where i is the index of the classifier. The gating function allows higher weights in the ensemble to be given to classifiers that have outperformed the others in the past under similar conditions. For this purpose, the performance of the classifiers is learned from manually annotated data as described below.

Initially, a labelled training set, containing the record of the classifier detections at different detection distances is used to build a collection of confusion matrices, i.e. one matrix for each of the distance ranges. Subsequently, from the above confusion matrices a function of accuracy of the detector over distance is constructed by using non-parametric fitting, i.e. the localised regression smoothing technique (Cleveland & Loader 2001). This approach results in a model that can interpolate the accuracy of the classifier for unobserved distances. In run time, the above model is used to determine the value of the historical gating function $\mathrm{g}_{\mathrm{h}}^{\mathrm{i}}$ by using the following equation:

$$g_h^i(d) = A_d * (Acc(d) - A_B) , \tag{4.12}$$

where A_d is a coefficient, given as input parameter, $Acc(d)$ is the predicted accuracy of the classifier given by the regression model for the specific detection distance, A_B is another input parameter, which adjusts the value of g_h^i to the base desired level.

Over time, as the number of collected detection points increases, data about the performance of the individual classifiers is collected continuously and used to improve the statistical model and the accuracy of the classifier. The labelling of data is performed online by comparing the result of the individual classifier with the result given by the ensemble. Based on this comparison, the output of each classifier is labelled accordingly and added to the data training set of the model. Utilising the above mechanism, the model about the historical performance of the classifier evolves gradually to match the changes in the environment.

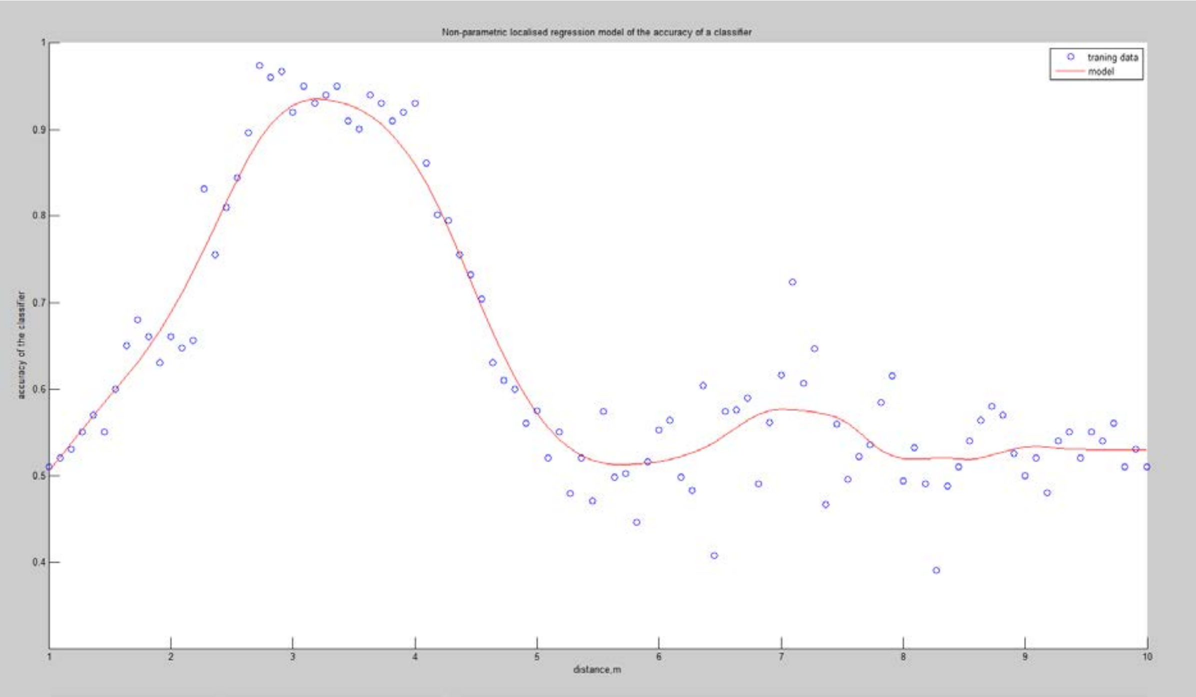


Figure 4-21: Typical distribution of the accuracy of a classifier over distance and the localised regression smoothing model

In addition, a Map of Prohibited Areas (MPA), g_o, a spatial pdf containing the areas where human presence is unlikely, is computed from the position of the obstacles on an occupancy map of the environment. MPA is then used as the third gating function transforming (4.10) into:

$$p(\Omega|\boldsymbol{X},\boldsymbol{\vartheta}) = \sum_{i=1}^{n_c} g_o(X).\, g_d^i(X).\ g_h^i(X).\, p(y|\boldsymbol{X},\vartheta_i)\,, \tag{4.13}$$

where $g_o(X)$, derived from MPA, results in very low probability in places where human presence is unlikely and relatively high probability in areas that human presence is possible.

Finally, the probability for presence of a human, $p(\Omega|\boldsymbol{X},\boldsymbol{\vartheta})$, is computed per cluster of detections, defined in 4.7.1, by applying (4.14) to all detection belonging to the cluster.

Finally, when overall ensemble the probability,$p(\Omega|\boldsymbol{X},\boldsymbol{\vartheta})$, reaches a certain pre-defined threshold P_{det}, it is identified by the method that a human detection is made at the position of the centroid of the cluster. In particular, the following condition is used to establish the presence of a human:

$$p(\Omega|\boldsymbol{X},\boldsymbol{\vartheta}) \;\geq\; P_{det}\begin{cases} true - positive\ human\ detection \\ false - negative\ human\ detection \end{cases} \tag{4.14}$$

Applying (4.14) sequentially, appropriate decisions are made about the human detections over the whole cluster of detections.

The position of the detected person is refined by applying the update step of the cluster's Kalman filter, described in 4.6, sequentially for each detection from the cluster. The position of the person reported by the filter's posterior is used internally by the ensemble detector as a reference position of the person.

4.8 Experimental results

The evaluation strategy is based on examination of the accuracy of the proposed human detectors, i.e. human leg and RGB-D human body detectors, as well as the evaluation of the improvement introduced by the proposed classifier fusion method in comparison with the individual detectors. This procedure is carried out by comparing the reported human detections with the ground truth, as reported by the motion capture system, recording the human motion sequences in parallel as described below.

4.8.1 Experiments with Single Classifiers

4.8.1.1 Evaluation of human leg detection from laser range finder data

The experiments were conducted in a simulated home environment, measuring 10mx7 m. Figure 4-2 shows the floor plan of the environment and the typical measurements of the LRF sensor when a person is present in the environment. An experiment was conducted to establish the error rate of the leg detection algorithm using a SICK 300 laser range finder, mounted on the robotic platform at 10 cm above ground and operating at 2Hz scan rate. In the experiment, five sequences of 60 seconds observations by the laser range finder were recorded. Each sequence consists of 120 scans.

The human legs in the scene during the experiments were detected in real time on a frame by frame basis by the human leg detection method, described in 2.1.3, using the laser range finder signal as input. The frames recorded were later analysed to build the confusion matrices of the leg detector for ten different ranges of the detection distance, i.e. distances starting from one metre with an increase of one metre. The function of the false positive (FP) and false negative rates (FN) rates over distance, shown in Figure 4-22 below, were used for training of the historical gating function of the ensemble classifier as described in 4.7.

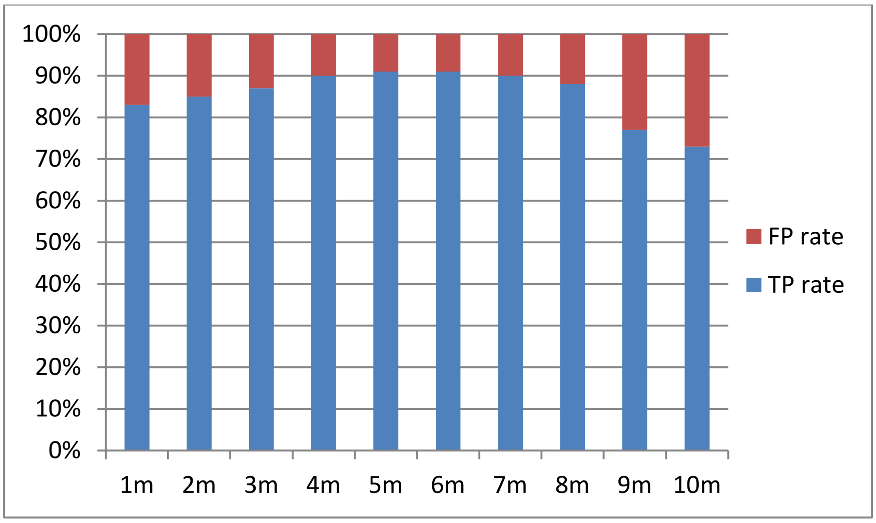


Figure 4-22: Performance of the leg detector as a function of the detection distance

Finally, by setting up the random forest classifier, used in the leg detector, to output the fraction of the class in the leaf, the probability of class membership of each class can was evaluated. The probability is matched against a variable threshold to construct the ROC function for the detector, shown in Figure 4-23.

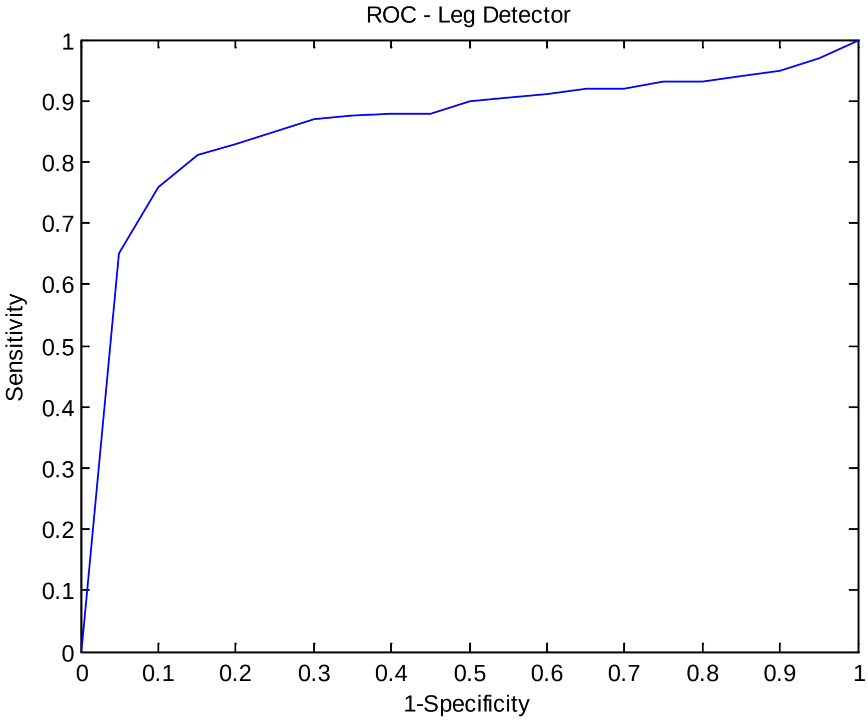


Figure 4-23: ROC curve of the leg detector

A tracker, based on the standard Kalman filter, was automatically associated with each detected person, as explained in Section 4.6. The output of the tracker was compared with the "ground truth" human position readings reported by the MOCAP system, which was recording the position of the human legs at a rate of 120 Hz. In total, five experiments were conducted. In the first one a single person was asked to walk in a straight line, in parallel to the direction of motion of the robot. The resulting accuracy, i.e. unsigned absolute error, over the recorded sequence for this experiment presented graphically in Figure 4-24.

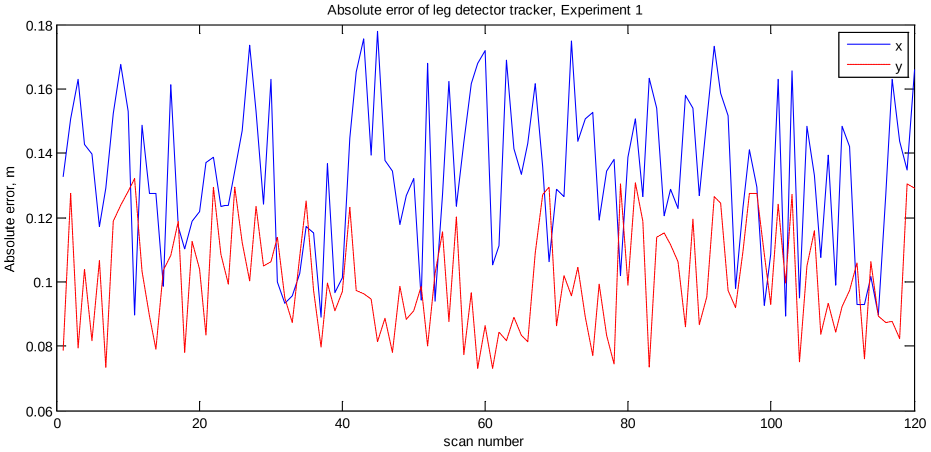


Figure 4-24: Evaluation of the error of the leg detector tracker

As seen from the above figure, slightly different accuracy was computed for the human detection in x and y directions. This observation can be explained with the much higher depth accuracy of the laser range finder, used in the experiments, in comparison with its angular resolution. The lower angular resolution resulted in a higher lateral error at longer distances, which, in this case, coincided with the x axis of the MOCAP system. This above effect is reduced when the error is averaged for several people walking randomly as demonstrated in the additional four experiments, conducted with varying number of participants. The evaluation of tracker error rates for these experiments is presented in Appendix F.

From the data collected from all experiments, the mean error and the standard deviation of the position reported by the tracker were calculated and are presented in Table 4-2.

		Mean error, m	Standard deviation, m
Experiment 1 (one person, far)	x	0.13	0.024
	y	0.10	0.017
Experiment 2 (two people, far)	x	0.13	0.037
	y	0.17	0.047
Experiment 3 (two people, near)	x	0.13	0.047
	y	0.13	0.0313

Experiment 4 (three people, near)	x	0.12	0.045
	y	0.15	0.037
Experiment 5 (three people, far)	x	0.12	0.025
	y	0.11	0.028
Overall evaluation (all experiments)	x	0.13	0.037
	y	0.13	0.041

Table 4-2: Experiment results – human leg detector

4.8.1.2 Evaluation of human body detection in RGB-D data

The proposed method for detection of human body in RGB-D data was evaluated by observing a typical home environment with a Kinect sensor and manually annotating a few of the resulting depth images with regions containing separate human body parts. Subsequently, after classification by the proposed algorithm, each pixel was assigned a label from one of the following classes by the algorithm: Head (dark red), Arm and Leg (purple), Thigh (light red), Flat Surface (green) and Unknown (black), as shown in Figure 4-25 below. Initially, the evaluation of the accuracy of the classifier was carried out by manually observing the colour annotated images.

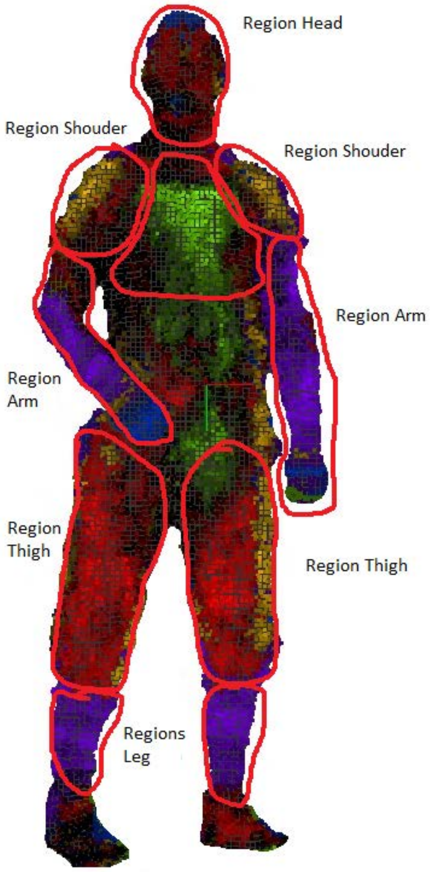


Figure 4-25: Evaluation of the RGB-D human body classifier (best viewed in colour)

As can be seen from Figure 4-25, the human body part classifier exhibits relatively good sensitivity properties by being able to detect and classify pixels from different body parts relatively accurately. However, it suffers from relatively low specificity, as it shows a tendency to misclassify parts of everyday objects as human body parts, a typical example of which is shown in Figure 4-26 below. This results in a relatively high number of false positive detections. Therefore, the human body detector is considered to be a weak detector and cannot be used independently for human detection. However, it contributes to the overall output of the combined classifier, described later in Section 4.7 .

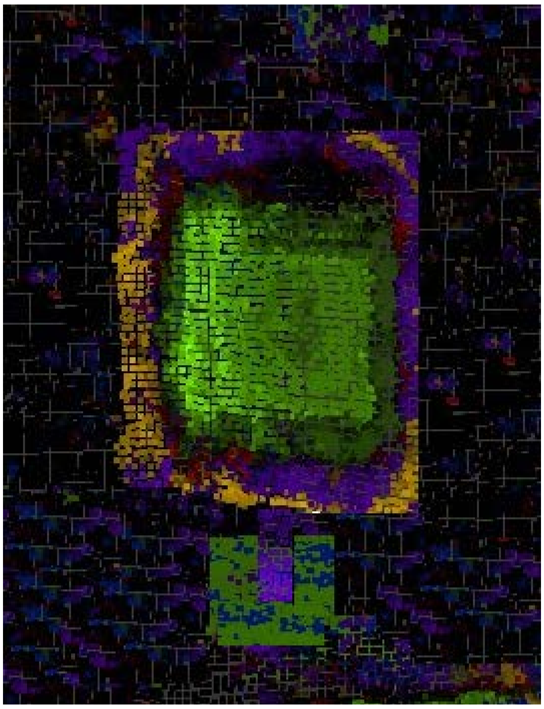


Figure 4-26: An image of a monitor wrongly classified as human (best viewed in colour)

Finally, after applying the voting criteria for human detection using the spatial distribution of the detected human body parts, described in 4.5.3, the detected human bodies were labelled manually by a human observer. By varying the threshold in the above voting criteria several confusion matrices of the RGB-D detector were constructed. Finally, the confusion matrices were used to construct the ROC curve for the detector, shown in Figure 4-27.

Given the fact that only a small set of depth images has been used in training of the classifier, i.e. 20 images in total, it is likely that the relatively low specificity is a result of insufficient training. It is considered that by substantially increasing the number of the labelled training images it is possible that a substantial improvement in the performance of the classifier can be achieved. Such an improvement will be investigated as a part of future work.

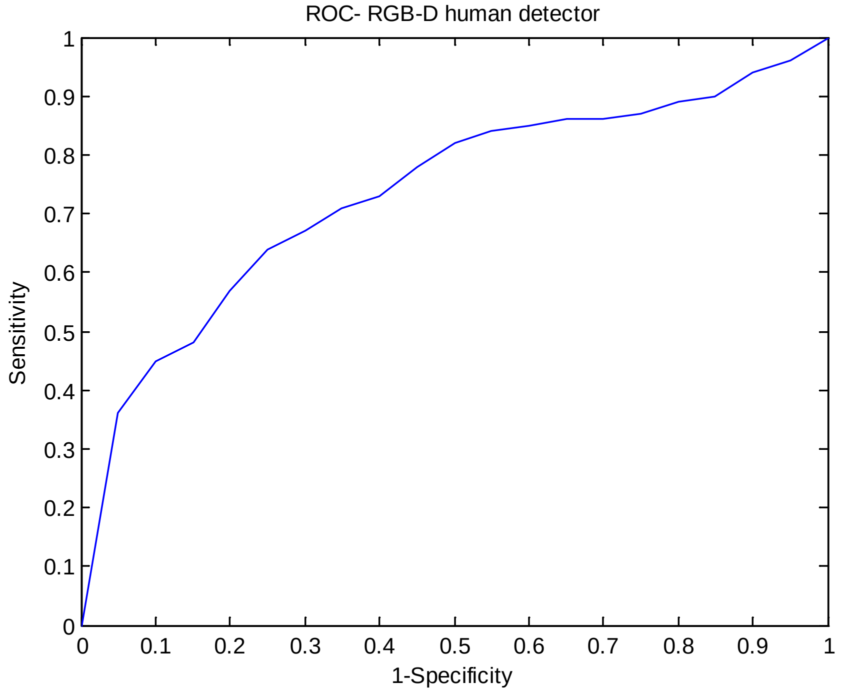


Figure 4-27: ROC curve of the RGB-D human detector

4.8.2 Classifier Fusion Experiments

The first experiment is designed to establish the characteristics of the ensemble classifier. In this experiment, a person was asked to walk randomly in the environment while data was collected by a laser range finder and a RGB-D sensor. The data was fed to the specific detectors which provided input to the ensemble classifier. The reported detections were compared manually with the ground truth human position, reported by the MOCAP system. If a detection was within a range of 0.25m from the real position of the individual, the detection was considered as a True Positive (TP). Otherwise, if the detection was outside the specified area, it was considered as a False Positive (FP). Through varying the P_{det} parameter of the algorithm, given in (4.14), the relevant confusion matrices were computed, which were later used for constructing of the ROC curve, shown in Figure 4-28 below.

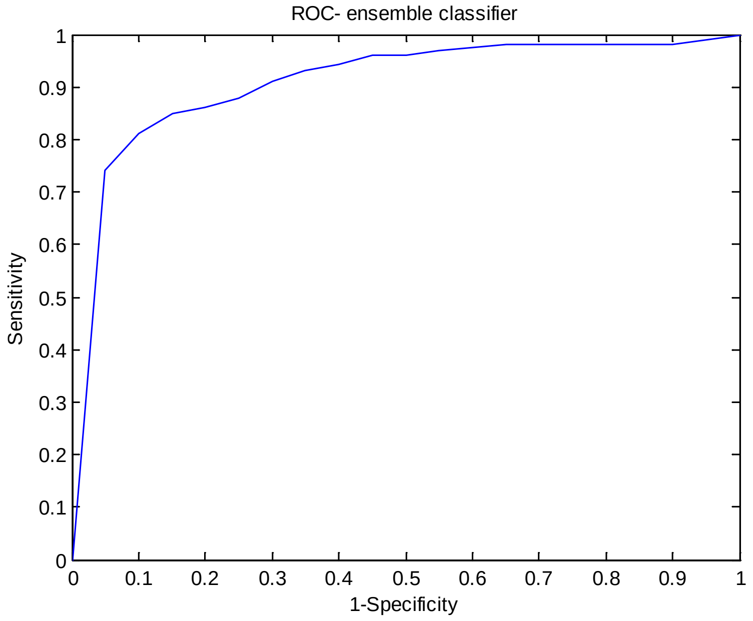


Figure 4-28: ROC curve of the ensemble classifier

Finally, the ROC curves of the three classifiers were compared by overlaying, as shown in Figure 4-29, to establish the improvement of the ensemble classifier over the individual human presence classifiers.

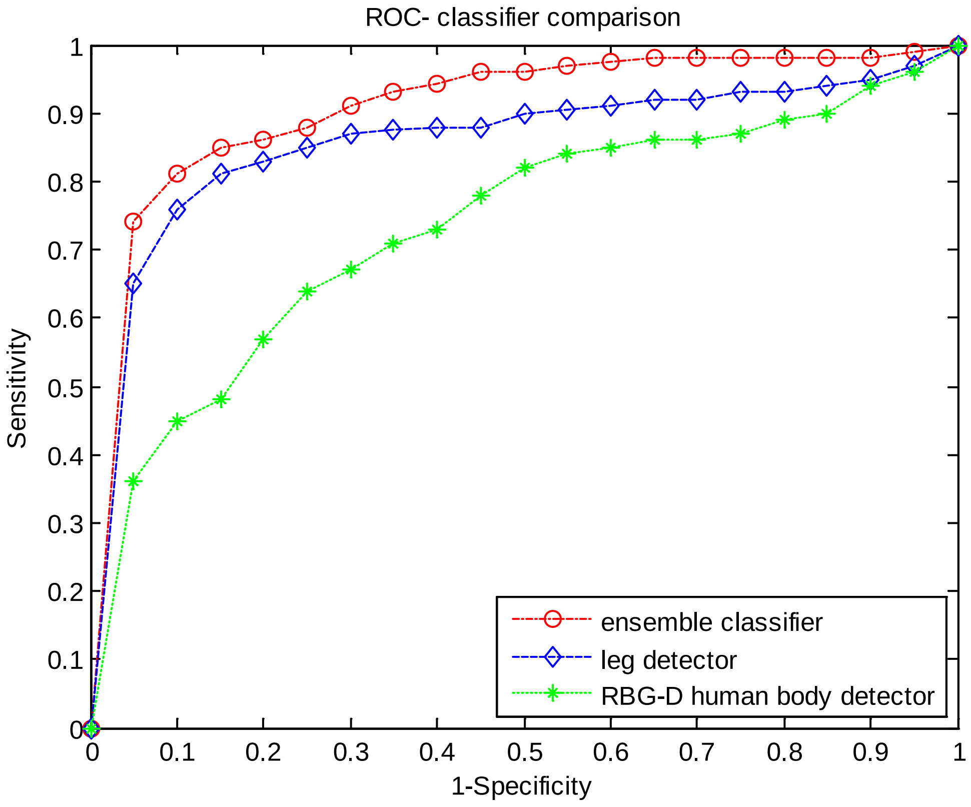


Figure 4-29: Comparison between the ROC curves of the classifiers

The resulting ensemble detector is tested by comparing its results with the ground truth data, originating from the MOCAP system. The error in the detection of a human is presented by the absolute error of the associated trackers.

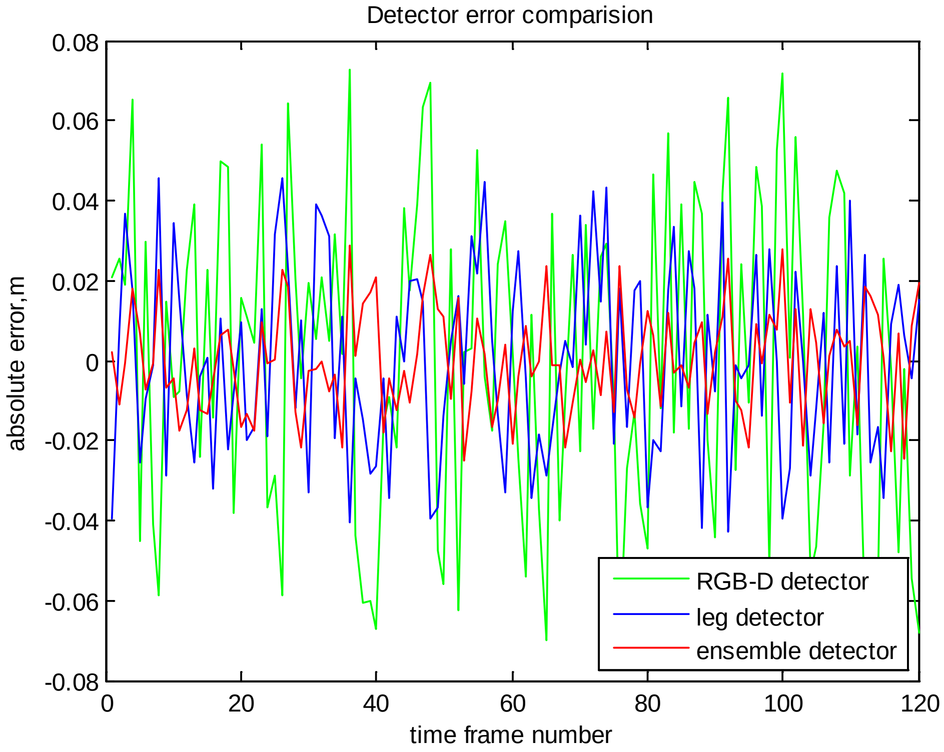


Figure 4-30: Error comparison between the leg, RGB-D and the ensemble detectors

Subsequently, a comparison is made by the distribution of the error of the leg detector, RGB-D detector and the ensemble detector as shown in Figure 4-31 below.

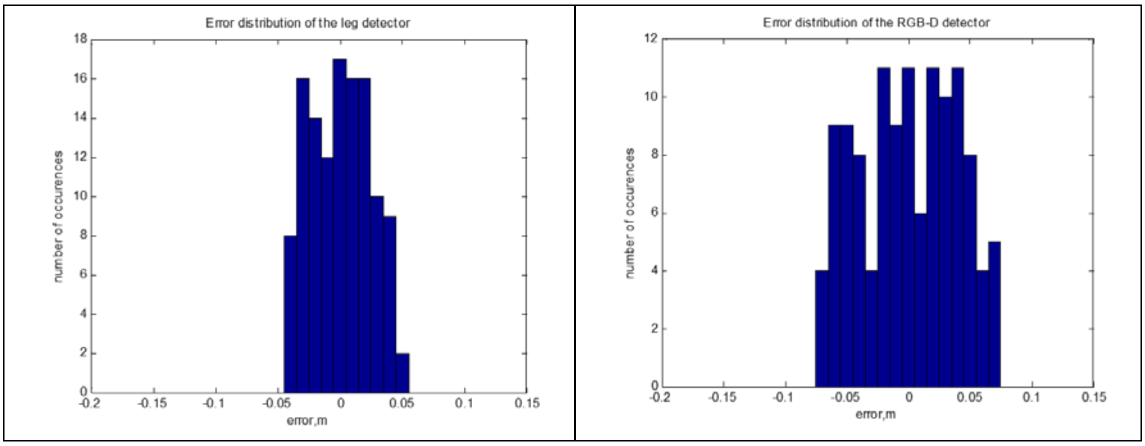


Figure 4-31: Comparison between distribution of the error of the leg and RGB-D detectors

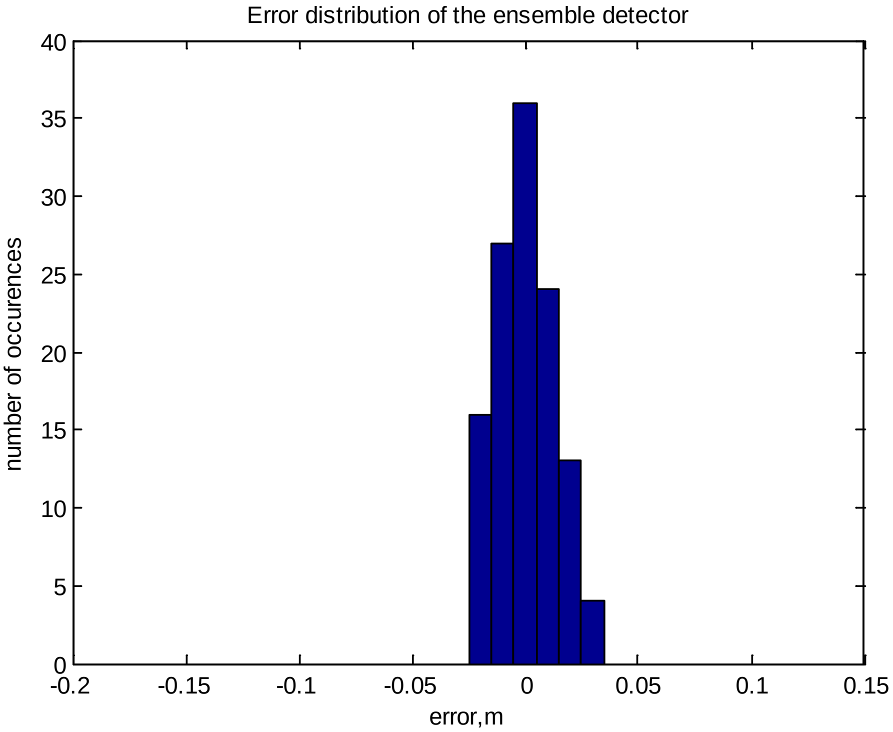


Figure 4-32: Distribution of the error of the ensemble detector

As seen from the above figures, the ensemble detector offers a narrower distribution of the error in comparison with the individual classifiers. The above observation is also confirmed by the standard deviation computed for each of the detectors, shown in Table 4-3 below.

	Standard deviation, m
Leg detector	0.024
RGB-D detector	0.037
Ensemble detector	0.016

Table 4-3: Experiment results – comparison between the standard deviation of the leg, RGB-D and ensemble detectors

Evaluation of the results

As seen from the above experiments, the ensemble detector improves human presence detection characteristics of the single modality detectors by using classifier fusion method that is based on learned historical parameters, related to the accuracy of the classifier over the detection distance range, and a dynamic measure, related to the temporal consistence of the particular detector. The above combination for the importance of the tracker guarantees stable performance both in the initial stage, when a new target enters the observable space, and in later stages when the associated trackers had converged and are able to provide a target tracking with a high confidence. As demonstrated by the experiment, using the ensemble detector results in considerable reduction of the standard deviation error in tracking of the target. For example, in the experiment, an improvement of approximately 30% was achieved by the ensemble detector over the performance of the RGB-D detector.

Additionally, by increasing the sensor redundancy and applying the proposed detector fusion method, it is possible to continue tracking with relatively good accuracy even when some of the sensors are temporary occluded. This feature is considered an important factor for increasing the resilience of the operation of a service robot, deployed at the elderly home, due to the unpredictability of the home environment.

The error in tracking can be further reduced by increasing the number of detectors and including new modalities, e.g. thermal, ultrasonic, that contribute to the overall performance of the human detection module.

4.9 Summary

Laser range sensing, depth and colour images belong to the richest sensory modalities, currently available for a robot to detect people. This chapter has addressed the problem of how to extract useful informational cues from a number of different modalities using sensors that are typically available on board a mobile robot. It also addressed how to combine the extracted cues using an ensemble classifier to achieve an improvement in classification accuracy over the accuracy of detectors operating on individual modalities. In addition, the above data fusion method brings an increased robustness in case of a transient occlusion to one or more sensors used in human detection. In addition, as demonstrated by the experiments, the proposed fusion method improves the performance of the best single classifier. In particular, the experiments with two detectors showed that the error of the most accurate detector, i.e. the leg detector based on a laser range finder, was reduced by 33% as seen from Table 4-3.

Although only two detectors were developed and tested in this chapter, the proposed framework for human detection allows addition of new human detectors, developed at a later stage. This enables a number of third-party detectors and sensors to be added in the future which will improve the performance and the redundancy of the human detection even further. For example, it is considered beneficial that an infrared thermal based human detector is added in future. This will allow heat sensitive detection of human bodies and will further increase the overall robustness of the human detection. Further, in Chapter 6, the detection of people and their location, reported by the methods in this chapter, are analysed by the methods proposed in Chapter 6 to estimate human tracks, which are used for planning of socially aware navigation paths.

In summary, several new methods and techniques have been proposed in this chapter aimed at improving the existing methods that are suitable for human detection and localisation. In particular the main contributions include:

Contributions

- a method for estimation of distortion in images that are the result of sequential scans by a laser range finder of moving targets, as described in 4.4.5;
- an algorithm for human leg detection from laser range finder data based on random forest machine learning, described in 4.4.7;
- a Human Interpretable Signature of Points (HISP) local feature descriptor and an associated method for human detection using HISP for human body shape detection, described in 4.5.2;
- a method for human body part detection through classification of HISPs, calculated from the point cloud dataset.
- a method for human detection through classifier fusion that combines several informational cues from different sensing modalities. The method is based on a three-layer gating procedure utilising learned historical knowledge, about the individual sensor detection characteristics, and the dynamic indicators about the temporal consistency of each detection in 4.7.

5 Human Pose Estimation and Tracking in 3D

5.1 Background

In service robotics, a closer interaction between robots and humans has the potential to improve the overall human experience by enabling closer, more natural and wider range of physical interactions between people and robots. Currently, any motion of the robot arm in proximity of people is avoided or severely restricted due to the strict passive safety rules, imposed to prevent harm to people. However, accurate and reliable human location and pose information is envisiged to enable the developers to focus on closer and more fine tuned HRI, instead on imposing rudimentary “no-motion near people” passive restrictions. In particular, such human information will enable more accurate arm planning with smaller safety margins, which is expected to lead to a range of new interaction patterns, like the direct passing of objects, physical human-robot collaboration and teaching by guiding of the arm of the robot directly by the human.

The human location information can be made more reliable by combining data from the available sensors through a classifier fusion method, as proposed in Chapter 4. However, for improved HRI, a reliable human pose estimation is also essential as this will enable operation of the robot’s arm near the person. This chapter addresses the problem of improving the reliability of human pose estimation and tracking.

In particular, human pose estimation deals with the problem of establishing the spatial locations of human body parts at the same time. Human pose tracking deals with the issue of continuous estimation of the location of body parts over time from a sequence of observations. There are many challenges making the reliable human pose estimation and tracking difficult. The biggest challenge is the high dimensionality of the kinematic model of the human body. For large redundant kinematic systems, as the human body, the formulation of motion, especially if body dynamics is considered, becomes a very broad and complex for optimisation problem. Moreover, self-occlusion, loose clothing and imperfect sensor properties, like noise and insufficient resolution, hinder the accurate dynamic pose estimation by introducing additional uncertainties in the detection process. A method addressing the above challenges is proposed bellow.

5.2 Problem Definition

The method proposed in this chapter uses 3D point cloud data, originating from the sensors of a typical service robot, to achieve improved tracking of the human pose.

Definition: A human pose tracking algorithm, that takes as input a dataset, made of point cloud measurements Y_t, from a monocular depth sensor, at each time t, $Y_t = \{y_t^j\}_{j=1}^{M_t}$, where M_t is the number of depth detections, at time t, and j, in superscript, is the index of the detections within the point cloud at time t, is required. The algorithm is required to produce an estimate of the human pose configuration state at each time t, $X_t = \{x_t^i\}_{i=1}^{N_t}$, where x_t^i is the i^{th}state variable at time t and i in superscript is the index of the state variables.

The first phase in tracking of the human pose is the modelling of the human body. The main target of the modelling process is to develop a simple but a sufficiently accurate model with a relatively small number of state variables that require minimal computational load for computation of state. Therefore, the following design goals have been identified to guide the development of the proposed human body skeletal (5.3.1) and appearance (5.3.2) models:

- Design Goal 1: Minimum number of parameters that are able to encode sufficiently accurately the human pose configurations in respect to the needs of adequate human-robot interaction;

- Design Goal 2: Human motion constraints embedded in the kinematic model to avoids additional computational procedures for motion constraint enforcement;

- Design Goal 3: Simplified rendering of surfaces – to allow visualisation and run-time pose optimisation at a reduced computational cost;

- Design Goal 4: Uncomplicated computation of derivatives for the positions of body parts in regard to the parameters – to allow less complicated mathematical procedures for pose optimisation in order to achieve fit with the measurements;

- Design Goal 5: Convenient motion concatenation of the separate articulated body parts, required for the forward kinematic computation.

The Global Local Articulated Interactive Closest Point (GLAICP) algorithm, proposed in this chapter, adaptively merges the local human pose configuration, computed through ICP, with the pose configuration resulting from the inverse kinematic computation of the most probable pose, which achieves a match in the associated key-point target pair. The result of the adaptive merging process is an increased resilience of the proposed GLAICP algorithm to temporary disturbances, like partial occlusions, in comparison with ICP. The GLAICP algorithm can be considered an extension to the articulated iterative closed point algorithm (AICP) (Mündermann et al.

2006;Pellegrini et al. 2008). In addition to the original AICP, the proposed method benefits from a mechanism for avoidance of local minima convergence problem.

The key idea behind the proposed method is to use an adaptive global optimisation procedure, guiding the conversion process of the local pose optimisation. The guidance mitigates the tendency of AICP to convergence to a wrong pose associated with the local minima. Overall, avoiding the local minima trap results in reduced probability of tracking failure and more robust tracking performance. In particular, the global optimisation part of GLAICP utilises a local features descriptor search in the point cloud to identify the target positions for several predefined key-points from the surface of the human body. Then, the method computes an optimal body configuration that brings the key points into the target position. Typically, as the search space of human pose configurations is enormous, due to the high number of variables, such a search process requires a long time to enumerate all the possible combinations. Instead, the proposed algorithm utilises inverse kinematics to compute the number of joint angles combinations that bring the kinematic skeleton into the optimal posture configuration, i.e. the one that minimises the distance between the key-points and the identified respective targets. As in some cases it is very likely that more than one pose configurations exist that satisfy the above condition for a pose match, a Method for Selection of the Most Realistic Pose Configuration (MSMRPC), proposed in 5.5.3, is used to compute the optimal pose configuration among all posible. The selection of the optimal pose configuration is based on criteria, representing simplified human motion dynamics and the human effort needed to achieve a pose, resulting in a match between key point and key point targets. Finally, the computed optimal pose is used to guide the local pose optimisation convergence process, as proposed by the mechanism for Guidance of Local Pose Optimisation, described in 5.5.4. Overall, the method is based on adaptive merging of the global and local pose configuration adjustments, executed iteratively for each linear segment of the kinematic chain until a convergence is reached.

It is claimed that, due to the increased robustness of the proposed GLAICP, in critical HRI applications, like those requiring close physical interaction, potentially, it improves the safety for the human by providing a more reliable human pose tracking.

The main contribution of this chapter is the GLAICP algorithm. An additional contribution is the proposed CHISP descriptor (5.5.6), which is a generic geometric local feature descriptor that can be utilised in a broad range of point cloud related applications.

5.3 Kinematic Parameterisation

5.3.1 Kinematic Skeletal Model

The human body is a highly complex mechanical system of bones, muscles and soft tissues. The joints are connecting the main bones defining the human pose configuration. The complexity of the human body, which without simplifications can reach a significant number of degrees-of-freedom (DOF), comes mainly from the fact that the joints are not rigid structures. For example, in addition to the rotation, a cartilage can compress and expand, ligaments can stretch, thus transforming a simple ball-and-joint 3 DOF joint into a complex joint with 6 DOF. Additionally, joints normally do not rotate about a fix set of orthogonal axes but around axes which change orientation as a complex function of the pose (Catani et al. 1996). This motion complexity adds to the complexity of the overall human body kinematics.

In the context of personal robotics, achieving sub-millimetre precision of the human pose tracking is not essential in service robotics. Choosing a simplified model, over a very accurate, but complex, one, reduces the tracking precision to a small degree. However, it enables the human pose tracking using only the limited computation resources of a mobile robot. At the same time, an over-simplified model could not represent all body pose configurations with sufficient accuracy. It is also likely that a simple model can restrict the ability of the tracking algorithm to reach a convergence and produce a realistic pose. Therefore, a compromise between a complex and an over-simplified model of the human body is considered an appropriate solution for the purpose of human pose tracking as explained below.

As a result of the investigation of a number of human locomotion models, e.g. (Brubaker & Fleet 2008; Kuo 2001), the model that meets the above requirements and objectives, shown in Figure 5-1, is selected and further adapted to the needs of GLAICP. In this model, the human body is represented by eleven joints, encoded as local coordinate systems, and thirteen linear elements with fixed lengths, representing the bones in the human body. The default lengths of the linear segments are selected based on average values reported in anthropomorphic studies, as explained below.

Enforcing limited motion freedom in the skeleton joints is considered a rather computationally demanding task, which requires additional modelling of the relative periaxial rotation of the human bones and the non-rigidity of the joints. In particular, modelling of the non-rigidity properties of a joint requires a non-trivial mathematical representation of the freedom of motion in the soft joint tissues. Using the above approximation disregards the effect of the soft tissues and enables representing all joints with as having only the full 3 DOF. The simplification reduces the amount of calculations needed for pose tracking, through applying the same unified computational procedure when concatenating the motion for sequential joints. The undesired effect of enabling the full 3 DOF motion in all joints is that some unnatural human poses become possible. These unrealistic pose configurations, however, do not pose a significant problem for tracking, as the later procedures of the GLAICP reject them as not matching the point cloud data.

The pelvis position is used as the main reference point in body tracking, i.e. the root of the skeletal model that links human position to the outside world. The link between the room coordinate system and the pelvis position is denoted by a global transform, T, and a global rotation, R_g. The position of the pelvis defines the base coordinate system of the human skeleton in which the pose configuration variables, i.e. the state of the kinematic skeleton, are recorded.

There are different design options for the parameterisation of rotations of the joints of the skeletal model. These include: rotation matrices, Euler angles, quaternions and axis-angle representations. The most suitable option for the GLAICP algorithm is the axis-angle representation, which, unlike quaternions, requires only three parameters ϑw to describe a rotation and does not suffer from the gimbal lock (Moeslung et al. 2011). A brief description of the angle-axis representation, used both for the forward kinematic chain and inverse kinematic chain calculations in GLAICP, is given below. In particular, the axis - angle representation of a rotation is necessary in order to simplify the calculations in the global optimisation part of the GLAICP. This representation uses an axis in space, specified as unit vector $\omega \in \mathbb{R}^3$ and an angle of rotation in radians ϑ, where ϑ determines the amount of rotation about ω.

Every rotation in GLAICP is also represented by its exponential form, based on the axis of rotation ω and the angle of rotation ϑ:

$$R(\omega, \vartheta) = e^{\vartheta\widehat{\omega}} , \tag{5.1}$$

where $\widehat{\omega} \in so(3)$, is a skew symmetric matrix, i.e. satisfies $\widehat{\omega}^T = -\widehat{\omega}$, which is constructed from ω using the wedge operator, ^, given the vector of the 3D scaled rotation axes $\vartheta\omega = \vartheta \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix}$, as:

$$\boldsymbol{\vartheta}\widehat{\omega} = \boldsymbol{\vartheta} \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}, \tag{5.2}$$

where $\omega_1, \omega_2, \omega_3$ are the unit rotation axis.

In (5.1), $\vartheta\widehat{\omega}$, is the matrix representation of a twist (Murray et al. 1994**).** The exponential form, $\boldsymbol{e^{\widehat{\omega}\vartheta}}$ is computed easily using the Rodrigues' formula given in (5.3). It requires only the square of the matrix $\widehat{\omega}$, $\sin(\vartheta)$ and $\cos(\vartheta)$ to calculate the exponential form of the rigid motion (Murray et al. 1994) :

$$e^{\widehat{\omega}\vartheta} = I + \widehat{\omega}\, sin(\vartheta) + \widehat{\omega}^2(1 - \cos(\vartheta)) , \tag{5.3}$$

where I is the identity matrix. Equation (5.3) provides an efficient way for GLAICP to compute the exponential for of rotation from the angle θ and the axis w.

If a translation is required, e.g. whe calculating the position of. the root element of the kinematic chain , then this is achieved using the vector $\theta v \in \mathbb{R}^3$, which represents the translation along the axis of the rotation. Finally, the six parameters representing a twist $\vartheta\xi \in \mathbb{R}^6$ are given as:

$$\vartheta\xi = \vartheta(\omega_1, \omega_2, \omega_3, v_1, v_2, v_3) \tag{5.4}$$

Analogously to (5.2), the twist action $\vartheta\hat{\xi} \in se(3)$ is constructed in GLAICP from the twist coordinates $\vartheta\xi$, given in (5.4), and using the wedge operator (Murray et al. 1994):

$$[\vartheta\xi]^\wedge = \vartheta\hat{\xi} = \begin{bmatrix} 0 & -\omega_3 & \omega_2 & v_1 \\ \omega_3 & 0 & -\omega_1 & v_2 \\ -\omega_2 & \omega_1 & 0 & v_3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{5.5}$$

For a full 6 DOF joint, i.e. the root joint in the skeleton, i.e. joint R_0 in Figure 5-1, all six twist parameters are required. However, for all other joints in the human body, the required number of parameters is lower due to their restricted degrees of freedom. In GLAICP, where only ball joints are used, capable only of rotation but no translation motion, only three parameters are required to encode the motion of each joint, therefore, it is modelled as a twist with a known joint location and an unknown rotation $\vartheta\omega$ (Moll & Rosenhahn 2009).

The state vector of the human pose in GLAICP consists of 3 parameters for each ball joints, i.e. joints $\mathrm{R_i}$, $\mathrm{i} = 1,..,10$, shown in Figure 5-1, and six parameters for the root joint $\mathrm{R_0}$, and is given as a vector X:

$$X = \\ (w_1^0, w_2^0, w_3^0, v_1^0, v_2^0, v_3^0,\ \theta_1^1, \theta_2^1, \theta_3^1, \theta_1^2, \theta_2^2, \theta_3^2, \dots, \theta_1^{10}, \theta_2^{10}, \theta_3^{10}) \tag{5.6}$$

The motion along the kinematic chain is presented as concatenation of a series of rigid motions over the joints, involved along the chain, resulting in a concatenation of the twist transforms for the joints. This representation, expressed with twists associated with the respective joints, results in the following equation for the spatial coordinates of an arbitrary point in the body as a function of the joint angles in the chain before the point (Moeslung et al. 2011):

$$p_a = \bar{p}_s(\xi_1, \xi_2.., \xi_n) = e^{\hat{\xi}_1} e^{\hat{\xi}_2} \dots e^{\hat{\xi}_n} G_{init} \bar{p}_b\ , \tag{5.7}$$

where n is the number of joints, ξ_1, ξ_2, …, ξ_n are the twists associated with the joints along the kinematic chain, ϑ_1, ϑ_2 … , ϑ_n are the angles associated with the joints and G_{init} is the rigid transformation at the initial pose. The equation (5.7), known in robotics as forward kinematics map (Yiu & Li 2003), is used in GLACP for computation of the position of key-points of the human body as explained later.

5.3.2 The volumetric appearance model

The volumetric appearance model, also further referred as the appearance model, is computed from the skeletal model by replacing the linear elements between joints with geometric primitives, i.e. cylinders with the appropriate size. The above simplified representation guarantees fast computation of the body surfaces. The distances between the body surfaces and points from the point cloud are then used for computation of the amount of fit between the model and the data in the local optimisation part of GLAICP. In particular, the ecliptic cylinders used for body parts with a non-circular cross-section, e.g. human body trunk, allow a relatively accurate representation of the shape of these parts at a very low computational cost. Moreover, due to the identical base areas of the elliptical cylinders, only three parameters are required to define each cylinder. These parameters are: the lengths a, b of the axes of the ellipse of the base area and the height of the cylinder, l. The above parameters of the appearance model are specific to each individual person and, once established, they stay constant during the tracking process. Default approximate values, determined from anthropometric data described in 5.3.3, are used, if no individual values are provided during the initialisation phase. In particular, the matrix $A = (P_1, \dots, P_n)$, $n = 12$ is used to store the parameters of the appearance model of the human body, where $\boldsymbol{P_i}$ is the column vector, $\boldsymbol{P_i} = (a_i, b_i, l_i)^T$, that denotes the parameters of the ith segment in appearance model. Currently, the system relies on initial calibration. However, it is possible that stored specific values for each person are retrieved after identification of the person, e.g. a face detection. As future work, an improvement of the algorithm will be investigated to allow dynamic adaptation of the human body parameters during tracking. This will remove the need for initial calibration of the tracking system.

Although the above appearance model may be considered rather simplistic for the realistic rendering of a person's body surface, it provides sufficient accuracy for the purpose of estimation of the extent of fit between the human body and the surrounding point cloud. Moreover, with a simplified model, the computation of the distances to a simple geometric shape is done more efficiently, which enables real time operation of the method.

Both models, used in GLAICP, i.e. the kinematic skeleton model and the appearance model, are depicted in Figure 5-1 below. On the left, the articulated human body model, used for in the forward and inverse computations of the pose configuration, is shown. On the right, the volumetric appearance model, used for computation of the extent of matching of the pose configuration to the point cloud, is shown. In the articulated model, the joints and their coordinate systems are marked with the symbols: $\{R_1, R_2, R_3, \dots, R_{10}\}$. The linear segments, representing the bones of the human skeleton are marked with the numbers {1,2,3,..., 10}. Typical lengths for the linear elements are used in the method as explained below in 5.3.3 . In the volumetric appearance model, an average mass measure is associated with each linear segment, representing the inertia of the human body parts. This measure provides more realistic tracking by preventing any sudden motion from happening as the appearance model is matched to noisy point cloud data.

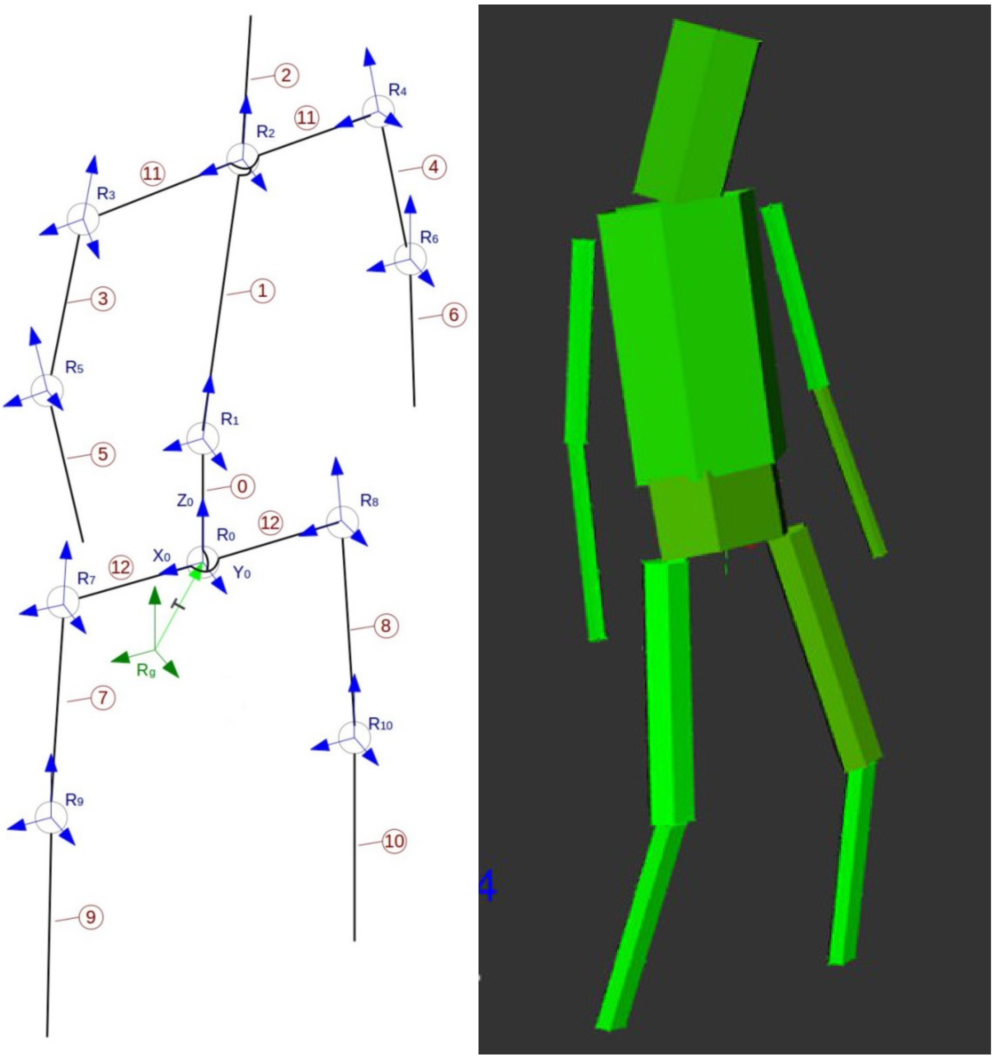


Figure 5-1: The simplified models used by the human pose estimation subsystem – left side, the articulated human body model; right side, the appearance model

5.3.3 Anthropometric considerations

The linear elements between the joints in the skeleton, representing the bones of the human body, have an important role in computation of the pose tracking. They, together with the joint angles, are part of the rigid transforms matrices that define the pose configuration. However, unlike the joint angles, the linear elements, stored in the vector P_b, are constant and need to be estimated only once during initialisation. In particular, appropriate values are used for the initialisation of the vector $\boldsymbol{P_b}$, based on the average reported human size and weight in the anthropometrics studies (de Leva 1996), listed in Table 5-1 below.

Ref. Number	**Segment description**	**Female,**	**Male,**
(refer to		**length**	**length**
Figure 5-1)		**cm**	**cm**
12	Hip	19.10	19.70
0	Lower Trunk	10.46	8.37
1	Upper Trunk	24.97	26.29
11	Shoulder	17.10	18.20
2	Head and Neck	14.05	13.95
3,4	Upper Arm	15.86	16.18
5,6	Forearm	14.50	15.45
7,8	Thigh	21.24	24.25
9,10	Calf	24.92	24.93

Table 5-1 Average sizes of human body segments

5.4 ICP and its Generalisation for Articulated Bodies

As remarked, the proposed GLAICP represents an extension to the Articulated Interactive Closest Point (AICP) algorithm (Grest et al. 2005; Plankers & Fua 2003; Demirdjian et al. 2003). This extension increases the robustness of the pose tracking to temporary disturbances like transient occlusions and reduces the tendency for convergence to a local minima instead of the correct pose. As AICP itself is based on the ICP algorithm (Besl & McKay 1992), a short analysis of strengths and the weeknesess of both ICP and AICP, provided below, is considered necessary to justify the proposed extension.

The ICP method (Besl & McKay 1992) is a computationally efficient method for registration of 3D shapes. It is based on the minimisation of the distances between the points belonging to two separate point clouds. It is especially efficient when the nearest neighbour search, used internally, is based on the kd-threes method (Bentley 1975). The main issue with the ICP algorithm is that it always converges monolithically to the nearest local minimum of a mean-square distance metric. According to its definition (Besl & McKay 1992), ICP registers a data shape P to be in the best alignment with a model shape X. Although the initial data and the model shapes can be represented in one of the many allowable forms, including 1) sets of points, 2) sets of line segments, 3) sets of parametric curves, 4) sets of implicit curves 5) sets of triangles 6) sets of parametric surfaces and 7) sets of implicit surfaces, that have to be decomposed into point sets prior to processing.

The ICP algorithm, in its original formulation, consists of four steps executed iteratively until convergence within tolerance τ is reached (Besl & McKay 1992):

- For every point in P_k $c$ompute the closest points yielding a set of correspondences Y_k, $Y_k = C(P_k, X)$, where index k denotes the time frame, C is the closes point operator;

- Compute the local registration $\vec{\mathrm{q}}_{\mathrm{k}}$ that minimises the summed square distances between the points in $\mathrm{P_k}$ and their correspondents in $\mathrm{Y_k}$;

- Apply the registration: $\mathrm{P_{k+1}} = \vec{\mathrm{q}}_{\mathrm{k}}(\mathrm{P_0})$, where $\mathrm{P_{k+1}}$ is the resulting data shape point set at time $\mathrm{k} + 1$, $\vec{\mathrm{q}}_{\mathrm{k}}$ the registration state vector calculated in the previous step, $\mathrm{P_0}$ is the initial data shape point set;

- Terminate the iteration when the change in the mean-square error falls below the pre-set threshold $\tau > 0$: $d_k - d_{k+1}$, otherwise repeat from step 1.

At the end of the convergence, the algorithm returns a rigid body transformation, $\vec{q}_k$, that gives an optimal fit of the data shape P to the model shape X.

Due to the efficiency of the algorithm, many different variants have been proposed (Rusinkiewicz & Levoy 2001). In particular, several closed form solutions have been reported for calculation of the local registration in step 2 of the above algorithm. Most commonly they include singular value decomposition (Arun et al. 1987) or a representation in quaternion space (Benjemaa & Schmitt 1998). However, all of these solutions are restricted only to rigid bodies and, therefore, not directly applicable to the problem of human pose estimation.

Although ICP offers an efficient solution for registration of 3D shapes, by definition, it is unsuitable when any degree of change in the shapes is present. Typical examples for such a change include deformable shapes, high level of outliers or discretisation errors (Segal et al. 2009). As these are exactly the problems that the typical low-cost depth sensors, used in today's service robots, as analysed in 4.5.1, exhibit, ICP is not directly applicable to solving the problem of human pose tracking.

As an alternative approach, a collective optimisation of all pose parameters that guarantees a minimal alignment error is a possible. For example, such a collective optimisation can be achieved through the Levenberg-Marquardt (LM) optimisation (Dewaele et al. 2006). However, this approach results in a non-closed form solution, which cannot be solved directly. In addition, the above approach and other similar approaches based on local optimisers suffer heavily from the local minima problem.

The articulated ICP (AICP) (Pellegrini et al. 2008), a generalisation of the original ICP for articulated bodies, divides the articulated human body into rigid segments that are then aligned using the original ICP. The approach, proposed in this work, is based on the AICP method for local optimisation of the pose. However, unlike AICP, the proposed approach uses a global pose optimisation method in parallel to the local pose optimisation to achive improved convergence. In particular, the proposed algorithm has improved robustness to temporary disturbances in the convergence process, e.g. occlusions and reduced tendency to converge to the local minima. Overall, the combination of global and local pose optimisations benefits the human tracking by preserving the simplicity and the efficiency of the original ICP while reducing its tendency to converge to the local minima.

5.5 Global-Local Articulated ICP

The proposed Global-Local Articulated ICP (GLAISP) algorithm relies on a combination of alternating global and local optimisation cycles to guide the kinematic model to the optimal human pose configuration. The key idea behind the GLAICP is to use an inverse kinematics mechanism to compute the body configuration change needed to bring the appearance model into alighment with the data, i.e. a configuration that minimises the sum of squared distances between points from the model and the masurement data.

Initially, the method searches for a number of known key point signatures in the point cloud data and, if fount, it matches them with their positions from the surface of the human body. Then, using inverse kinematics, it computes the optimal pose configuration that brings the model in alignment with the data. The new pose configuration is then used as guidance to the convergence process of the local iterative pose optimisation, significantly reducing the risk of the local minima problem in the convergence between the model and the data.

The proposed method addresses the three main shortcomings of the Articulated ICP (AICP) (Grest et al. 2005; Plankers & Fua 2003). Firstly, as AICP searches only the nearest neighbourhoods to establish local correspondences it exhibits a tendency to converge to the local minima. Secondly, it is possible that the human pose, due to some disturbance, e.g. occlusion or noise, gets into a wrong state that is very distant from the data in the point cloud. The resulting big displacements make it impossible for any appropriate associations to be established. Then, as the final state in each time frame is used as the initial state for the next frame, the inability to establish local correspondences in one time frame makes it even less likely to happen in later time frames. The accumulation of error leads eventually to an unrecoverable tracking failure due to inability of ICP to establish local point associations. The recovery from a failure can be achieved only by re-initialisation of the tracking process. Finally, ICP is not able to achieve an instantaneous match between the skeletal model and the measurements. Instead, it relies on numerous converging iterations, each bringing the model a step closer to alignment. For complicated non-rigid models, like the human body, the iterations take place in sequence along the kinematic chain and require substantial time to converge. If in meantime the dataset changes significantly, e.g. fast limb motions, the convergence is unable to keep up with the change and loses the local correspondences, which results in a tracking failure.

The proposed method reduces the shortcomings of the AICP by providing global guidance to the convergence process. In particular, the guidance is computed from the global pose prediction, based on a key point signature matching in the point cloud. After a satisfactory match is found between a pre-selected key point and a point from the point cloud, referred to as a target point, the optimal kinematic chain configuration, i.e. the one that achieves alignment between both is computed through inverse kinematics. Then, the resulting pose configuration, i.e. the guidance pose, is used to guide the local adjustments to achieve the correct alignment of the linear elements from the kinematic chain. The guidance pose, computed in each subsequent time frame, has three effects on the pose convergence. Firstly, it drives the convergence process away from the local minima. Secondly, it acts as a partial initialisation at each time frame,Finally, it accelerates the convergence process by providing a pose configuration that is sufficiently close to the correct pose, reducing in this way the number of iterations needed. These three effects counteract directly the deficiencies of AICP.

In particular, the key operation principle of the proposed GLAICP is based on provision of adaptive guidance to the local optimisation process. The amount of the guidance depends on the ratio between the particular key point – target distance and the extent of fitting between the appearance model and the local point set. The rationale for the introduction of the adaptive guidance is to allow stronger global pose influence in the initial phase of the conversion process as well as when the local matching fails, i.e. the number of local association between points from the model and measurement data set falls below a predefined threshold. In the first case, a stronger global influence in the initial stages of the process is intended to accelerating the convergence. In the second case, it the global pose guidance increases the resilience of the convergence process to any transient disturbance. On the opposite, when the distance between the key-points and targets becomes relatively small, as this happens in the final stages, the global influence is reduced proportionally to allow final adjustments to the pose to be carried out only by the local optimisation. The reason for the reduced influence is that the risk of a local minima problem is much lower in the final stages of the process. Also, the only the local pose optimisation cycles alone can achieve sufficient accuracy of convergence.

Overview of the local and global cycles of GLAICP is depicted in Figure 5-2. During initialisation, the trunk of the human body is first positioned in space by using initialisation information from the human localisation method proposed in Chapter 4. Then, the position of the human body trunk is aligned through ICP with the measurements from the point cloud. The above initialisation process positions the pelvis, i.e. the root element of the four kinematic chains, representing the four human limbs, in a fixed position relative to the room coordinate system. After the initialisation is finished, a separate GLAICP process is started for each of the limbs in parallel. In each GLAICP process, the global optimisation cycle guides the local optimisation through the corrections to the joint angles adjustment as described below.

After a number of key points are matched to their target positions, using a search for known CHISP descriptor signatures in the point cloud, as described in 5.5.6, they are used as a basis for computation of several possible guidance pose configurations. After ranking of the above pose configurations for suitability, using the criteria described below, one of them is selected to be the guidance pose. Subsequently, using the selected guidance pose, the local and global angle adjustments are computed and merged into a single correction for each segment from the kinematic chain, as described in 5.5.4. The segment position is updated according to the calculated correction. The process is repeated for each subsequent segment until the end of the kinematic chain is reached. At the end of the kinematic chain, the alignment between key points and targets is re-accessed by computing the alignment measure, described in 5.5.1. The above process is repeated several until the alignment measure falls below the predefined threshold. If successful convergence cannot be achieved within a predefined number of cycles, the next guidance pose in the above ranking, is selected. In this way, the process is repeated until a satisfactory alignment is reached or a timeout occurs.

In summary, as seen from the high-level overview of the GLAICP cyclic process depicted in Figure 5-2, in each time period a number of alternating local and global optimisation cycles are executed. In these cycles, the guidance to the local pose optimisation is provided in the form of joint angle corrections, computed from inverse kinematic pose adjustment needed to bring a number of key points into alignment with their identified target positions in the point cloud.

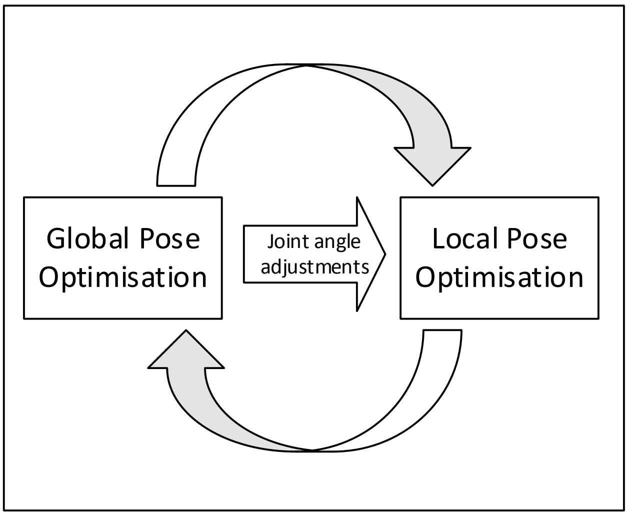


Figure 5-2: Iterative global and local cycles in GLAICP

5.5.1 Alignment Measure

The overall goal of GLAISP is to minimise the error function, $\mathrm{e}\colon \mathbb{R}^3 \mapsto \mathbb{R}$, which is defined in Euclidian space as the sum of squared re-projection errors between the i$^{\mathrm{th}}$ key point position on the skeletal model $\hat{\mathrm{d}}_{\mathrm{i}}(\mathrm{x}_{\mathrm{t}})$ and its target position $\tilde{\mathrm{d}}_{\mathrm{i}}$, as identified in the point cloud:

$$\boldsymbol{e}(\boldsymbol{x_t}) = \sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{e_i^2}(\boldsymbol{x_t}) = \sum_{\boldsymbol{i}}^{\boldsymbol{N}} \left\| \widehat{\boldsymbol{d}}_{\boldsymbol{i}}(\boldsymbol{x_t}) - \widetilde{\boldsymbol{d}}_{\boldsymbol{i}} \right\|^{\mathbf{2}} \tag{5.8}$$

In Figure 5-3 below, one of the key- point- to-target transformation, $\widehat{\boldsymbol{d}}_{\boldsymbol{i}}(\boldsymbol{x_t}) \mapsto \widetilde{\boldsymbol{d}}_{\boldsymbol{i}}$, is illustrated with the pair consisting of the key point p_A and its target position p_B. It is possible that more than a single pose configuration exists that brings point p_A into alignment with point p_B. Therefore, in GLAICP, the pose configurations are ranked for suitability, as described in 5.5.3, and the pose with the highest score is used as a guidance pose. If computing time and resources allow, e.g. accelerated computation using parallel computing, several possible pose configurations, are accessed and the one that achieves the lowest error $\boldsymbol{e}(\boldsymbol{x_t})$ of alighment is selected as the final guidance pose.

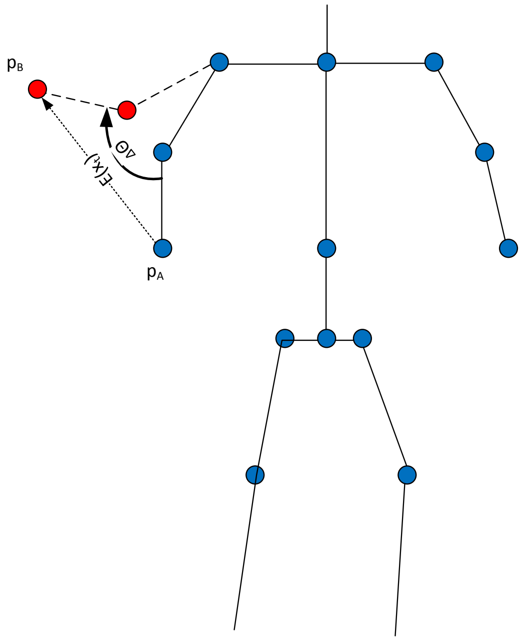


Figure 5-3: Achieving alignment between a key point p_A and its target position p_B

5.5.2 Inverse Kinematics Computation

After finding the target position of a pre-defined key point for the particular limb, the GLAISP algorithm computes the changes needed for the current configuration of the skeletal model to achieve alignment between the key point and its target position. A similar problem, known as inverse kinematics, has been investigated in robotics extensively. The inverse kinematics computes the required changes of configuration of the robot's kinematic chain that guarantee that the end effector reaches the desired position (Richard 1981). Because of the similarity between the problem of computing the correct skeletal pose and the robot's manipulator, IK is adopted for use in the proposed method. Unlike Forward Kinematics, which manipulates the whole structure sequentially, IK operates directly through solving the kinematics equation.

In GLAICP, IK is applied for each of the four kinematic chains in the human skeleton model, to compute the likely pose configurations that bring the key point- target pair in aliment for the particular kinematic chain. Later, the above configurations are ranked using the proposed mechanism for selection of the most realistic pose. Based on the results, the most appropriate human pose configuration is selected to be used to guide the convergence process. The process of evaluation of several pose configurations is suitable for parallelisation which could provide substantial acceleration.

The goal of both IK and the optimal guidance pose selection parts is to find the most probable state of the kinematic model that brings the known key point positions of the human body surface as close as possible to the identified target points. The computation uses the given parameters of the skeletal model of the human body, its current state and the associations between resulting local features and key points on the surface of the appearance model. For a single key point, $\mathrm{p_A}$, and a single identified local feature, a target $\mathrm{p_B}$ the above goal is defined as:

$$arg \min_{\vartheta_1,\dots,\vartheta_n} \left\| e^{\hat{\xi}_1\vartheta_1} e^{\hat{\xi}_2\vartheta_2} \dots e^{\hat{\xi}_n\vartheta_n} G_{init}\bar{p}_A - p_B \right\|^2 , \tag{5.9}$$

where $\vartheta_1, \dots, \vartheta_n$ are the joint angles in the kinematic chain, $\hat{\xi}_1$,... $\hat{\xi}_n$ are the associated rotational twists and G_{init} is the transformation between the root coordinate frame and point p_A at zero pose.

For simplification, the kinematic chains and the key-point positions are specifically selected to avoid the multiple objective optimisation problem, arising when several key points per chain have to be matched with their identified position in the point cloud and the distances between them to be minimised. In particular, only one key point per limb, positioned at the end of the kinematic chain, i.e. on the fingertips, is used to avoid any cross-interference when the kinematic chain state is optimised. Additional key points can be used as a backup when the main one cannot be located, e.g. due to occlusion. In the worst case, i.e. when no local feature signatures can be matched to the a pre-stored key point signature, or the certainty in the match in the signatures is not sufficient for a positive association between a key point and a stored signature to be made, the IK computation part is skipped in the current time frame and the algorithm reverts to the standard AICP. Similarly, when more than one local signature matches the pre-recorded signature of a key, the optimal pose configuration is not computed to reduce the risk of using a wrong guidance pose.

Inverse Kinematics Part

There are several different possible approaches possible for solving the above IK task. One possibility is to use the closed form or an analytical solution. Although this is a fast method that returns a precise result, its drawbacks include increased complexity for a high number of DOF in the kinematic chain. Another downside of the close solution approach is the lack of any result when the target is unreachable. This behaviour, although mathematically correct, is not beneficial for the GLAICP algorithm as even in such cases the pose configuration has to be adjusted. In fact, it is required that in a case of an unreachable target at least an approximate configuration that brings the key point as close as possible to its target is found. Moreover, only an approximate guidance pose is needed for the purpose of GLAICP as the final stages of the convergence process are mainly controlled by the local optimisation process. Therefore, a better suited IK approach is selected to be used in GLAICP. In particular, this is an approximate approach that is based on computation of the Jacobian of the forward kinematic chain as described below.

Jacobian based Inverse Kinematics

Considering that the position of a key point of the body can be given as a function of the state, i.e. the angles of the kinematic chain:

$$(x, y, z) = f(\theta)\ , \tag{5.10}$$

where x, y, z are the displacement in the coordinates of the point and θ is a vector made of the joint parameters, i.e. the angles of the kinematic chain, inverse kinematics (IK) is used to solve the problem of finding the θ from x, y, z :

$$\theta = f^{-1}(x, y, z) \tag{5.11}$$

Because of the nonlinear nature of the above function, as discussed earlier, it is not feasible for the problem to be solved directly. Therefore, it is linearized as:

$$(\Delta x, \Delta y, \Delta z) = J_p \Delta\theta\ , \tag{5.12}$$

where J_p is the Jacobian matrix; $\Delta\theta$ are small changes to the joint angles and $(\Delta x, \Delta y, \Delta z)$ is the displacement of the point's position, i.e. the small changes of its position along x, y, z axes. From (5.11) it follows that:

$$\Delta\theta = J^{-1}(\Delta x, \Delta y, \Delta z), \tag{5.13}$$

where J^{-1} is the inverse Jacobian matrix. However, since the Jacobian matrix is non-square, its inverse cannot be defined. Instead, the pseudoinverse is used. The pseudoinverse Jacobian matrix is computed as:

$$J^{+} = J^{T}(J_{p}J^{T})^{-1}, \tag{5.14}$$

where, J^{+} is the pseudoinverse Jacobian matrix and J^{T} is the transposed Jacobian matrix.

A typical human body kinematic chain, i.e. a human arm, is given in Figure 5-4 below.

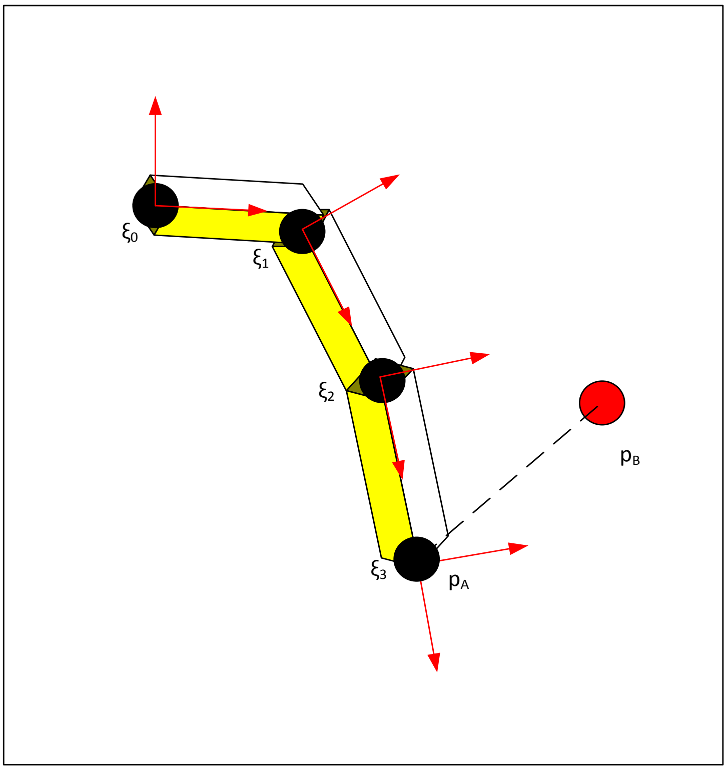


Figure 5-4: IK based on Jacobian of the Forward Kinematic Chain

Using the Forward Kinematic based on the twist transformations $(\boldsymbol{\xi_0}, \boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \boldsymbol{\xi_3})$ for the joints in the kinematic chain from the figure, the Inverse kinematic equations are derived as explained below.

Constraints

There are a few constraints that have to be taken into consideration when solving the minimisation problem, presented by (5.9). The constrains, imposed by the fixed lengths of the rigid linear elements in the skeleton, are embedded into the twist transformations for the joints as follows:

$$\left|\begin{array}{l}\boldsymbol{\xi_0} = (\omega_0, T_0) \\ \boldsymbol{\xi_1} = (\omega_1, (l_1, 0{,}0)) \\ \boldsymbol{\xi_2} = (\omega_2, (l_2, 0{,}0)) \\ \boldsymbol{\xi_3} = (\omega_3, (l_3, 0{,}0)) \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{array}\right. , \tag{5.15}$$

where l_1, l_2, l_3 are the lengths of the fixed elements and T_0 is the global transform for the root element of the kinematic chain. In addition to the constraints resulting from the fixed lengths l_1, l_2, l_3 and T_0, there are restrictions imposed by the joint angles that are unreachable or uncomfortable for a human, e.g. a knee bent backwards. However, these joint restrictions are imposed through the model but at a later stage of the computation process when the pose configurations are ranked based on the proposed Mechanism for Selection of the Most Realistic Pose Configuration.

Number of solutions

The above IK problem is under-constrained and equivalent to a non-linear square optimisation task. Depending on the position of the target point p_B, there are three possible cases: a) the point is inside the sphere of reachable space, b) it is outside the reachable space sphere or c) it is lying just on the outer edge surface of the reachable space sphere. The scenarios are analysed below:

CASE A: The target point p_B is within the reachable space. Then, the chances are that, if no joint restriction exist, multiple pose configurations exist that satisfy (5.9) and the joint angle restrictions. In reality, this case corresponds to the situation when the target point can be reached by the human arm in several different ways. In fact, the solution space in such this case could reach an infinite number of solutions, i.e. pose configurations, without any of them to be mathematically inferior when compared with the rest. However, from a human point of view, some of these pose configurations are less realistic as they require more strain and effort to achieve. Also, when people are in motion they try to minimise the amount of energy required. The above observation, applied as an optimisation to the kinematic chain, limits the number of likely pose configurations that can be used from an energy minimisation point of view to reach a certain point in space.

In similarity to the inverse kinematic method based on minimising discomfort (Abdel-Malek et al. 2004), the mechanism for selection of the most realistic pose configuration, proposed in 5.5.3, takes into account the level of discomfort experienced by the subject and the energy minimisation considerations to limit the number of considered poses. In particular, these factors are combined into a cost function, described in 5.5.3, which is then used to rank the possible poses and return the most realistic guidance pose configuration.

Case B: In the case when the target is unreachable, minimisation of (5.9) will never be able to bring the key point into its destination point. However, as it can be observed, (5.9) is formulated as minimisation that aims to bring p_A as close as possible to p_B. The result is a single solution, representing the pose configuration that positions the key point, p_A as close as possible to its target, p_B. In this case, the above approximate pose configuration is used as the guidance pose for the local optimisation part of GLAICP.

CASE C: There are number of points in space that are positioned between the reachable and unreachable space for the kinematic chain. Provided that there are no joint restrictions, the collection of all these points is a sphere with a centre positioned at the root of the kinematic chain and radius equal to the sum of the length of its constituent linear elements. When the target point p_B lies on the sphere it can be reached be a single pose configuration, i.e. a straight line of the linear elements of the kinematic chain.

The Pose Jacobian

As seen from (5.13) and (5.14), the Jacobian of the forward kinematic chain is needed in the Cases A above for the computation of the change in the joint angles from the required displacement of the key point. From the forward kinematic chain equation in (5.7) and the definition of the Jacobian, the following equation is given as:

$$J_p = [\frac{\delta p}{\delta\theta_0^x}\ \frac{\delta p}{\delta\theta_0^y}\ \frac{\delta p}{\delta\theta_0^z}\ \frac{\delta p}{\delta\theta_1^x}\ \frac{\delta p}{\delta\theta_1^y}\ \frac{\delta p}{\delta\theta_1^z}\ \frac{\delta p}{\delta\theta_2^x}\ \frac{\delta p}{\delta\theta_2^y}\ \frac{\delta p}{\delta\theta_2^z}] \tag{5.16}$$

where $\frac{\delta p}{\delta\theta_i^x}, \frac{\delta p}{\delta\theta_i^y}, \frac{\delta p}{\delta\theta_i^z}$ are the partial derivatives of the function p_s over the $\theta^x, \theta^y, \theta^z$ angles of i^{th} joint. The derivation of J_p is presented in Appendix A.

Finally, the pseudoinverse J^+ of the Jacobian is computed by applying (5.14).

After the adjustment of the joint angles is computed, a new position of the key point is generated using forward kinematics. The error, i.e. the distance between points A and B is re-calculated and compared against a pre-defined threshold. The process is repeated until the error becomes sufficiently small as presented below:

IK Part of the GLAICP

```
Input: Detected key-point positions: $x, y, z$
Output: New corrected joint angles: $\theta_t$
until $e(x, y, z) > \Sigma$ do
   calculate $J$
   $J^+ = J^T(JJ^T)^{-1}$
     subdivide $\Delta x, \Delta y, \Delta z$
     $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + J^+(\Delta x, \Delta y, \Delta z)$
     re-calculate $e(x, y, z)$
end
```

Algorithm 5.1: Overall algorithm for the IK part of GLAICP

5.5.3 Selection of the Most Realistic Pose Configuration

The proposed below Mechanism for Selection of the Most Realistic Pose Configuration (MSMRPC) is used to select the most appropriate pose configuration when more than one configuration are returned by solving the IK problem, i.e. CASE A described in 5.5.2. The key principles for evaluation of the returned poses are that it has to match the point cloud data and to be the most likely pose configuration that a person is likely to be. This is presented as an optimisation problem using a pose ranking function as explained below.

Furthermore, as in some cases it is likely that the number of solutions to the IK problem is infinite, the MSMRPC deploys a random sampling to reduce the number of pose configurations down to a manageable number. The above sampling mechanism selects only a few of all possible pose configurations for evaluation, which results in obtaining only an approximation for the optimal pose. However, the above approximation is sufficient for the needs of the local optimisation part of the GLAICP algorithm as the local pose optimisation can start from an approximate pose and converge to the optimal pose configuration. In Case A, when multiple pose configurations are returned by the IK, the proposed ranking criteria, modelling human motion behaviour, is used as explained below.

Pose Ranking Function, $\boldsymbol{F_{PR}}$

If more than one pose configurations minimising the alignment error are returned by the IK, a pose ranking function, F_{PR}, is used to rank them according to the probability of a human using each particular configuration to reach to the target position. The configuration probability is determined by using estimation of the measures of human comfort and effort required. Then, the pose with the highest ranking is selected as the guiding pose to be used by GLAICP for guidance of the local optimisation part. If, after a number of iterations, satisfactory alignment cannot be achieved using the selected guidance pose, then the next pose configuration in the ranking list is selected as the guidance pose. This process is repeated until sufficient alignment is achieved between key-point and its target or a time-out event occurs.

In particular, the pose ranking function is directly proportional to the extent of match between the appearance model, given the particular pose configuration, and the point cloud data, as measured by the point cloud matching function, $\mathrm{F_{PC}}$. Also, $\mathrm{F_{PR}}$ is inversely proportional to the difficulty for the person to reach the particular pose, as measured by the discomfort function, $\mathrm{F_D}$ as given bellow:

$$F_{PR} = \mu_{PC} \frac{F_{PC}}{F_D} \ , \tag{5.17}$$

where F_{PC} is the point cloud data match function; F_D is the discomfort function; μ_{PC} is a coefficient.

Discomfort Function, F_D

Inspired by the principle of least effort and the dynamic principles of human gait (Kuo et al., 2010), the proposed discomfort function F_D is designed to model the amount of effort required for people to change their pose. The function F_D is dependent on two factors, the effort, f_{eff}, and change in the potential energy f_{energy} of the human body:

$$F_D = \frac{\mu_{eff} f_{eff} + \mu_{energy} f_{energy}}{\mu_{eff} + \mu_{energy}} , \tag{5.18}$$

where μ_{eff} and μ_{energy} are weight coefficients.

The human effort is considered to be linked to the amount of motion, e.g. expansion or contraction, of the human muscles to reach the new pose. Therefore, the effort factor, f_{eff}, is modelled to be proportional to the angular distance, which the joints accumulatively have to travel in order for the skeleton to reach the target pose configuration from the initial pose configuration. The effort factor is given as:

$$f_{eff} = \sum_{i=1}^{n} |\vartheta_i - \vartheta_i'| , \tag{5.19}$$

where, i is the joint index, ϑ_i' is the initial angle of the i^{th} joint, ϑ_i is the joint angle for the evaluated pose configuration.

The energy factor, $\mathrm{f}_{\mathrm{energy}}$, measures the change in the potential energy between two different poses. The change of the potential energy is proportional to the change in altitude of the centres of mass for each of the limbs in the kinematic chain. Overall, the change in the potential energy level for the whole skeleton, i.e. its energy factor for change of pose, is given as:

$$f_{energy} = \sum_{i=1}^{k} (P_i' - P_i) , \tag{5.20}$$

where k is the number of the considered linear elements, P_i' and P_i are the potential energies of the new and the initial pose configurations of the i^{th} linear element respectively.

A negative $\mathrm{f}_{\mathrm{energy}}$ value indicates that the new pose configuration has lower potential energy in comparison with the initial pose. The pose configuration with the lowest energy, with all other factors equal, would be the preferred option for the pose selection algorithm as it requires the least energy to reach by a human.

In (5.20), the potential energy of a linear element from the kinematic chain can be represented by the mass of the element, the gravity and the altitude, i.e. $P = mgh$:

$$f_{energy} = \sum_{i=1}^{k} m_i g(h_i - h_i') , \tag{5.21}$$

where m_i is the mass of the linear element, g is the gravitational field strength constant (9.8 N/Kg), h_i and h_i' are the heights of the original and the evaluated pose configurations. The average mass of the separate linear elements, used in the GLAICP, is taken from the physical ergonomics literature (Chaffin et al. 2006). If required, they can also be adjusted in the initialisation step of the method.

As the centres of mass of the elements are points that are at fixed position in relation to the respective elements of the skeleton, their heights, h_i and h_i', are represented as functions of the joint angles of the skeleton, i.e. by using the forward kinematic map, given in (5.7). This allows the change in the potential energy f_{energy} to be represented as a function of the joint angles, θ, and calculated directly for each pose configuration returned by the IK.

Point Cloud Data Match Function, $\mathbf{F_{PC}}$

The point cloud data match function, $\mathrm{F_{PC}}$, corresponds to the likelihood that the local point cloud data is generated by the evaluated pose. It serves as a check that the considered pose configuration corresponds to the measurement data. In particular, the function $\mathrm{F_{PC}}$ measures the extent of fit between the appearance model of the evaluated pose configuration and the data from the point cloud. The function is computed as the ratio between the number of the point from the cloud that can be associated with points from the appearance model, given the maximum matching distance $\mathrm{R_{PC}}$, and the total number of the sampled points from the appearance model:

$$\boldsymbol{F_{PC}} = \frac{N_{matched}}{N_{total}} , \tag{5.22}$$

where N_{total} is the total number of sampled points from the appearance model, $N_{matching}$ is the number of the points associated with at least one point from the point cloud within the specified distance, R_{PC}.

5.5.4 Guidance to the Local Pose Optimisation

The output from the global part of GLAICP is a set of ranked pose configurations. One of them is selected as described above to be used as a guidance pose in the local optimisation part of the algorithm. The guidance pose is represented by a set of joint angles, $\Delta\theta = (\Delta\vartheta_1{}^{k}, \dots, \Delta\vartheta_n{}^{k})$, $k = 1..3$, where n is the number of joints.

The local optimisation part of GLAICP computes the angle changes $\Delta\Lambda = (\Delta\lambda_1{}^{k}, \dots, \Delta\lambda_n{}^{k})$, $k = 1..3$, needed to minimise locally the accumulative distance between the segment from the appearance model and the associated points from the point cloud dataset. Due to a number of external factors, e.g. discretisation error, outliers, wrong data associations and IK approximations, it is unlikely that the angles in $\Delta\theta$ and $\Delta\Lambda$ match completely. Therefore, a pose configuratoin guidance mechanism, using information contained in both the local and the global pose optimisation is proposed below. The mechanism is used to combine adaptively both local and global optimisation correction angles into a single correction angle, which is then used to update the position of the linear segment. In particular, the correction angle for the segment is computed as a weighted mean of both local and global correction angles. The weights of each component depend on the certainty in computation of each of them. The formula used for computation of the resulting angle for joint correction of ith joint is given as:

$$\Delta\varphi_i{}^{k} = \frac{w_\vartheta \Delta\vartheta_i{}^{k} + w_{\lambda_i}\Delta\lambda_i{}^{k}}{w_\vartheta + w_{\lambda_i}}, \tag{5.23}$$

where, i is the index of the joint, $k = 1..3$ is the index of the angle within the joint, $\Delta\vartheta_i{}^{k}$ and $\Delta\lambda_i{}^{k}$ are respectively the global and the local correction angle components associated with the i^{th} joint and the k angle, w_ϑ is the adaptive weight for the IK correction, which is the same for all elements in the kinematic chain and w_{λ_i} is the adaptive weight for the ICP correction elements $\Delta\lambda_i{}^{k}$, $k = 1..3$, for the ith joint.

In the next step, the combined joint adjustment angles, $\Delta\varphi_i{}^{k}$, are used instead of the angles $\Delta\boldsymbol{\lambda}_i{}^{k}$ in the standard articulated ICP for computing the iterative correction of the i^{th} joint. The weights w_ϑ and w_{λ_i} depend on the certainty in identification of the targets of the key points for the kinematic chain and of the certainty of the local point-to-point associations as explained below.

Adaptive global optimisation weight

In principle, if the confidence in target position identification is high, i.e. there is a good match between the stored signature pattern and the signature computed from the point cloud, then the weight for the global part, w_ϑ, is increased and the IK component takes a more significant role in the overall correction angle, computed in (5.23). The weight function of the global optimisation part of GLAICP, w_ϑ, is given as:

$$w_\vartheta(C_\vartheta, x_s) = \frac{C_\vartheta}{(1 + C_\vartheta + {x_s}^2)}, \tag{5.24}$$

where C_ϑ is a parameter of the GLAICP algorithm, $0 \leqslant C_\vartheta \leqslant 1$, which determines the dependence of the certainty on the weight, w_ϑ, change, and x_s is the measurement of the confidence in the match between the point cloud with the stored reference signature. In particular, the variable x_s, $0 \leqslant x_s \leqslant 1$, depends on the ratio of the number of matching bins between the stored reference signature and the point cloud data, as defined later in 5.5.6. For example, when there is a full match between both signatures then $\mathrm{x_s} = 1$. On the opposite, when the signatures are completely different and there is no match between them, then, $x_s = 0$, resulting in a minimal weight of the global optimisation part.

By modifying the parameter C_ϑ it is possible to adjust the weight of the IK element in the GLAICP algorithm. For example, a dependence of the global correction on the matching certainty for a number of typical values of C_ϑ is given in Figure 5-5 below.

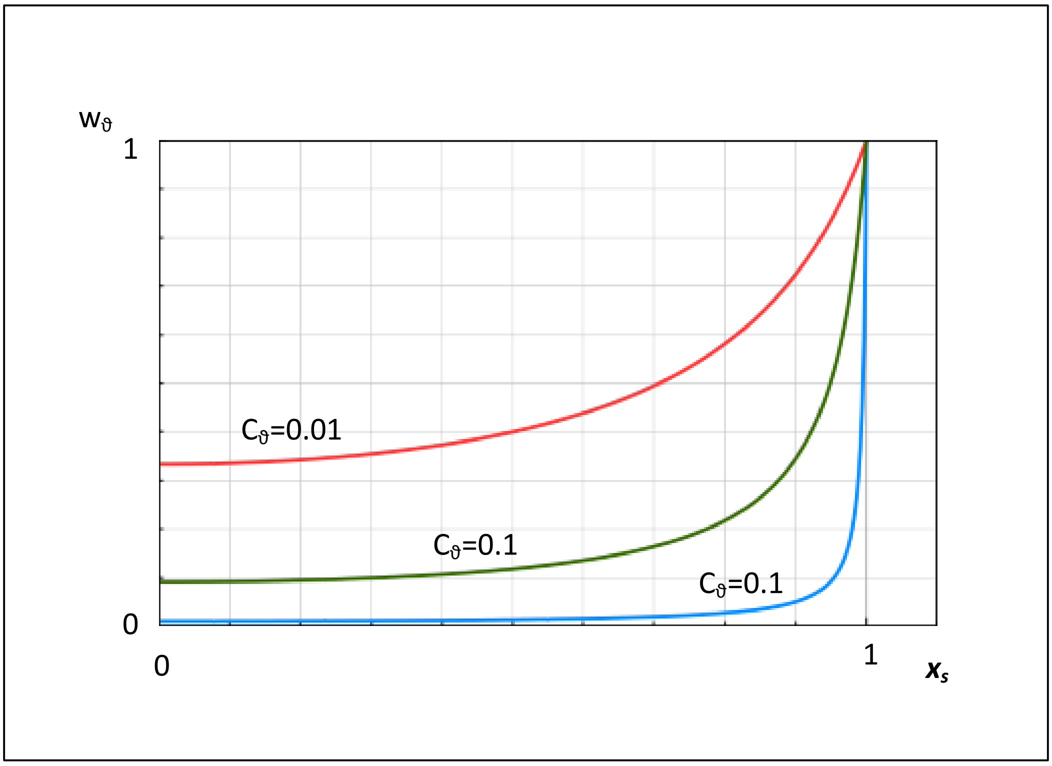


Figure 5-5: The weight function of the global optimisation part.

As seen from the graph in the figure, the maximum value of the global weight function, w_ϑ, is the value of 1, which represents the weight when there is a full match in the signatures. The minimum value for w_ϑ depends on the C_ϑ parameter. Also, it should be noted that the weight function $w_{\vartheta}(C_{\vartheta}, x_s)$ is a non-linear function which drops significantly at lower confidences x_s to reduce the impact that partially matched signatures have on the convergence process.

Adaptive local optimisation weight

The joint adjustment from the local optimisation part of GLAICP is dependent on the establishing of point-to-point associations between points on the surface of the appearance model and the points from the point cloud. The certainty in the computation of the correction angle increases with the increase of the number of successful associations. For example, when all the points from the segment of the appearance model are close to points from the point cloud, then the certainty in the computed adjustment angle is high and the algorithm increases the weight of the local optimisation part. In particular, the importance weight of the local part is calculated based on the certainty in the point-to-point associations. The dependence is given by the following weight function $w_{\lambda}(C_{\lambda}, x_i)$ for the local optimisation adjustment:

$$w_{\lambda}(C_{\lambda}, x_I) = \frac{C_{\lambda}}{(1 + C_{\lambda} + {x_I}^2)} , \tag{5.25}$$

where C_{λ} , $0 < C_{\lambda} < 1$, is a parameter provided to the GLAICP algorithm, which determines how the weight changes with the number of successful point-to-point associations, and x_i, $0 \leqslant x_i \leqslant 1$, is the measurement of the certainty of the local optimisation part. The measurement x_i is computed as the ratio of number of established point-to-point associations and the total number of points available for such associations:

$$x_i = \frac{N_i}{M_i} , \tag{5.26}$$

where M_i is the number of points sampled uniformly from the appearance model of the i^{th} linear element, and N_i is the number of point-to-point associations established between points from the appearance model and points from the point cloud. Due to a limit in the maximum distance, R_{max}, in which the closest neighbour search is operates, it is normal that there are not enough suitable points that can be associated with points from the appearance model and therefore $N_i \leqslant M_i$. With the increase of distance of the closest neighbour search, R_{max}, the certainty will increase potentially due to the increased number of point-to-point associations.

The importance of the local optimisation part is controlled by modifying the parameter C_{λ}. It should be noted that, similarly to C_{ϑ} in Figure 5-5, the function of local optimisation weight $w_{\lambda}(C_{\lambda}, x_I)$ is a non-linear function that diminishes rapidly with the increase of the uncertainty in the local optimisation part. This reduces the undesired effect that a small number random local associations can have on the overall operation of GLAICP.

5.5.5 Identification of Regions of Interest in the Point Cloud

Due to the relatively high computational overhead in the comparison of feature signatures and the large number of points in the point cloud, the real-time computation of the HISP local shape descriptor for every point from point cloud is a challenging task that requires substantial parrallel computational resources. Therefore, in GLAICP, the CHISP descriptors are only computed for points belonging to relatively small regions of the point cloud, i.e. regions of interest (ROI). This optimisation allows significant reduction in the computational overhead of the algorithm. ROI are identified by applying a method for computation of the discrete conformal factor (Ben-Chen et al. 2008) and selecting the subset from the point cloud that has values of the conformal factor bigger than the pre-defined threshold value, Φ_{min}.

The conformal factor, Φ, is a pose-invariant measure, based on the conformal geometry. In principle, it represents the amount of local work required to transform the mesh into a sphere. The conformal factor is computed using the following equation (Ben-chen & Gotsman 2008):

$$L\Phi = K^{T} - K^{orig} \ , \tag{5.27}$$

where L is the discrete Laplace-Beltrani operator with cotangent weights (Hildebrandt et al. 2007), K^{T} is a vector containing the target Gaussian curvature and K^{orig} is a vector containing the Gaussian curvature of the mesh.

The discreet Gaussian curvature, k_v^{orig}, of a vertex v in the mesh, is defined by Meyer et al. (2002) as:

$$k_v^{orig} = \begin{cases} 2\pi - \sum_{t \in T_v} \theta_t & , v \notin B \\ \pi - \sum_{t \in T_v} \theta_t & , v \in B \end{cases} \ , \tag{5.28}$$

where θ_t is the angle near the vertex v in triangle t, T_v is the set of triangles connected to v, and B is the set of all vertices on the mesh boundary.

As the target is a uniform Gaussian curvature, each vertex can be assigned a portion of the total curvature. The target curvature k_v^t at a vertex v is then given as:

$$k_v^t = \left(\sum_{i \in V} k_i^{orig}\right) \frac{\sum_{t \in T_v} Area(t)/3}{\sum_{t \in T} Area(t)} \ , \tag{5.29}$$

where V is the set of all vertices and T is the set of all triangles in the mesh. As seen from (5.29), the portion of the total curvature assigned to the vertex depends from the "influence area" of the vertex, i.e. a third of the area of the faces near the vertex, divided by the total surface area of the mesh.

As proposed by Ben-chen & Gotsman (2008) (5.27) can be solved efficiently through Cholesky factorisation. It can be observed from the results that the conformal factor gradually increases along the length of the mesh extrusions. For a mesh representing a human pose, the highest values of Φ are at the extremities of the limbs, i.e. the tip of the fingers. The above observed property of the conformal factor is a very interesting feature that is directly applicable to GLAICP. If the key points used for the signature match are positioned where the conformal factor reaches its maximum value then the search for matching signatures can be accelerated considerably by reducing the number of points for which signatures are computed. Moreover, it is possible to attach a global position information to a signature making it unique within the human body. This largely eliminates the risk of mismatch between points from different parts of the human body that have similar signatures.

In conclusion, the combination of a local descriptor and global body position location information, i.e. as given by the conformal factor, augments significantly the local descriptor by adding an extra layer of global body position information. The combination, referred further as a hybrid descriptor, reduces significantly the false positive rate in comparison of descriptors. The additional computational overhead reduction, caused by reduction in the number of computed signatures, addresses one of the main challenges in tracking of human pose through local features identification, i.e. the computation overhead. The proposed below hybrid descriptor, named **Conformal Human Interpretable Signature of Points (CHISP)** is considered as one of the contributions of this work.

The conformal factor information is used in GLAICP to identify key ROI from the human body. Subsequently, the local feature signatures are computed only for those regions of interest. The optimisation, based on separation of the computation of the global and local parts of the CHISP descriptor, contributes to reducing the required computation resources for the overall human tracking algorithm by avoiding computation of the local feature signatures for the majority of points.

In practice, the CHISP descriptor is computed by concatenation of [8x1] vector representing the conformal factor, $(h^c[1], \dots, h^c[8])$ and the vector $h^H[9], \dots, h^H[n]$ representing the standard HISP descriptor as shown in Figure 5-6. The resulting CHISP descriptor is given as:

$$H_{CHISP} = (h^c[1], \dots, h^c[8], h^H[9], \dots, h^H[n]) \ , \tag{5.30}$$

where n is length of the HISP descriptor, which depends on the input parameters determining the space division of the local feature, as described in 4.5.2.2.

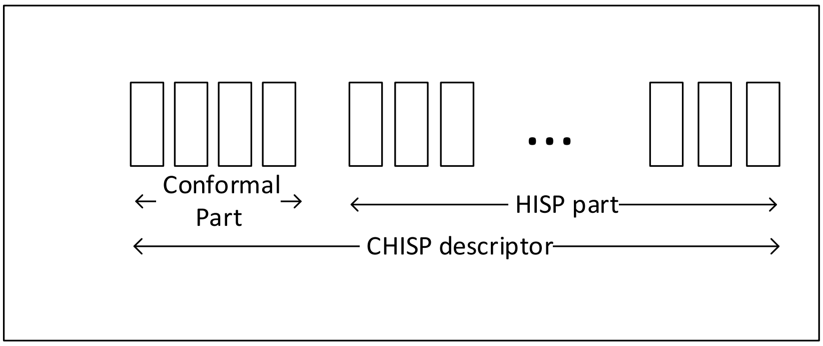


Figure 5-6: CHISP descriptor

Subsequently, the CHISP descriptor is computed for each point within ROI, as described in 4.5.2. The resulting signature is compared against previously stored descriptors of key points from the human limbs, described in 5.5.6. If a positive match is identified, the key point-target pair is established and subsequently used in the inverse kinematics part of GLAICP, as described in 5.5.2.

5.5.6 CHISP descriptor matching

The inverse kinematics part of GLAICP is based on identification of key point-target pairs that are used for the calculation of the guidance pose configuration. The following process steps are required for finding the key point-target pairs.

Initially, the presence and location of people is established using the method proposed in Chapter 4. Then, the point cloud is segmented, using human location information, to leave only the parts of the point cloud that contain shapes of human bodies. Next, a mesh is constructed from each of the point cloud segments to identify the regions of interest, as described in 5.5.5. Then, the CHISP signatures are computed for the points belonging to identified ROI and compared with pre-recorded reference CHISP signatures. The results from the comparison represent a collection of known key points from the human body, which are later used as the targets of the IK part of GLAICP as described below.

Due to the noise in the measurement data, it is likely that some points migrate between neighbouring bins of the signature. If a precise matching between signatures is targeted, the point migration would result in the inability to achieve a full match between signatures originating from the same position in the human body. This signature fluctuation will lead to an increased rate of false negatives. Therefore, to mitigate the above issue, an approximate matching method is proposed for comparing CHISP signatures. The method deals separately with the conformal part and the signature (HISP) part of the CHISP descriptor as described below:

Global comparison (conformal part):

Initially, the conformal part of the current signature is compared against the conformal parts of pre-stored reference signatures, representing key points of the human body. In particular, the comparison consists of subtraction of the numbers, h_1^c and h_2^c representing the conformal parts of the two signatures:

$$D_{Global}\,(h_1^c, h_2^c) \; = \; |h_1^c - h_2^c| \;\; , \tag{5.31}$$

Then, the difference, $D_{Global}\,(h_1^c, h_2^c)$, is compared with a threshold value, given as an input parameter, and a decision is made for continuation of the comparison of the rest of the signatures, as given by the rule below:

$$if\ D_{Global}\,(h_1^c, h_2^c) \leqslant\ Tr_{HISP} \begin{cases} true, & Continue \\ false, & Abort\ comparison \end{cases} \tag{5.32}$$

If the conformal parts differ significantly, suggesting that the signatures belong to different parts of the human body, the result of the comparison in (5.32) is negative and the comparison terminated. Otherwise, the local part of the signature is computed from the point cloud and compared with the stored reference signature as described below.

Local comparison (signature part): If the conformal parts of the CHISP signatures are found to be within a certain distance measure of similarity, defined by the given threshold, then the second phase of the comparison is executed. The second phase of the comparison is based on a particular property of the HISP descriptor. This property is related to the fact that HISP disregards the position of the slices within a cone, as described in 4.5.2. When computing the signature, HISP sorts the slices according to the number of points and presents them in increasing order. In the graphical representation of the descriptor, the above feature presents in a number of distinctive shapes, illustrated in the top diagram in Figure 5-8. These trapezium shapes represent the underlying surface topology of the point cloud and are used for fast comparison of signatures.

In particular, the idea behind the accelerated comparison method is based on fitting a right trapezium template, illustrated in Figure 5-7, over the separate segments in the HISP signature, as shown in Figure 5-8. The width of the distinctive placeholders within the HISP signature is constant as it is linked to the fixed number of slices within the spheres. Therefore, only two parameters are sufficient to encode the position of each fitted template, i.e. the length of the smaller side of the triangle and the angle between the leg and the small side, α. For computation time optimisation purpose, the angle α is presented by its tangent, i.e. the ratio of its opposite and adjacent sides of the ABC triangle, i.e. $\tan\alpha = \frac{BC}{AB}$.

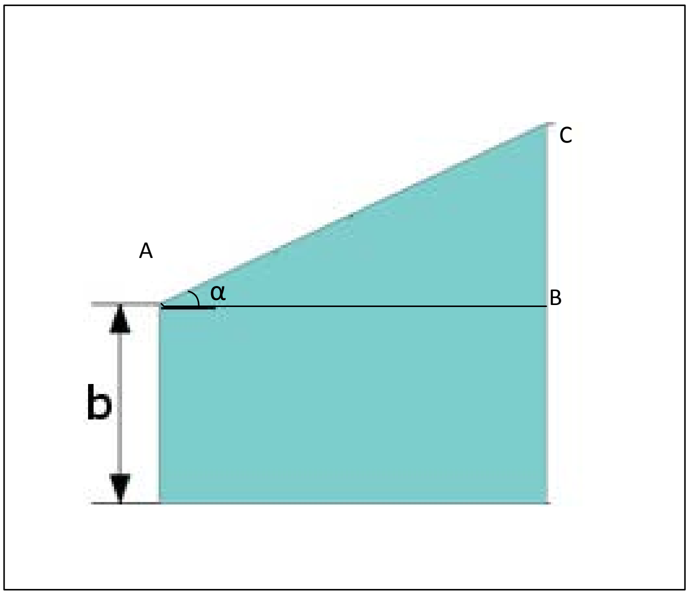


Figure 5-7: Right trapezium template used to fit segments in the HISP signature

The procedure of conversion of HISP signature into its simplified form, i.e. Simplified SHISP descriptor (SHISP), is illustrated in Figure 5-8. The result is the SHISP signature given by a [2xn] matrix:

$$M_{SHISP} = \begin{bmatrix} \tan\alpha_1 & \dots & \tan\alpha_n \\ b_1 & & b_n \end{bmatrix}, \tag{5.33}$$

where α_i and b_i, $i = 1 \dots n$, are the parameters of i^{th} trapezium shape in the signature, and n is the total number of cones in the signature, which is provided as an input parameter to the algorithm.

For example, the HISP signature part, given by following vector:

⟨0.0,0.3,0.6,0.8,0.2,0.3,0.4,0.6,0.3,0.4,0.5,0.6,0.0,0.0,0.0,0.0,0.5,0.5,0.5,0.5,0.2,0.4,0.6,0.8,0.1,0.2,0.4,0.5⟩

can be simplified, using the above process, illustrated in Figure 5-8, into the [2x7] matrix M_{SHISP_1}:

$$M_{SHISP\_1} = \begin{bmatrix} 1.0\ 0.7\ 0.6\ 0.0\ 0.0\ 0.9\ 0.6 \\ 0.0\ 0.1\ 0.2\ 0.0\ 0.5\ 0.1\ 0.0 \end{bmatrix} \tag{5.34}$$

The serialised form of the matrix M_{HISP_1} is a vector H_s, representing the Simplified SHISP descriptor (SHISP):

$$H_{SHISP} = (h^{\alpha}[1], \dots, h^{\alpha}[n], h^{b}[n+1], \dots, h^{b}[2n]) \ , \tag{5.35}$$

Using SHISP allows a light-weight key point-target comparison, as it is considerably shorter in length and also allows a fast approximate comparison of two signatures.

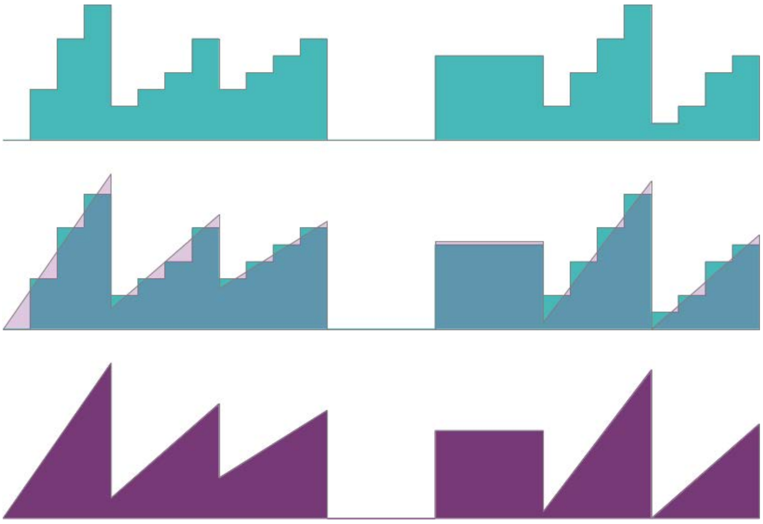


Figure 5-8: Representation of the HISP descriptor with simplified geometric shapes

In particular, the comparison between two SHISP vectors, H_{SHISP_1} and H_{SHISP_2}, is done through computation of the Bhattacharyya coefficient (Bhattacharyya 1943) on the normalised α,and b parts of the H_{SHISP_1} and H_{SHISP_2} vectors. Let $\widetilde{H}^{\alpha}_{SHISP}$ be the normalised vector of α part of a H_{SHISP} vector:

$$\widetilde{H}^{\alpha}_{SHISP}\ [k] = \frac{H^{\alpha}_{SHISP}\ [k]}{\sum_{k=1}^{n} H^{\alpha}_{SHISP}\ [k]} \tag{5.36}$$

,

Then, the Bhattacharyya coefficient of $H^{\alpha}_{SHISP\_1}$ and $H^{\alpha}_{SHISP\_2}$ is computed as:

$$S_{\alpha}\left(H^{\alpha}_{SHISP_1}, H^{\alpha}_{SHISP_2}\right) = \sum_{k=1}^{n} \sqrt{\widetilde{H}^{\alpha}_{SHISP\_1}\ [k]\ \widetilde{H}^{\alpha}_{SHISP\_2}\ [k]} \tag{5.37}$$

The Bhattacharyya coefficient of $H^{b}_{SHISP\_1}$ and $H^{b}_{SHISP\_2}$ is computed in similar way to (5.37) for the b-parts of the H_{SHISP_1} and H_{SHISP_2} vectors:

$$S_{b}\left(H^{b}_{SHISP_1}, H^{b}_{SHISP_2}\right) = \sum_{k=1}^{n} \sqrt{\widetilde{H}^{b}_{SHISP\_1}\ [k]\ \widetilde{H}^{b}_{SHISP\_2}\ [k]} \tag{5.38}$$

Finally, taking into account (5.37) and (5.38), the combined similarity between two CHISP feature vectors is evaluated as a two-dimensional similarity vector:

$$S\left(H_{SHISP\_1}, H_{SHISP\_2}\right) = (S_{\alpha}, S_{b}) \tag{5.39}$$

After computation of the degree of similarity of the signatures of all points in the region of interest against the stored reference signatures is completed, the point with the highest degree of similarity, i.e. resulting in similarity vector with the higher norm $\|S\|$, is established. Then, if the elements S_α and S_b of this point are higher than the pre-defined thresholds, S_{α_min} and S_{b_min}, provided as input parameters, the point is selected as a target position of the respective key point and used in GLAICP for computation of the guidance pose.

5.5.7 GLAICP algorithm

A general overview of the GLAICP algorithm is provided below. Also the overview is depicted graphically by the block diagram shown in Figure 5-2.

In each timeframe, after acquisition of the latest point cloud, first the GLAICP algorithm segments it, using the Euclidean Cluster Extraction algorithm (Rusu 2009b), to break it down into its constituent point clusters. Then, the clusters are processed to identify the presence of any human body shapes in the clusters. In particular, the identification is based on the human presence and localisation information, provided by the algorithms from Chapter 4, by isolating the points in cylinders positioned at the centre of reported detections. Next, GLAICP generates a mesh from the edge points in the segmented dataset, identifies the regions of interest in the mesh, as explained in 5.5.5, searches the region of interest for point descriptors matching those of the pre-stored key-point signatures of the limbs using the three-stage procedure described in 5.5.6. Next, GLAICP aligns the trunk of the human body with the point cloud, as described in 5.5. Then, it computes one or more probable pose configurations for each of the kinematic chains, representing the human limbs, through the method described in 5.5.2. Next, the guidance pose is selected for each of the kinematics chains, based on the ranking of the poses, as described in 5.5.3. Subsequently, the selected guidance pose is used to guide the local convergence process, using the method described in 5.5.4. The overall result of the current cycle is evaluated using the alignment measure function, described in 5.5.1. If the alignment error is higher than the pre-set threshold, the cycle is repeated using the next pose configuration in the ranking. If none of the poses identified by IK results in an alignment measurement smaller than the threshold, then the configuration that produces the smallest error is selected and returned as the final output from the GLAICP for the current timeframe. Subsequently, the final output pose configuration is used as a starting pose configuration in the next time frame.

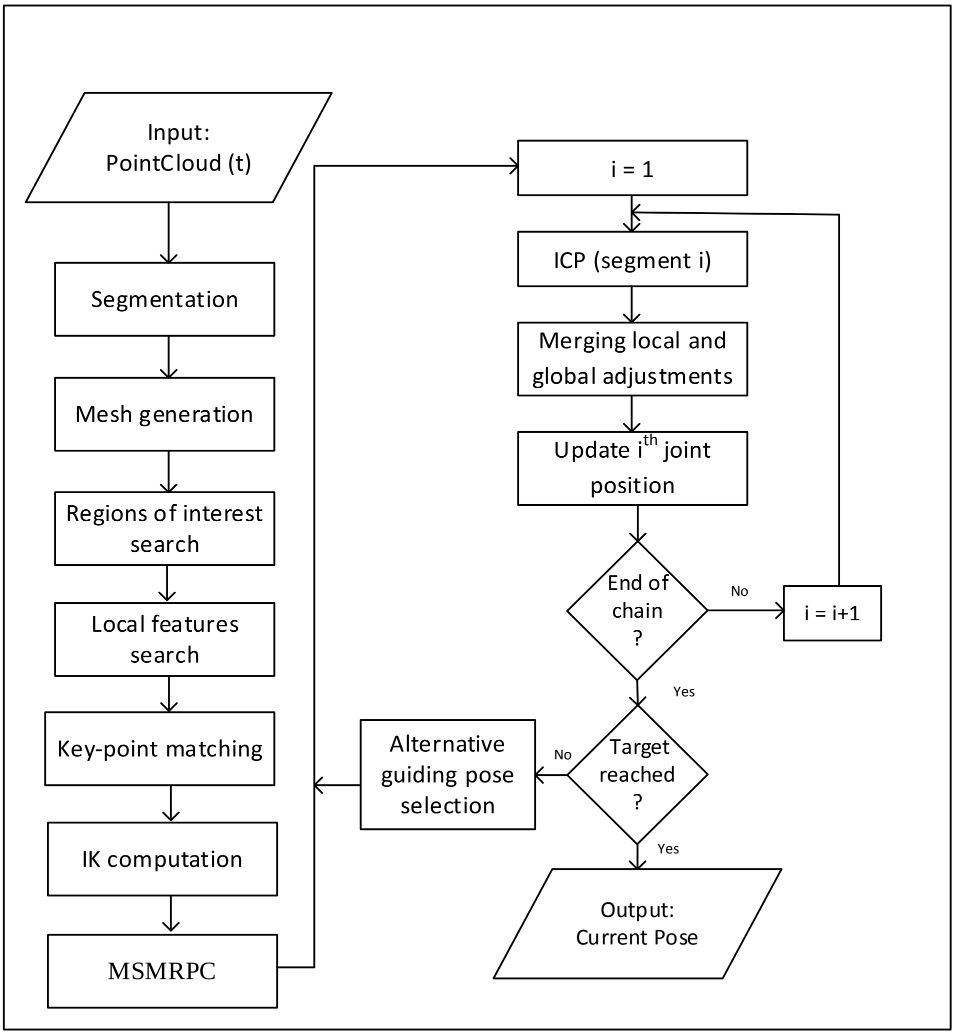


Figure 5-9: The GLAICP algorithm

A potential for optimisation of the above algorithm is possible by the introduction of parallel parts of the computation process. The most appropriate candidates for parallelisation are the four separate kinematic chains, the search for a key point in the regions of interest to establish origin-target pairs and the evaluation of the optimal convergence for different guidance pose configurations. The acceleration resulting from parallelisation will be explored as a part of future work.

5.6 Experimental results

Evaluation of the GLAICP algorithm was carried out by comparing its tracking accuracy against the accuracy of the original AICP algorithm. A dataset, consisting of recorded timestamp point cloud depth images, at resolution 640 x 480, was collected using an RGB-D sensor, i.e. Microsoft Kinect. Simultaneously, the 3D space positions of 12 markers, positioned at key human body joints, was acquired using a MOCAP system, i.e. PhaseSpace Impulse X2 (Phasespace 2014).

In the experiment, the MOCAP system was configured to use the 12 available cameras located around the recording area and record the position of the 12 active LED markers, as shown in Figure 5-10 below. The positions were recorded with average accuracy of less than 0.5mm within the measurement volume, as specified by the manufacturer. Also the recording was done at a rate of 420 frames/second and maximum latency of 10ms. The recorded time-stamped measurements of the MOCAP system were used as the ground truth for the evaluation of the tracking error of both, AICP and GLAICP algorithms. The goal of the experiment was to establish the relative improvement, if any, of GLAICP over the standard AICP.

In the experiment, a person carried out random motion sequences including torso, arm and leg movements. The sequences were recorded by a Kinect sensor and the MOCAP system running simultaneously. Subsequently, the registered point cloud datasets were processed both by the GLAICP and the standard ICP algorithms and the results were compared against the MOCAP data to calculate the absolute error in tracking. In particular, in each time frame, the resulting absolute human joint positions of the tracked skeleton were compared with the recorded positions of the markers, located at the monitored human body joints by the MOCAP system, i.e. the ground truth. The absolute distance error e_d, defined as the Euclidean distance between the position of the MOCAP marker and the positions reported by the proposed GLAICP and the standard AICP algorithms, were computed and the average values are shown in Figure 5-11 below .

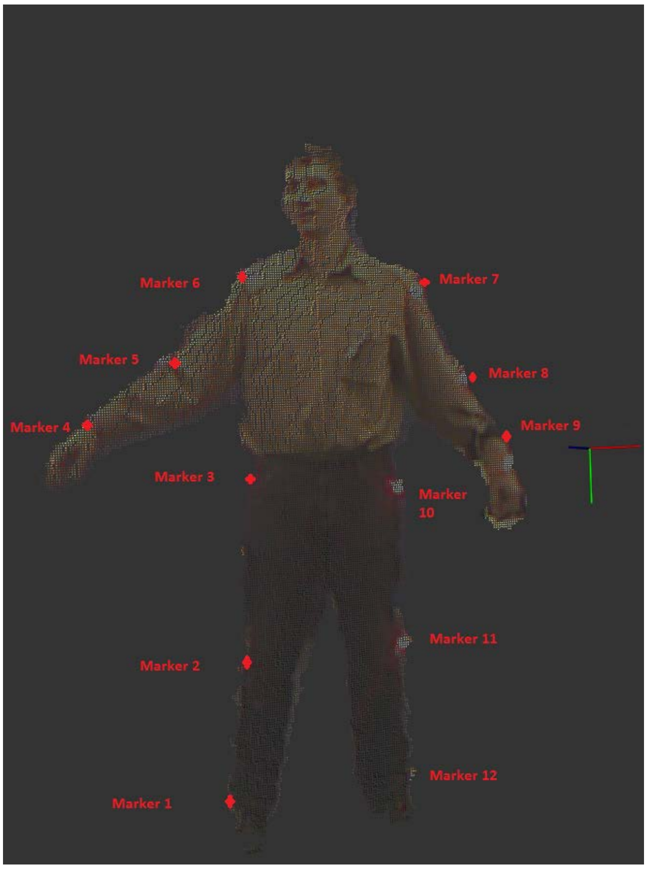


Figure 5-10: Position of the markers in the experimental setup (point cloud visualisation)

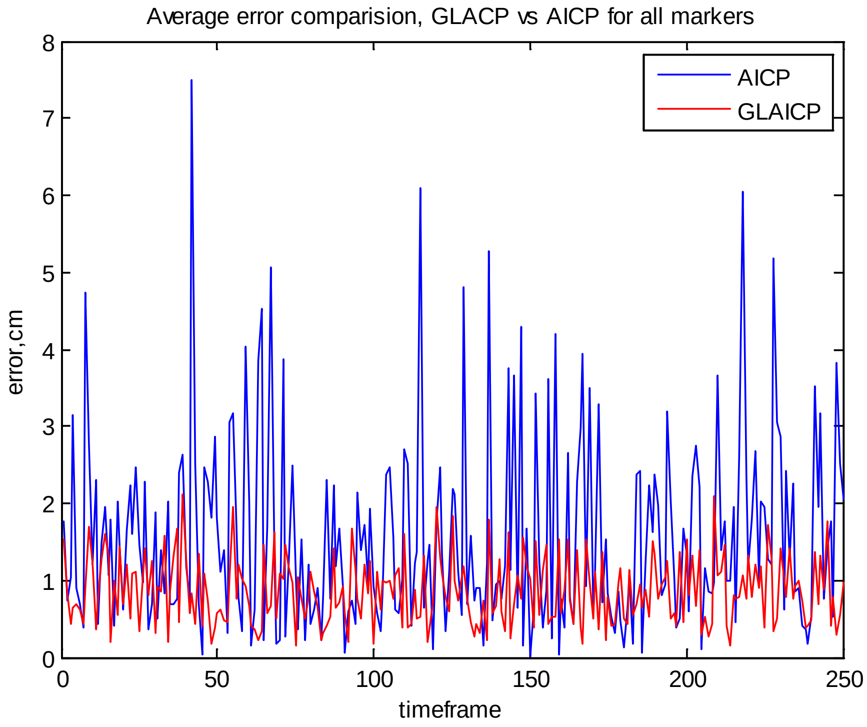


Figure 5-11: Distance average error comparison between GLAICP and AICP algorithms

As can be seen from the above figure, the GLAICP algorithm reduces both the average error and the tendency of the standard AICP to converge to the local minima, manifested with sharp peaks in the average error distance of the standard AICP. The above observation is also confirmed by the probability density function of the distance error, given by the histogram of the distance error averaged for all human body joints, shown in Figure 5-12.

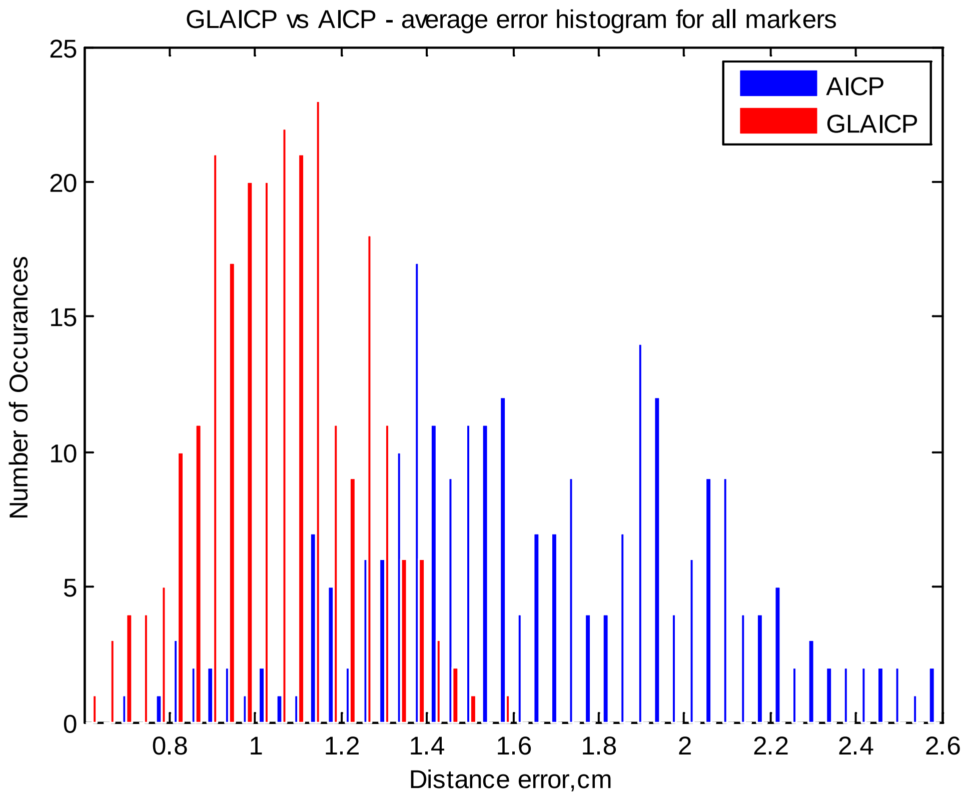


Figure 5-12: Distance error distributions of GLAICP and AICP algorithms

As seen from Figure 5-12, the error distribution of the proposed GLAICP algorithm is narrower in comparison with the distribution of the error of the standard AICP algorithm. This confirms the earlier observation that GLAICP improves the performance of the standard AICP.

An interesting observation is the non-zero offset of the mean value of the error for GLAICP. A possible explanation for this could be the fact that the MOCAP markers are not positioned at the geometric centre of the human body joint. Instead they have a small offset due to the way they are attached to the human body surface. The detailed errors for each joint are provided in Appendix G. The resulting mean error and standard deviations are given in Table 5-2 below.

	GLAICP		AICP	
	Mean error, cm	Std. dev. ,cm	Mean error, cm	Std. dev. ,cm
Marker 1	*1.0420*	*0.5309*	*1.5803*	*1.1159*
Marker 2	1.3719	0.7515	1.4570	1.2197
Marker 3	1.0374	0.5603	1.6637	1.5158
Marker 4	1.0356	0.5114	1.5289	1.5573
Marker 5	1.1303	0.6133	1.6758	1.1337
Marker 6	1.1845	0.5982	1.6689	1.6184
Marker 7	1.1216	0.8562	1.7336	1.0351
Marker 8	1.0436	0.6143	1.8459	1.7475
Marker 9	0.8745	0.7666	1.6206	1.0590
Marker 10	0.9780	0.6234	1.8984	0.9499
Marker 11	1.0542	0.5358	1.7762	1.2341
Marker 12	0.8587	0.4373	1.5654	1.2461
All markers	1.0610	0.1808	1.6679	0.3938

Table 5-2: Evaluation of the improvement of GLAICP over AICP

5.7 Evaluation of the results

As seen from the results from the above experiments, comparing the proposed GLAICP with the standard AICP algorithm, the GLAICP algorithm reduces the average error in human pose tracking with approximately 30%. Moreover, the probability of big tracking errors, often leading to a total tracking failure is also reduced, which leads to an improved robustness of tracking. In particular, the increased robustness of human pose tracking was demonstrated by an approximately 40% drop in the standard deviation of the tracking error. Due to the time saved at the beginning of the convergence process, more time is available in the final stages of the process, enabling a closer match between a model and sensor data.

In conclusion, the reduction in tracking error enables more accurate and robust tracking of the human pose configuration as demonstrated by the experiments. In the context of service robotics, knowing the human pose more precisely and with greater confidence enables closer physical interactions between a robot arm and the user which ultimately leads to enhanced HRI.

5.8 Summary

The proposed GLAICP algorithm offers the advantage of reducing the vulnerability of human pose tracking in its tendency to suffer a local minima problem, a typical issue for the popular class of local iterative human pose tracking algorithms. As demonstrated in the experiments, the human pose tracking by GLAICP is characterised by a reduced distance error in comparison with the standard AICP.

The GLAICP algorithm is based on a combination of two methods. The first method is the global pose optimisation, using local descriptor search in point cloud data. The second method is the standard AICP, which is based on local optimisation. It is claimed that the proposed combination of both improves the human pose tracking characteristics, i.e. tracking with smaller error and increased robustness. In particular, GLAICP inherits from the global optimisation part the robustness to local minima problems as well as the ability to jump directly to a state that is closer to the real pose configuration. Similarly, GLAICP inherits from the local optimisation approaches the accuracy of the correct pose configuration in the final stages of the convergence process.

The overall reduced uncertainty in the human pose configuration tracking is a major factor for enabling closer physical interaction patterns between a robot and a human. In addition to achieving sufficient reliability and accuracy in human pose tracking, other measures can be employed for increased safety of a person engaging in close physical interaction with a robot. These include, for example, using a lightweight robotic arm with force feedback, pressure sensitive robotic arm surface and capacitive touch sensors. It is considered that improved human pose tracking and the additional safety measures allow much closer physical interaction patterns between a robot and a human. These patterns resemble the natural human to human interactions and, as a result, are more likely to meet the user needs for more meaningful and lifelike interaction with a service robot.

In conclusion, by reducing the risk of local minima problem and a tracking failure the GLAICP guarantees higher availability of human pose information needed for the decision-making algorithms of the robot by reducing the probability of tracking failures. As a tracking failure recovery is a very time-consuming operation,i.e. re-initialisation, this improvement is considered to be an important factor for increasing the overall tracking information availability. The constant information availability about the human status is envisaged to play an important role in enabling close physical human-robot interaction. It is believed that in service robotics, applications where the robot can provide physical assistance to the user through close physical HRI have the potential to revolutionise the home care and support services for the elderly and in firm.

Contributions

Two contributions are made in this chapter as follows:

- GLAICP - a novel human pose tracking method that combines both global and local human pose optimisation to improve the robustness of the human pose tracking;
- CHISP - Conformal Human Interpretable Signature of Points.

6 Social Group Aware Navigation Planning through Human Detection Analysis

6.1 Background

Navigation path planning is another important aspect of any service robotics application. A good navigation path guarantees that the robot can reach the target while avoiding obstacles and achieving at the same time one or more optimisation goals, e.g. minimal energy or shortest time to destination. Since the navigation path is an integral part of the HRI patterns, it is important when planning a navigation path for a service robot to take into account the social norms of acceptable behaviour. It is also important to understand the dynamics and the context of the environment in order to adapt the navigation strategy accordingly. In crowded environments, one of the accepted norms of social behaviour is to avoid interruption to the ongoing human interactions, e.g. by avoiding the space where the interaction takes place. One of the key premises in this work is that that the robot can achieve increased user acceptance by complying with the social norms of behaviour. Therefore, a new optimisation strategy, aimed at minimisation of the interruption of interpersonal interactions within the social group during navigation to a target, is investigated in this chapter.

Computing an optimal navigation path to a person, especially in a crowded environment, is a challenging task that requires in-depth reasoning about the context of the environment. In this chapter, a novel approach that extracts context information using analysis of the human tracks is proposed. In particular, the method establishes the boundaries between the social groups and estimates the intensity of the interpersonal interaction between group members.

Initially, it was intended that human tracks were to be constructed directly from the human detections, made by the classifier fusion method, described in Chapter 4. However, such a strategy works well only when there are a limited number of people in the environment and when they are not close to each other. With the increase in the number of people, e.g. in a crowded environment, due to the increased ambiguity in data associations between tracks and detections, direct detection to tracks rapidly becomes a challenging task. Therefore, in the proposed method, track estimation is carried out by Bayesian inference of the most likely association between a track and detections and the position of the people. Subsequently, the newly estimated human tracks are analysed to establish the human group boundaries and the active interaction between the group members. Finally, a minimally intrusive navigation path is generated using the above information about the human interaction with the social groups. This thesis proposes that the generated navigation path minimises disruption to the human interactions by avoiding the areas where these interactions take place. A conceptual overview of the whole process is represented graphically in Figure 6-1 below.

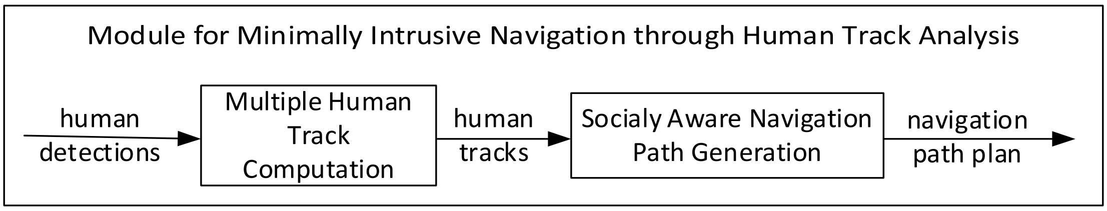

Module for Minimally Intrusive Navigation through Human Track Analysis
human detections
Multiple Human Track Computation
human tracks
Socialy Aware Navigation Path Generation
navigation path plan

Figure 6-1: Conceptual overview of the Minimally Intrusive Navigation Subsystem

Multi-object tracking, also known as multi-target tracking, plays a significant role in many important domains of life such as surveillance (Chai et al. 2011; Xie et al. 2004) , pedestrian detection (Dollar et al. 2012) and air traffic monitoring (Meng et al. 2002) . Typically, in such tracking scenarios some objects, with indistinguishable identity, move continuously within the observed space. In principle, from an external observer's point of view, new targets can appear at random in space and time, remain for a random period, and then disappear at random. Here, the sequence of positions that a target follows through its lifetime in the observed space is referred to as a human track. The position of the targets within the space is either measured by the sensors, either at random intervals or periodically in scans, or is the result of sensor/classifier fusion. The main challenges for the proposed method include: a) accounting for tracks that have newly emerged or ceased to exist for estimation of the number of tracks at each time frame, b) finding the most probable associations between tracks and measurements and c) estimating the position of the people in the Euclidian space from the available measurements.

In service robotics, multiple human tracking, i.e. establishing the number and the position of correct human tracks from a number of individual detections without knowing specifically the identity of the targets, is considered to be an important contributing factor for achieving richer, more meaningful and natural HRI. Indeed, knowing the correct number of people and their tracks at any time is a beneficial factor to the overall efficiency, safety and usability of the robotic system. The advantages are multi-faceted and include: user-friendly navigation, improved human-robot interaction (HRI) patterns and improved passive safety. Although the task of multiple human tracking has similarities to the multi-target tracking of objects, it differs significantly due to the specific motion patterns of people.

Currently, as observed by (Kahn et al. 2008), the service robots are not capable of engaging in sufficiently meaningful interactions with their users. It could be argued that this is mainly due to their inability to follow the appropriate socially established patterns for the particular context. It is considered that due to the lack of sufficient context awareness and cognitive capabilities, the service robots of today fail to approach the targeted person while efficiently observing the social norms of behaviour.

Moreover, human safety in HRI can also benefit from the ability of the robot to analyse the human behaviour patterns and make appropriate decisions about its actions.

This chapter addresses the issue of finding a suitable navigation path for a service robot. The optimisation goal is to minimise the disturbance to the existing social interactions between the members of a human group.

In particular, the above problem is tackled by the proposed method for multiple human tracking and minimally intrusive navigation path generation. The following benefits are anticipated as a result of the improved navigation. Firstly, improved availability of correct human tracking information in crowded environments. As already discussed, in such busy environments the standard distance-based data associations fail, resulting in the inability of the robot to generate the correct track for a person. Secondly, by utilising the above track information, the robot will be able to associate several past detection events and information cues to the human track and make appropriate decisions based on the above association links. For example, an intermittent detection, like a face detection, which is available only when the human faces the camera, can provide a solid information cue about the identity of the person. By combining reliable human tracking in crowded environments with the intermittent face detection, the robot can assign identity to the track and use it appropriately later even when face detection is no longer available, e.g. the person is no longer facing the robot. Finally, analysing the generated human tracks to extract context information, like interpersonal interaction affiliations, is a powerful source of context information that enables estimation of the interpersonal interaction and computation of a minimally disruptive navigation path as proposed later. Navigation in a socially compliant manner is considered to be an important aspect of the social interaction skill-set of future service robots. This chapter addresses the problem of multiple track generation from unreliable human detections, subsequent estimation of human interaction through analysis of the human tracks and finally computation of human-aware navigation paths that complies with the interaction links within the social group.

The standard approach to the problem of multiple human tracking, used in Chapter 4, is to assign an individual tracker to each target and then associate the detections within a certain radius to the track. However, when the number of people increases this approach is unable to cope with the correct detection-track associations. The failure is caused by the probability of establishing wrong data associations between tracks and detections. A single wrong association can take the track in a wrong direction and lead to further wrong associations, as demonstrated in Figure 6-2 below. Tackling the above shortcoming necessitates a more sophisticated probabilistic approach. The proposed approach relies on information from adjacent time frames to infer the correct track-detection associations. The following challenges are addressed:

Challenge one - the measurements from the sensors are noisy and uncertain, i.e. they occur with a detection probability of less than one. Also, there is always a probability of background noise and spurious detections that are close to the state of the track, e.g. clutter, or missing detections, that as a result can interfere with the correct data associations.

Challenge two - when people move in close proximity to each other, associating detection correctly to the right tracks proves to be a challenging task. This issue, known in the literature as the data association problem (Xie et al. 2004), introduces ambiguity and, potentially, leads to a wrong estimation of the track's state in the subsequent timeframes. In typical cases, the estimated track, after a wrong data association, takes a wrong direction until it becomes impossible for any further correct associations to be made, due to the increased distance from the correct position. Inability of an estimated track to recover, following a wrong association, will most likely result in a tracking failure.

Challenge three – the effects of missed detections and clutter, i.e. sensor noise have to be minimised given the irregular human motion patterns. In contrast to the unrestricted motion of the targets in a typical multi-object tracking scenario, e.g. airplanes tracked by a radar in airspace, the human motion is typically restricted by the room layout, obstacles and other people. Ignoring the above restrictions and approximating the human motion with the motion of a free object moving in space, would result in a suboptimal tracking. Therefore, in the proposed method, knowledge about the human motion patterns is included in the model to improve the tracking performance.

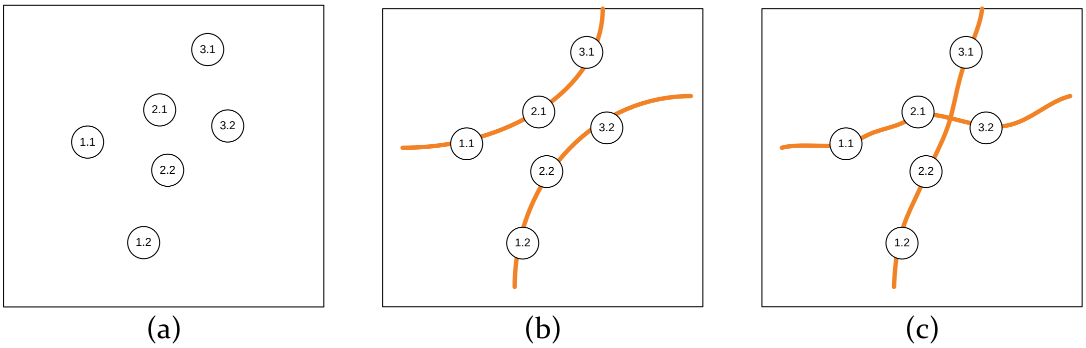


Figure 6-2: The effect of a single wrong data association on the track state progression:

(a) Detection from two people at three consecutive time frames (circles represent measurements and numbers with format {t.n} represent the time frame, t, and the index of the measurement n within the timeframe); (b) The generated tracks after correct data associations between measurements and tracks (solid line represent a track), (c) The generated tracks after a wrong data association in the third time frame.

Formally, the task, given the imperfect detections from the sensors and the intrinsic ambiguity of data associations in a typical service robotics scenario, is identified as:

- estimation of the number of targets, i.e. human tracks, that are present on the scene at each time frame;
- estimation of the data associations while eliminating or reducing the effect of the noise;
- a decision as to when to end a track in the correct time due to lack of detections without continuing to associate it with clutter;
- research of a mechanism for analysis of the estimated tracks for establishing the human group boundaries;
- research of a mechanism for estimation of the interpersonal interactions;
- computation of a navigation path, allowing the service robot to approach a particular person in a crowded environment without interrupting the interaction within the social group.

In this chapter, a framework for multi-human tracking, suitable for tracking of multiple persons in service robotics environments, and generation of an appropriate minimally intrusive path to the target person, is proposed. The input information for the framework is the set of detections generated by the multimodal human detection and localisation algorithm, described in Chapter 4. The output from the framework, i.e. the minimally invasive navigation path is used by the service robot's run-time control modules as defined by the proposed paradigm for situational awareness of the robots in regard to people.

The following contributions are made in this chapter as follows:

- a Bayesian framework for multi-human tracking from imperfect human detections that incorporates detection association into the inference process and uses human motion knowledge to eliminate the effects of clutter;
- a method for analysis of human tracks to estimate interpersonal human interaction within a human group;
- a method for the planning of a minimally intrusive path to reaching a target person within a group of people while minimising disruption to the social interaction.

This chapter is structured as follows: in section 6.2 the problem of multiple human tracking is defined, together with a definition of the system parameters that are used throughout the chapter; section 6.3 contains discussion of the assumptions made in the proposed model; in 6.4, the overall stochastic approach, taken to address the challenges is presented; in 6.5 the general system models are presented; in 6.6 the specific models, embedded into the tracking algorithm, are discussed; in 6.7 the details of a method for enhancing the reliability of the human tracking, for reducing clutter by applying human walking pattern knowledge, are given; in 6.9 the method for human group boundary identification and human track analysis is described; finally, in 6.10 a method for computation of a minimally intrusive navigation path in crowded scenes is proposed.

6.2 Problem Definition

Let T be the duration of the observation time window, during which, at each time frame, t, the robot receives a set Y_t of human detections from the human localisation classifier fusion method. In case of continuous observations, T can be a sliding window containing the latest m observations. In case of analysis of recorded past movements of the people, T is a fixed time window. Let K be the maximum number of people that appear in the observation area, denoted by R, during the observation time window T. Let a person k move in R for a fixed duration, marked with the events of birth and death of the associated track $\left[t_b^k, t_d^k\right] \subset [1, T]$. Each detection of a person is assumed to start at a random position in R at time t_b^k, i.e. the track's birth, to move independently in R until time t_d^k, the track's death, and finally, to disappear, discontinuing the track. Let p_z be the probability of the death of the track at each timeframe. Then, at each single time frame, an existing track persists with probability $1 - p_z$ and disappears with probability p_z. Let the number of new people appearing at each time over R have a Poisson distribution with a parameter $\lambda_b S$, where λ_b is the birth rate of tracks, i.e. new people observations per unit time, per unit area, and S is the area of R. Assuming that the initial position of a new human observation is uniformly distributed over R, let p_d denote the detection probability, further referred to as detection sensitivity. The false positive detections, further referred to as clutter, are modelled as a homogeneous spatial Poisson distribution with a parameter $\lambda_c S$, where λ_c is the clutter rate and S is the area of R. Finally, let N be a curve in Euclidian space. The curve N, constructed from a set of vectors $\{n_1, n_2, \dots, n_m\}$, defines uniquely the navigation path of the robot to reach the defined target. The curve N is the required output of the algorithm.

Formally, given the above definitions, the problem statement is presented as the following set of definitions:

Definition 1*: A multiple human tracking algorithm is required to take as input a set of detections from multiple sensors, at each time t, $Y_t = \{y_t^j\}_{j=1}^{M_t}$, where M_t is the number of detections at time t and j is the index of the detection within the set of detections at time t, and to produce a number of most probable tracks, $X_t = \{x_t^i\}_{i=1}^{N_t}$, where X_t is the track state, containing the position coordinates and the associations with detection, at each time; N_t is the number of active tracks at time t and i is the index of the track within the set of tracks.*

The first goal is to estimate simultaneously the distribution over all track states X_t and the data associations β_t given to the detections $Y_{1:t}$. It should be noted that the states X_t of the track over the whole time period T fully defines the birth, t_b^k, and the death, t_d^k, of each track.

In practice, the algorithm needs to cope not only with the isolated case when only the correct positive detections Y_t are present, i.e. when the presence of a single person generates only one single detection with the correct coordinates and no clutter, but also with the more general case when Y_t dataset contains some of the true positive detections intermixed with a relatively large number of false positive detections, e.g.

noise and clutter. Moreover, the number of false negatives e.g. missed detections, presents an additional challenge for the estimation of the tracks. The above complication increases the difficulty of the task estimation significantly. As explained earlier, in order to estimate the state of each track, it is necessary that the estimation mechanism takes into account the detections over the window T, as the information contained in a single time frame is insufficient for the separation of the false detections from the true positive ones. In particular, the algorithm, relying on existing knowledge of the human motion, infers stochastically the number of the active tracks at each time, their birth and death times, t_b^k and t_d^k, and the most probable human track positions, i.e. the states of the tracks.

***Definition 2**: A minimally intrusive path planning algorithm is required to take as input a set of the human track states, as defined by $X_t = \{x_t^i\}_{i=1}^{N_t}$, where N_t is the number of active tracks at time t and i is the index of the track within the set of tracks, and to generate as output a navigation plan, defined by $\{n_1, n_2, \dots, n_m\}$, for a mobile robot that minimises the interruption of the social interactions.*

The generated path is then used by the robot to initiate a human friendly navigation to the selected target person.

6.3 Assumptions

Several assumptions are made to simplify the above problem. In the context of the service robotics, these assumptions are considered to be an acceptable approximation of the reality. Moreover, making the correct decision about future actions of the robot within the specified time limit is more important than achieving an absolute accuracy of the navigation path.

- Assumption 1: It is assumed that people exist at singular points in the Euclidean space, i.e. people are represented by the coordinates of the projection of their centre of the mass on the floor;

- Assumption 2: It is assumed that all human tracks are independent of each other – in reality this will equate to the situation that each person is walking independently from the rest of the people in the environment. Although more complex social scenarios, e.g. people walking together or following a leader, are possible, these are considered outside the scope of this work;

- Assumption 3: It is assumed that clutter is a result from a random process - the random nature of clutter detection is an important feature, used to distinguish it from the true positive detections, which are dependent on the track state;

- Assumption 4: It is assumed that future track states depend only upon the present state, not on the sequence of events that preceded it, i.e. the Markovian property is valid for all tracks.

6.4 Bayesian Inference of Human Tracks

The task here is to estimate the position of the human tracks using the set of detections, available as an output from the classifier fusion method proposed in Chapter 4. Initially, the inverse task is addressed, i.e. building a probabilistic likelihood model describing how the detections depend on the underlying state. Then, the application of the Bayes rule (Bayes & Price 1763), which links the likelihood, the prior and the posterior, becomes possible by using the above model. Bayes rule allows inferring the unknown quantities, i.e. the state of the tracks at different times in the particular case, using the known values, i.e. the human detections. Fortunately, knowledge about the human motion, i.e. a model of the way in which people move and the characteristics of the sensors, can be embedded into the likelihood model in order to improve the inference process. In particular, adding specific human motion knowledge helps to eliminate some of the unlikely states, which are otherwise possible. The elimination of the unlikely states reduces the computational load of the algorithm and contributes to a more precise estimation of the unknown state.

Similarly to (Sittler 1964), a detection association event is defined as a partition of detections such that each element of the partition is a collection of detections resulting from the presence of a person. The probability of occurrence of a detection association event is embedded in the likelihood model. In particular, let ψ be a parameter of the probabilistic model, representing the detection association event. Also, let the state of the human tracks, Y be a set of all human detections, resulting either from the presence of people in the scene, i.e. true positive detections, or from clutter, i.e. false positive detections - $P(Y|\psi)$ is the likelihood of observing the detections Y, given ψ. Then from the Bayes theorem the following equation is given as:

$$P(\psi|Y) = \frac{P(Y|\psi)P(\psi)}{P(Y)} \tag{6.1}$$

The equation (6.1) links the posterior $P(\psi|Y)$, the required output, with the prior $P(\psi)$ through a factor, $\frac{P(Y|\psi)}{P(Y)}$, representing the support that the detections, Y, provide for ψ. In the particular case of fixed detection associations, i.e. when partition of detections belong to the same track, the likelihood $P(Y|\psi)$ is reduced to that of a set of single filters. It should be noted that, in principle, no closed form formula for P(Y) is feasible to be derived due to complexity issues. Instead, the posterior P(ψ│Y) can be computed, up to a normalising constant through $P(Y|\psi)$ and $P(\psi)$ and (6.1). Also, it should be noted that the prior, $P(\psi)$, is independent of the detections Y – which allows application of the Bayes rule.

Due to the unknown number of people in a setting, the likelihood is required to accommodate variable sizes of state for targets and detections. Moreover, the parameter ψ has to be defined for all possible sizes of the detection vector as specified below.

Modelling of the Data Associations

In any period of the observation window, with length of T frames, $t = 1, \dots, T$, let M be a T-dimensional vector, representing the numbers of the true detections Y_t, i.e. $M = \{M_1, M_2, \dots, M_T\}$, where $M_t \in Z = \{0,1,2,3, \dots, MM\}$, where MM is a large number and $MM < \infty$. Let j_t be a set of sets of detection indices at time t, $j_t = \{j_1, j_2, \dots j_{M_t}\}$, where $j_{jj} = \{t, jj\}$, $jj \in Z = \{0,1,2,3, \dots, M_t\}$, is pointing to the to jj^{th} detection, denoted by y_t^j, at time t (referred to hereafter as Definition 1).

Let N be also a T-dimensional vector, i.e. $N = \{N_1, N_2, \dots, N_T\}$, representing the number of active tracks at times from $t = 1$ to $t = T$, where $N_t \in Z = \{0,1,2,3, \dots, NN\}$, NN is a large number and $NN < \infty$. Let i_t be a set of track indices at time t, $i_t = \{i_1, i_2, \dots i_{N_t}\}$, where $i_{ii} = \{t, ii\}$, $ii \in Z = \{0,1,2,3, \dots, N_t\}$, is pointing to the to ii^{th} track, identified by its state x_t^i, at time t, as defined in Definition 1.

For each time frame t, a set of links $\vartheta_t = \{\vartheta_{i,t}^j\}$ is defined, where $i \in \{1, .., N_t\}$, $j \in \{1, .. M_t\}$ and $\vartheta_{i,t}^j$ is the link between i^{th} active track ,i.e. x_t^i, and a j^{th}detection,i.e. y_t^j, at a time t, $\vartheta_{i,t}^j = \{t, ii, jj\}$, $jj \in Z = \{0,1,2,3, \dots, M_t\}$, $ii \in Z = \{0,1,2,3, \dots, N_t\}$. Let $\vartheta = \bigcup_{t=1}^{T} \vartheta_t$ be an index set of links whose size matches the number of possible combinations of tracks and detections, i.e. $M_t N_t$. It should be noted that only active tracks are allowed in the method to be associated with detections.

Then, the detection association set, ϐ, is defined as a collection of partitions of ϑ, i.e. ϐ $= \{\vartheta_{i,t}^j\}$, for which the following conditions are met:

1. $\vartheta_{i,t'}^j \neq \vartheta_{i,t''}^j$, for $t' \neq t''$, $t' = 1$ to T, $t'' = 1$ to T, $i \in Z = \{0,1,2,3, \dots, N_t\}$, $j \in Z = \{0,1,2,3, \dots, N_t\}$,i.e. a target at a specific time cannot be associated with detections from any other time-frame;
2. At time t, $\bigcap \theta_{i,t} = \emptyset$, $i \in Z = \{0,1,2,3, \dots, N_t\}$, i.e. in a single time-frame no two tracks can be associated with one detection;
3. For any target i, $\sum_{j,t} \vartheta_{i,t}^j \geq 2$, i.e. any track must be associated with at least two detections so it can exist.

An Illustrative Example

For the illustration of the problem of data association, a simplified hypothetical scenario is presented below. The scenario consists of an observation period of two time frames, i.e. $T = 2$, with the first one, at time $t = 1$, it has three detections and in the second one, at time $t = 2$, it has four detections, i.e. $M = \{2,4\}$. In this case the detection index sets are given as: $j_1 = \{(1,1), (1,2), (1,3),\}$ and $j_2 = \{(2,1), (2,2), (2,3), (2,4)\}$. Assuming that there are two tracks that start from time $t = 1$ and finish at time $t = 2$, i.e. both tracks are active during the whole observation period, the number of possible data associations can be listed. In Figure 6-3, one of the possible combinations for the association set is shown, i.e. $\beta = \{(1,1,1), (1,2,3), (2,1,3), (2,2,2)\}$, to illustrate the possible combinations. As it can be observed from Figure 6-3, there are 48 (2*3 combinations, for $t = 1$, and 2*4 combinations, for $t = 2$,) possible combination, without taking into account the likelihood for missed detections, i.e. both tracks are associated at any time. In this example, out of all 48 possible data association, there is a single association combination that connects the right detections with the right tracks. This combination maximises the joint probability from the likelihood model that these detections are generated from the tracks and the detection associations. In particular, the goal of the proposed method is twofold: a) estimation of the best data association, i.e. the one that maximizes the probability for generating the detections from tracks;

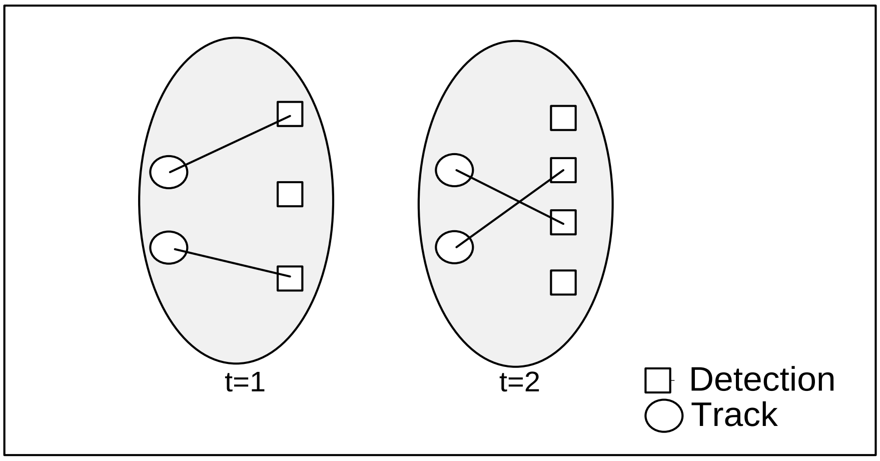


Figure 6-3: Illustrative example of data associations between two tracks and detections

b) inferring the best estimate for the state of the tracks, i.e. the position of the people in the scene that maximises the joint probability for the given detections. Both goals are interlinked and solved jointly as described below.

6.5 The General Tracking Model

In this section, the general tracking model of the multiple human tracking is described. The model serves as a foundation, on which the specific models, i.e. the state transition model, the observation model and a data association model are added at a later stage. However, the role of the general tracking model is to specify the relationship between the state, X, the parameters and the detections, Y.

A new track starts when a new cluster of detections, which cannot be associated with any of the existing tracks, appears in the observed area. In the model, the time of this event is recorded as the time of birth of the track, i.e. by using the variable t_b^k, where k is the index of the track. The track continues to exist, either in an active state, if there are detections which can be associated with it, or in an inactive state, when there are no such detections in existence. In particular, the track continues to exist until either the end of the observation window, T, is reached, or a sufficiently long sequence of missed detections is detected and a decision is made by the method to end the track. The maximum number of possible missed detections is linked in the model with the detection probability of the sensor and the required level of confidence is reached, as described in 6.7. When this happens, the time of death, t_d^k, of the track is reversed back to the time-frame that follows the last detection.

An illustrative example of the key principle of the above mechanism is shown in Figure 6-4. In the example, four people are present in the scene and they are observed by the sensors of the robot. However, as in reality their output contains noise, missed detections or spurious detections are possible. In the figure, there is a number of true positive detections, marked with the "+" symbol, as well as false negative detections, marked with the "−" symbol to illustrate the missed and the spurious detections. In addition to the detections originating from human presence, other spurious detections, i.e. clutter resulting from noise can also be reported falsely as detections, marked as squares in the figure to indicate clutter, are also present. As the clutter and the true positive detections are intermixed, they are indistinguishable in the input data, indicated by the continuous detection index. Due to the ambiguity introduced by mixing of clutter and true positive detections, a probabilistic reasoning mechanism, able to infer the most probable combination of assigned tracks and detections while avoiding clutter, is required. In (6.4), one such set of probable detection associations is illustrated using dotted lines.

In the above example, all probable data associations can be enumerated, as illustrated in the table of associations, shown in Table 6-1 below. The table lists all possible combinations between tracks and detections, both true positive and false positive, at time t. In the table, if there is a valid association between a track and a detection then the value of the corresponding cell is set to one, otherwise it is set to zero. A similar encoding format is used by the method to store the possible data association combinations when inferring the most probable track associations.

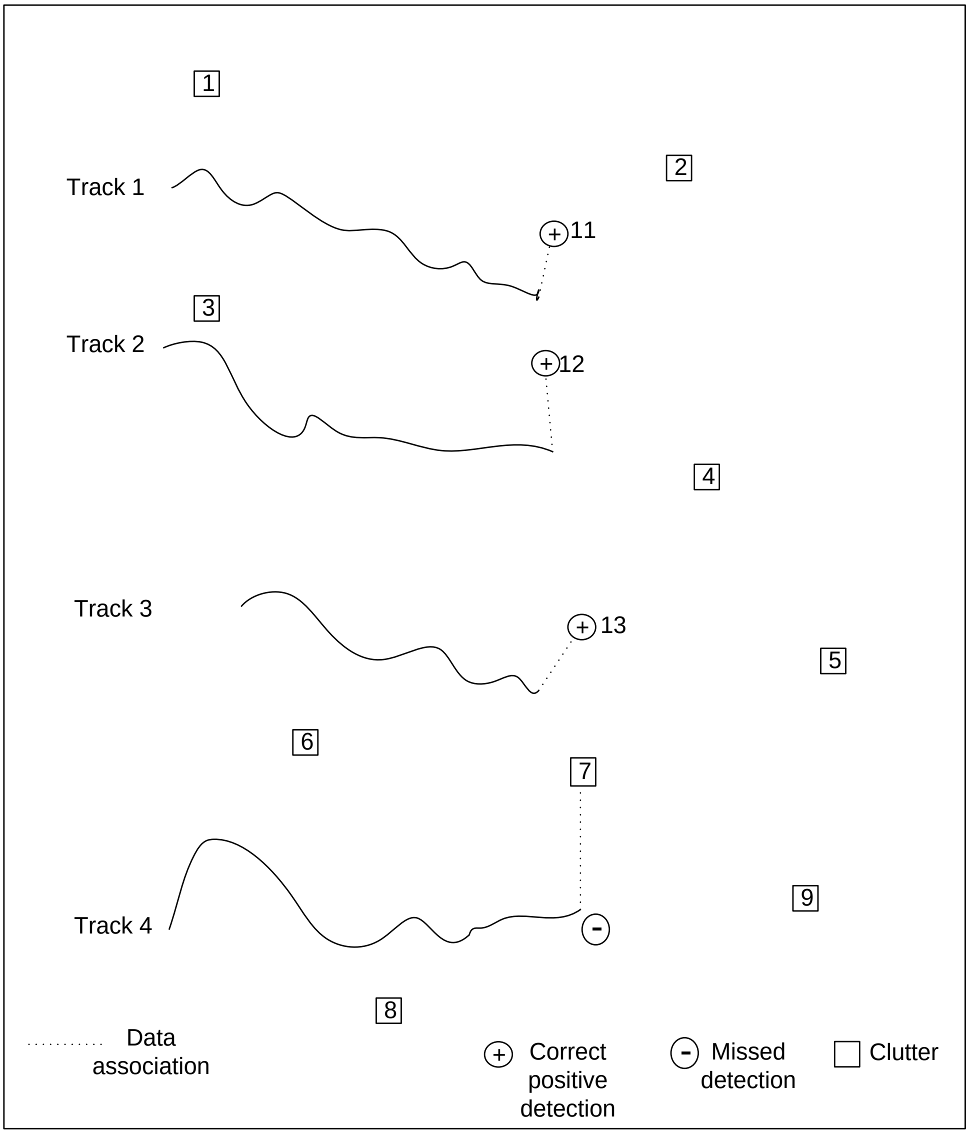


Figure 6-4: Example of tracks, observations, detections and possible associations

	Det. 1	Det. 2	Det. 3	Det. 4	Det. 5	Det. 6	Det. 7	Det. 8	Det. 9	Det. 10	Det. 11	Det. 12	Det. 13
Track 1	0	0	0	0	0	0	0	0	0	0	1	0	0
Track 2	0	0	0	0	0	0	0	0	0	0	0	1	0
Track 3	0	0	0	0	0	0	0	0	0	0	0	0	1
Track 4	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 6-1: The table of association consisting of binary association variables linking tracks with true positive and false positive detections

Since future states of the system depend only upon its current state, and not on the sequence of events that preceded it as defined in the assumptions earlier, a Hidden Markov Model (HMM) can be used to link the state of a track $\mathrm{X_t}$, and the observable detections, $\mathrm{Y_t}$, that have been associated with the track up to time t. In particular, in the model the state evolves according to the Markov process with an initial state $\mathrm{x_0 \sim p(x_0)}$ and transition kernel $\mathrm{P(X_t | X_{t-1})}$ for $\mathrm{t = 0, 1, \dots T}$ as given below:

$$X_t \sim P(X_t | X_{t-1}) \ , \tag{6.2}$$

where the transition kernel $\mathrm{P(X_t | X_{t-1})}$ is a variable that explicitly depends on the time index, t. Because of the Markovian property, a detection is conditionally independent of any previous detections, given the current state. The above property allows modelling of the likelihood as described below.

6.5.1 Model with Fixed Data Associations

In the ideal case, when the tracks are far apart from each other and the number of false positive detections is negligibly small and, therefore, unable to influence the track state to any significant extent, an approach based on fixed data associations is considered to be sufficient for the purpose of multiple human tracking. In this case, the following observation model using the likelihood distribution $P(Y_t|X_t)$ for $t = 0, 1, \dots T$ is given:

$$Y_t \sim P(Y_t|X_t), \tag{6.3}$$

where the observation likelihood Y_t varies with the time index t.

A graphical model of the HMM, displaying three time frames with the respective state and observations of the sequential progression, is shown in Figure 6-5 to illustrate the above simplified problem.

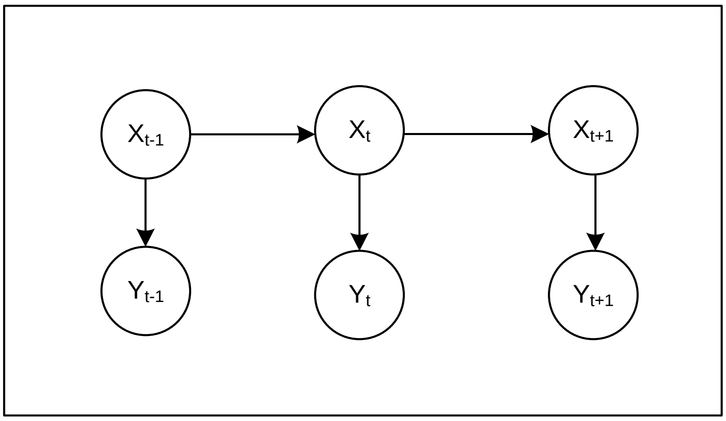


Figure 6-5: A graphical model of the fixed associations HMM

In this case, a model can be constructed based on recursive Bayesian estimation (Bergman 1999).

Although the above HMM, representing a data fixed association case, is not sufficient for real world multiple human estimations, it acts as a foundation for the improved model in the next section.

6.5.2 Model with Variable Data Associations

In reality, the data associations are not fixed and generally are not even known in advance. Therefore, an association variable is introduced in the model to account for the unknown variable data associations. As a result, in the new model, the link with the detections, Y, is influenced by the association variable, in addition to the state of the latent variable X. Therefore, instead of the HMM, shown in Figure 6-5, which models only the state under fixed data associations, a Dynamic Bayesian Network (DBN), shown in Figure 6-6 is used. Replacing the HMM with a DBN allows the inclusion of additional parameters into the model. The new parameters, θ, represent the variable detection associations, and μ, represents the influence of the environment on the state variable. As before, only three time-frames from the time series are shown on the figure, i.e. those with time index values: $-1, t, t+1$.

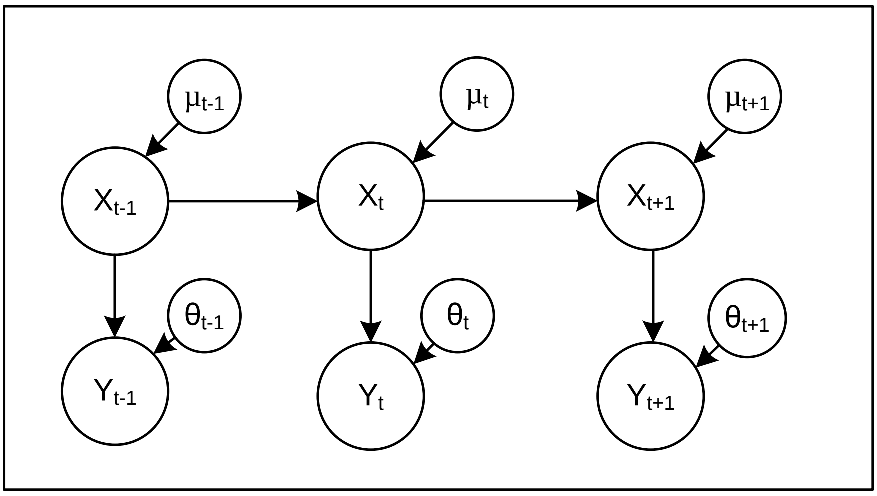


Figure 6-6: The graphical model of the system with variable associations.

After (9.63) and the joint probability are taken into account, the equation for the target posterior distribution is given as:

$$P(X_t, \theta_t \,|Y_{1:t}) = \frac{P(Y_t|X_t,\theta_t)P(X_t,|X_{t-1})P(\theta_t)P(X_{t-1}, \theta_{t-1} \,|Y_{1:t-1})}{P(Y_t|Y_{1:t-1})} , \qquad (6.4)$$

where $P(Y_t|X_t,\theta_t)$ is the likelihood given the associations, $P(X_t,|X_{t-1})$ is the transition density, $P(\theta_t)$ is the association likelihood, $P(X_{t-1},\theta_{t-1}\,|Y_{t-1})$ is the prior and $P(Y_t|Y_{1:t-1})$ is a normalising constant. The above terms are discussed in the following sections.

6.6 Specific Models of Multi-Human Tracking

6.6.1 Likelihood Model

The likelihood, given the associations, $P(Y_t|X_t,\theta_t)$ is modelled in considering what effects the data associations θ_t have on the detections, Y_t.

As the detections are independent of one another and also independent of clutter detections, the likelihood of a detection, $P(Y_t|X_t,\theta_t)$, can be factorised into two terms, a clutter likelihood term, $P_{clutter}$, and a real detection term, P_{real}, as given below:

$$P(Y_t|X_t,\theta_t) = P_{clutter}\,P_{real} \tag{6.5}$$

Since a false positive detection occurrence is the result of a random process, the clutter likelihood depends only on the number of clutter detections and the area of the field of view of the sensors. For example, when the observation area is increased, given a fixed number of false positive detections, the probability of clutter per unit area gets smaller as the same number of false positive detections, M_t are spread over the bigger area. Therefore, the clutter detection likelihood is modelled as:

$$P_{clutter} = \frac{1}{S}\,M_t, \tag{6.6}$$

where S is the area of the field of view of the sensors, M_t is the total number of false positive detections in the scan at time t.

In the above equation, as the observation area is a constant and the number of clutter detections is also fixed, M_t, the likelihood of clutter $P_{clutter}$ is a value which is not dependent on the state X_t.

As the tracks are independent of each other, the likelihood of a real detection is factorised as a product of the detection likelihoods of the separate tracks. It should be noted that only detections that belong to the specific track need to be included in real detection term P_{real}. This is achieved by inclusion of the association variable, θ, as an index to the detection, i.e. $y^{\vartheta^j_{i,t}}$. The reason is that $y^{\vartheta^j_{i,t}}$ represents the set of detections that the specific association variable $\vartheta^j_{i,t}$ links detection j with the track i at time t. Therefore, $P(y^{\vartheta^j_{i,t}}|x_{i,t})$ is the likelihood of a track i to generate the detection j to which the association variable $\vartheta^j_{i,t}$ is pointing to. It should be noted that only valid associations can be included, i.e. associations for which $\vartheta_{i,t}^{j} \neq 0$. Therefore, the likelihood for real detections, P_{real}, is given as:

$$P_{real} = \prod_{i=1,j=1,\vartheta\neq 0}^{N_t} P(y^{\vartheta_{i,t}^{j}}|x_{i,t}) \, , \tag{6.7}$$

where i is the track index, N_t is the number of valid tracks, $y^{\vartheta_{i,t}^{j}}$ is the detection with index j that is linked to $x_{i,t}$ by $\vartheta_{i,t}^{j}$.

After substitution (6.6) and (6.7) into (6.5) the following equation for the likelihood is given as:

$$P(Y_t|X_t, \vartheta_t) = \frac{M_t}{S} \prod_{i=1,j=1,\vartheta\neq 0}^{N_t} P(y^{\vartheta_{i,t}^{j}}|x_{i,t}) \, , \tag{6.8}$$

In the above equation, the first term, $\frac{M_t}{S}$, is a coefficient, which is independent of the state, X_t, and as such it cancels out in the subsequent MCMC-PF simulation, Therefore, the term $\frac{M_t}{S}$ is replaced by $Coeff_1$ in the subsequent equations.

When the track becomes inactive, e.g. in time frames where no detection can be associated with the particular track, the association variable is set to zero, i.e. $\vartheta = 0$. This occurence represents transient disconnection of the track state from the detections. In such a case, the track likelihood $P(Y_t|X_t, \theta_t)$ is reduced only to the likelihood for the clutter detections. This is introduced into the model as:

$$P(Y_t|X_t, \vartheta_t) = Coeff_1 \prod\nolimits_{i,j} \begin{cases} P(y^{\vartheta_{i,t}^{j}}|x_{i,t}) & \vartheta_{i,t}^{j} \neq 0 \\ 1/_S & \vartheta_{i,t}^{j} = 0 \end{cases} , \tag{6.9}$$

where i is the track index, j is the detection index and $\vartheta_{i,t}^{j}$ is the association variable.

6.6.2 Transition model

As the human tracks are independent of each other, described earlier in 6.3, the transition distribution, $P\,(X_t, X_{t-1})$, linking two consecutive states, is factorised as a product of the transition distributions of the separate tracks:

$$P\,(X_t, X_{t-1}) = \prod_{i=1}^{N_t} P\,(X_{i,t}, X_{i,t-1}) \, , \tag{6.10}$$

where i is the track index and N_t is the number of active tracks at time t.

6.6.3 Association likelihood

The association variable $\vartheta_{i,t}^{j}$ represents the link between the track i with the detection j. It also implicitly encodes information about the number of tracks and the number of detections that have been associated with the tracks, i.e. by the range of the indexes i and j.

Overall, there are three possible cases for the association variable $\vartheta_{i,t}^{j}$:

- Case A: (the ideal case) a correct detection is made from an existing track;
- Case B: a correct detection has not been made but a nearby clutter detection exists and can be assigned to the track
- Case C: neither a true positive nor a false positive detection, i.e. clutter, exists for this track and, therefore, no detection association is possible.

As the above cases are independent, the association likelihood can be factorised as:

$$P(\theta_t) = P_{detected\_tracks}(\theta_t)P_{undetected\_tracks}(\theta_t)P_{clutter}(\theta_t) \qquad (6.11)$$

Detailed considerations are presented below for the three possible cases for the association variable θ_t, by splitting them further into a number of subcases, as follows:

- ***Case A, Subcase* 1** (all tracks are detected and a valid detection is associated with each of them): the track i exists and the person has been successfully detected, i.e. a true positive detection. The probability for this happening is given by the parameter p_d, which is part of the sensor's characteristic. Because the assignment can be made successfully, the association variable $\vartheta_{i,t}^{j}$ is valid, i.e. $\vartheta_{i,t}^{j} \neq 0$, and links the i^{th}track with the j^{th} detection. In this case, the corresponding factor of the association prior in (6.11) is denoted by $P_{detected\ tracks}(\theta_t)$. For each correct detection, $P_{detected\ tracks}(\theta_t)$ can be calculated by dividing the detection probability by the number of the remaining detections, i.e. those in the scan that have not been assigned yet. For example, the likelihood for the detection probability of the detection of the first track will be $\frac{p_d}{M_t}$, for the second one the number of the observations is reduced by one so the probability is given as $\frac{p_d}{M_t-1}$, for the third $\frac{p_d}{M_t-2}$ and so on, until the last detected track when the detection probability will be $\frac{p_d}{M_t - M_{assigned}}$.

  The combined probability of detecting all tracks is given as a product of the separate probabilities for the individual tracks:

$$P_{detected\ tracks}(\theta_t) = \\ = \frac{p_d}{M_t}\frac{p_d}{M_t-1}\frac{p_d}{M_t-2}\dots\frac{p_d}{M_t - M_{assigned}} \qquad (6.12)$$

After rearrangement of (6.12):

$$P_{detected\ tracks}(\theta_t) = \\ = \frac{p_d{}^{(M_{assigned}+1)}}{M_t(M_t-1)(M_t-2)\ \ldots..(M_t-M_{assigned})}, \quad (6.13)$$

where M_t is the total number of tracks, $M_{assigned}$ is the number of assigned tracks.

- **Case A, Subcase 2** (there are one or more missed detections leading to associations with a clutter detection): The track i exists but the associated person is not detected by the sensors yet. However, a false positive detection has occurred in a close proximity and therefore is associated with the track. Similarly to the case A above, an association is made, although a wrong one, and therefore the association variable $\vartheta_{i,t}^{j}$ is valid, i.e. $\vartheta_{i,t}^{j}\ \neq 0$. The only difference is that it links the i^{th}track with the j^{th} detection, which happens to be a clutter detection. In this case the corresponding factor in the association likelihood is denoted by $P_{clutter}(\theta_t)$.

  Once all the tracks have been processed, the difference between the number of detections and the assigned detections as reported by the counters M_t and $M_{assigned}$ results in the number of clutter detections in the scene:

$$M_{clutter} = M_t - M_{assigned}, \quad (6.14)$$

  where M_t is the total number of tracks, $M_{assigned}$ is the number of assigned tracks.

  Assuming that clutter detections has a Poisson distribution with parameter λ_c, the following probability mass function can be given:

$$P_{clutter}(\theta_t) = \frac{e^{-\lambda_c}\lambda_c{}^{(M_{clutter})}}{M_{clutter}!}, \quad (6.15)$$

  where $M_{clutter}$ is the number of clutter detections.

- Case B (some of the human targets are undetected due to false negatives detections and no association is possible, neither with a real detection nor with a clutter detection): The track i exists but due to a missed detection it cannot be associated. The probability for this to happen is given as: $(1-p_d)$. Furthermore, there are no suitable clutter detections that can be associated with the track and the association variable is left unassigned, i.e. $\vartheta_{i,t}^{j} = 0$. In this case the corresponding factor of the association likelihood is denoted by $P_{undetected\ tracks}(\theta_t)$.

  The probability of a false negative detection for a track is given by:

$$p_{fn} = (1-p_d), \quad (6.16)$$

  where p_d is the sensitivity of the detector.

The number of tracks associated with false negative detections is denoted by $M_{m,t}$. In case of a false negative detection, no valid value can be assigned to the association variable $\vartheta_{i,t}^{j}$, as it is impossible to link the track with a non-existent detection.

In case of multiple consecutive false negative detections, the probability is given as a product of the individual probabilities for a missed detection event:

$$P_{undetected\ tracks}(\theta_t) = \prod_{m=1}^{M_{m,t}} (1 - p_d), \tag{6.17}$$

where $M_{m,t}$ is the number of tracks currently associated with false negative detections and p_d is the sensitivity of the detector.
Since the individual events have equal probabilities the above equation can be given as:

$$P_{undetected\ tracks}(\theta_t) = (1 - p_d)^{M_{m,t}} \tag{6.18}$$

After the substitution of (6.18),(6.15) and (6.13) into (6.11), the equation for the association likelihood is given as:

$$P(\theta_t) = \frac{{p_d}^{(M_{assigned}+1)}(1 - p_d)^{(M_{m,t})}}{(M_t - M_{assigned})M_t!} e^{-\lambda_c} {\lambda_c}^{(M_{clutter})}, \tag{6.19}$$

where λ_c is the parameter of Poisson distribution of clutter.

After re-arrangement of (6.19) the following equation can be given:

$$P(\theta_t) = \\ = \frac{e^{-\lambda_c}}{(M_t - M_{assigned})M_t!} \prod_{j=1}^{M_t} \begin{cases} p_d & j \in M_{m,t}, \vartheta_{i,t}^{j} \neq 0 \\ (1 - p_d) & j \notin M_{m,t}, \vartheta_{i,t}^{j} \neq 0 \\ \lambda_c & \vartheta_{i,t}^{j} = 0 \end{cases} \tag{6.20}$$

where $\vartheta_{i,t}^{j}$ is the association variable between j track and i detection at time t, $M_{m,t}$ is the number of tracks currently associated with false negative detections and M_t is the total number of tracks.

As the factor $\frac{e^{-\lambda_c}}{(M_t - M_{assigned})M_t!}$ is constant that does not depend on the state, after replacing it with a coefficient, $Coeff_2$, (6.21) can be given as a product of factors for each track.

$$P(\theta_t) = Coeff_2 \prod_{j=1}^{M_t} \begin{cases} p_d & \vartheta_{i,t}^{j} \neq 0 \\ (1 - p_d) & j \in M_{m,t}, \\ \lambda_c & \vartheta_{i,t}^{j} = 0 \end{cases} \tag{6.21}$$

The coefficient $Coeff_2$, like the $Coeff_1$ from the observation likelihood in (6.9), cancels out later in the MCMC-PF filter.

6.6.4 Human Physical Motion Model

When people are not obstructed by any obstacles, there is no restriction to the human motion pattern from the environment. In these cases, the human motion is modelled by the standard Brownian motion model, characterised by an acceleration which is a continuous Wiener time stochastic process with the following three properties (Peng & Hsu 2012):

- an initial value $W_0 = 0$
- a continuous function W_t
- independent increments of W_t with normal distribution that has expected value zero and variance $t - s$, where t and s are points in time so that $0 \leq s \leq t$ is true.

From the properties of the Wiener function, the following stochastic differential equation for the acceleration of the person is given:

$$\frac{d\dot{c}_t}{dt} = \varepsilon dW_t, \tag{6.22}$$

where c_t is the position in time t. After re-arrangement and integration (6.22) can be given as:

$$\dot{c}_t = \dot{c}_0 + \varepsilon W_t, \tag{6.23}$$

which is interpreted that the velocity starts with some initial value $\dot{c}_0$ and the temporal changes depend on the function εW_t.

Additionally, from the definition of the velocity $\dot{c}_t$:

$$\dot{c}_t = \frac{dc}{dt} \tag{6.24}$$

After substituting (6.24) into (6.23) and subsequent integration the following equation can be given:

$$c_t = c_0 + \dot{c}_0 t + \varepsilon \int_0^t W_t \, dt \tag{6.25}$$

Then, from (6.23) and (6.25) the following matrix form can be given:

$$\begin{bmatrix} c_t \\ \dot{c}_t \end{bmatrix} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_0 \\ \dot{c}_0 \end{bmatrix} + \varepsilon \begin{bmatrix} \int_0^t W_t dt \\ W_t \end{bmatrix} \tag{6.26}$$

The equation (6.26) is the Human Physical Motion Model given in a matrix form.

6.6.5 Observation Model

The position of people within the field of view, represented by the state X_t of the tracks as specified by Definition 1, is detected with probability defined by the parameter p_d. The resulting detections are stored in the vector Y_t. A Hidden Markov Model (HMM) is used to model the link between the hidden variables and the detections. In general, the observation model, linking the hidden states in the HMM with the detections, is given as:

$$Y_t \sim P(Y_t, X_t) \tag{6.27}$$

Then, the posterior distribution over the track states is calculated by marginalising the associated variables Θ_t, at time t:

$$P(Y_t, X_t) = \sum_{i,j} P(\,Y_t | X_t, \Theta_t) P(\Theta_t)\,, \tag{6.28}$$

where the summation is over the indexes of all valid values of the associated variables $\theta_{j,t}^{(i)}$, i.e. all valid i, j values from the detection and track indices.

A normal distribution, with a mean equal to the state of the track x_t and covariance R, is used to model the random noise in the classifier fusion method. Then, the observation model of the particular track is given as:

$$P(y_t | x_t) = \mathcal{N}\,(y_t |\, C\, x_t, R)\,, \tag{6.29}$$

where C is the following matrix:

$$C = \begin{Bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{Bmatrix} \tag{6.30}$$

6.7 Clutter Reduction through Human Motion Constraints

The human body follows the laws of physics, as any other physical body with mass. Therefore, there are certain constraints over the motion dynamics of the human body, i.e. its non-zero mass and limited power that restrict the maximum acceleration that the human body can reach. The above limitations allow elimination of the clutter that is positioned beyond the reachability limit, i.e. the boundary in the Euclidean space that can be reached by the human starting from its current estimated position.

Due to variations between individuals, each person has different physical characteristics that determine how fast they walk and, given their starting position, their limit of reachability. For example, a fit person can reach much further compared with a frail elderly person. Therefore, taking into account the specific individual characteristics, which differ between people and age groups, it is possible to customise the motion model to the individual physical abilities of the individual. This customisation is achieved by embedding dynamic parameters into the individual's motion model. These parameters play an important role in filtering out the majority of the clutter detections as explained below. Additionally, in a borderline case, e.g. detection that is still at a reachable but not very probable position, the joint posterior distribution for the particular combination of data associations and track state takes into account the reduced probability of the person reaching the particular location. Through the above parameters, prior knowledge about the real world characteristics of human motion is embedded into the probabilistic motion model and, as a result, contributes to improving the inference process.

In particular, under normal circumstances people tend to walk at a certain speed, often referred to as the preferred walking speed. Although humans are capable of walking at a wide range of speeds, e.g. from 0 to 2.5 m/s (9.0km/h; 5.6 mph), under normal circumstances, they tend to use only a narrow subrange of the above range of walking speeds. In particular, a number of factors, e.g. age and sex, have been reported in the literature to determine the preferred human walking speed (Ayis et al. 2007; Lord et al. 2005; Smith 1995; Ando at al.). The results of these studies are graphically summarised by (Spearpoint & MacLennan 2012) as shown in Figure 6-7.

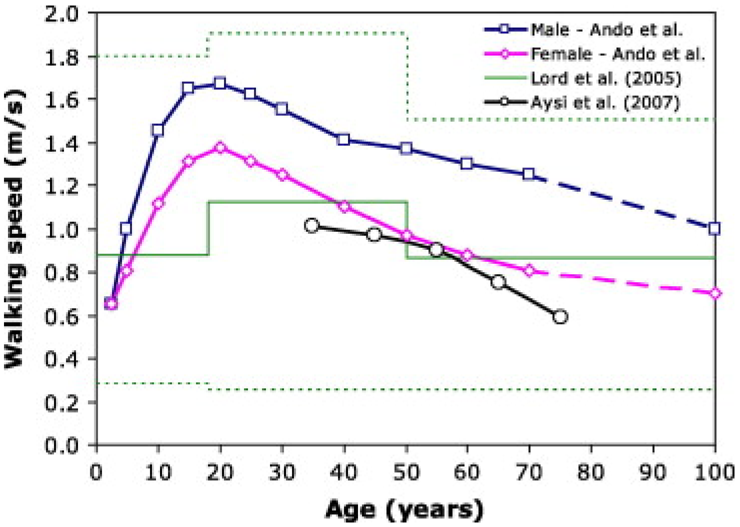


Figure 6-7: Comparison of walking speed as a function of gender and age from an adaptation of Ando et al.

Clutter Elimination

People do not always walk in a straight line. Moreover, they may pause at any moment for a random period of time and then continue to walk. Therefore, only the maximum limit of the walking speed can be used to check the validity of the detections. This check is done through the introduction of a parameter, V_{max}, representing the maximum walking speed of the tracked people. In particular, using this parameter the model is able to eliminate the false positive detections which are made beyond that person's reachable limit. This significantly simplifies the inference and reduces the computational overhead. The process of elimination is illustrated in Figure 6-8, which shows the estimated state of a track at time t and three detections: $y_{1,1}$, $y_{1,2}$, $y_{1.3}$. The detection, denoted by $y_{1,2}$, where the first index is the time frame, second one is the index of detection among all detection, is deemed invalid by the algorithm as it is positioned outside the circle of reachability and therefore discarded. The circle's centre is set to be at the current estimated state of the track and its radius is equal to the distance:

$$d_1 = V_{max}.t_{scan} + d_{error} , \tag{6.31}$$

where t_{scan} is the time between the two scans and d_{error} is a distance taking into account possible error margin in the estimation of the track state.

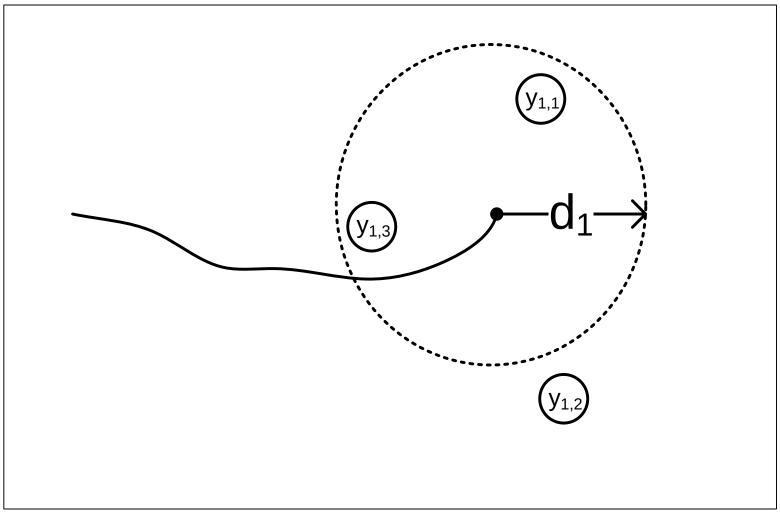


Figure 6-8: Maximum distance limitation after a single time frame.

Although the lack of suitable detections that can be associated with a track may suggest that the track has ceased to exist, this hypothesis cannot be confirmed or rejected immediately by the algorithm because of the possibility of missed detections. Therefore, the algorithm continues to compute its state for a period of time while trying to find possible data associations with detections, instead of ending the track immediately after the first failure to discover a suitable detection. In particular, when there are no detections, the state of the track is automatically updated by the filtering part of the algorithm. However, it is possible that a sudden change in the human motion, e.g. a change in the speed or direction of moving, causes the detection to occur at an unexpected place. Therefore, instead of the default behaviour of just redrawing the circle of maximum reach, i.e. the reachability boundary, using the predicted future state of the track, the circle of maximum reach is expanded without moving its centre. In particular, the algorithm increases the radius of the circle to correspond to the maximum distance that the person can move for the elapsed period, given the maximum speed V_{max} as shown for a circle with a radius d_1 in timeframe $t + 1$ in Figure 6-9 .

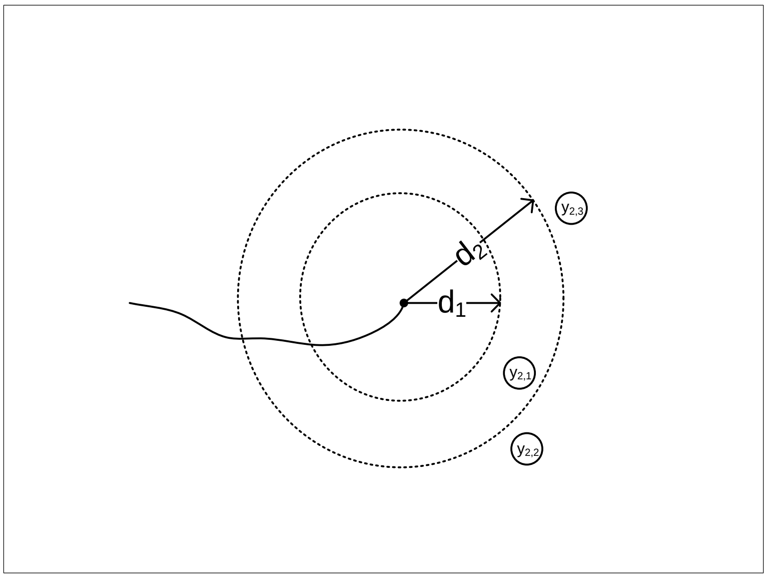


Figure 6-9: Maximum speed limitation after two timeframes

The radius of the circle of maximum reach is calculated as:

$$d_n = V_{max}.t_{scan}.n + d_{error} , \tag{6.32}$$

where n is the number of frames where no detections have been associated with the track.

Without associations with a detection, the above process of expansion of the maximum reach circle will be repeated until it covers the whole observable space. Therefore, a criterion is needed to make a decision as to when to end the track. The probabilistic method, taking into account the properties of the required confidence of the decision, is presented below.

Ending of a track

A method for making a decision regarding when to end a track, based on the sensitivity of the combined detector p_d, is proposed below.

If the sensitivity of the detector is denoted by p_d, then the probability of not detecting a target is represented by the probability $(1 - p_d)$. The joint probability for continuously missing detections in sequential frames, P_{ms}, depends on the sensitivity of the sensor and is given by:

$$\mathrm{P_{ms}} = (1 - \mathrm{p_d})^{M}, \quad (6.33)$$

where M is the number of sequential frames without detection and $\mathrm{p_d}$ is the sensitivity of the detector. As the probability, $\mathrm{p_d}$ ranges in value from zero to one, i.e. $0 < \mathrm{p_d} < 1$, it can be observed from (6.33) that the probability of the existence of a track after a number of consecutive frames without detections decreases rapidly over time. For example, if $\mathrm{p_d}$ is 0.7, the probability for two missed detections ($M = 2$) occurring in a succession is 0.09; for three it is 0.027; for four: 0.0081; for five: 0.00243 and so on.

As demonstrated by the above example, it becomes increasingly unlikely with every additional timeframe that the missed detections are the result of detector error. At the same time, with every new timeframe without a detection, the probability that the track has ended increases. This reasoning is the basis for the proposed method, based on Bayes theorem (Stuart & Ord 1987), for deciding when to end a track.

The key principle of the method is illustrated in Figure 6-10. In addition to the sensitivity of the detector, its specificity, p_{fd}, is also used in the decision of when to end the track. Although an ideal detector would be characterised as 100% sensitivity and 100% specificity, a real detector is characterised by parameters with lower values, resulting in a certain non-zero Bayes error rate (Stuart & Ord 1987).

The overall operation of the method is based on construction of a circle with a centre at the last estimated position of the track and radius based on (6.32). If no detection is made within the circle in the current time frame the posterior, i.e. the probability for existence of the track, is calculated using the Bayes theorem in (6.1) and illustrated by an example below. If the probability for existence of the track is below a certain predefined threshold the track is ended. Otherwise, in the next timeframe the circle is expanded and the process is repeated until the existence probability of the track becomes sufficiently small.

An Illustrative Example

Assuming that the prior, $P(Track)$ that exists at time t, has a value of 0.9. It is also known that the sensitivity of the combined detector is equal to 0.8, i.e. $p_d = 0.8$ and its specificity is equal to 0.9, i.e. $p_d = 0.9$. From the above the following four possible probability combinations about presence of a track can be given:

$$P(\,Detection|\,Track) = 0.8 \quad (6.34)$$
$$P(\urcorner Detection|\,Track) = 0.2 \quad (6.35)$$
$$P(\,Detection|\urcorner Track) = 0.1 \quad (6.36)$$
$$P(\urcorner Detection|\urcorner Track) = 0.9 \quad (6.37)$$

Also by using $P(Track)$ the following can be given for the probability of a track being not live:

$$P(\neg Track) = 1 - 0.9 = 0.1 \tag{6.38}$$

If no detections are made in the following scan, at time t, within the expanding search circle, it is possible to update the belief about the existence of a track by taking into account the above observation event. Indeed, the posterior $P(\neg Track|\neg Detection)$, which gives the necessary probability about the track ending, can be computed by applying the Bayes' rule for the circle as follows:

$$P(\neg Track|\neg Detection) = \frac{P(\neg Detection|\neg Track)P(\neg Track)}{(P(\neg Detection|Track)P(Track) + P(\neg Detection|\neg Track)P(\neg Track))} \tag{6.39}$$

After substitution of the values from equations (6.34),(6.35),(6.37),(6.38) into (6.39), the following can be given:

$$P(\neg Track|\neg Detection) = \frac{0.9 * 0.1}{(0.2 * 0.9 + 0.9 * 0.1)} = 0.333 \tag{6.40}$$

Similarly, the posterior for the existence of a track can be calculated as:

$$P(Track|\neg Detection) = \frac{P(\neg Detection|Track)P(Track)}{(P(\neg Detection|Track)P(Track) + P(\neg Detection|\neg Track)P(\neg Track))} \tag{6.41}$$

From (6.41) after substitution the following posterior can be calculated:

$$P(Track|\neg Detection) = \frac{0.2 * 0.9}{(0.2 * 0.9 + 0.9 * 0.1)} = 0.666 \tag{6.42}$$

As the sum $P(Track|\neg Detection) + P(\neg Track|\neg Detection)$ always equates to a probability of 1, the following equation can be given:

$$P(Track|\neg Detection) = 1 - P(\neg Track|\neg Detection) = 0.333 \tag{6.43}$$

By analysing (6.42) and (6.43), it can be observed that a scan without detection in any particular cell, i.e. a negative detection, reduces the belief in the continued existence of the track. In particular, the probability for existence of a track decreases from a prior, 0.9 (90%), to a posterior, 0.666 (66.6%) in the first frame without detection while the probability of absence of a track increases from 0.1 (10%) to 0.333 (33.3%). This trend continues if the period of no detections continues into the following frames as follows: at time $t + 1$, the posterior for a presence of a track becomes 0.3 (30%) and the one for absence of a track 0.7 (70%). The results from this example are graphically illustrated in Figure 6-10.

As demonstrated by the above example, it is possible to incorporate any combination of positive and negative detections from the sensors/detectors, by applying the Bayes rule sequentially. This approach results in a collection of posteriors giving the presence of the track at any of the timeframes. Based on the above probability for existence of a track, a decision to end the track is then made when a certain probability threshold, provided as a parameter to the method, is reached.

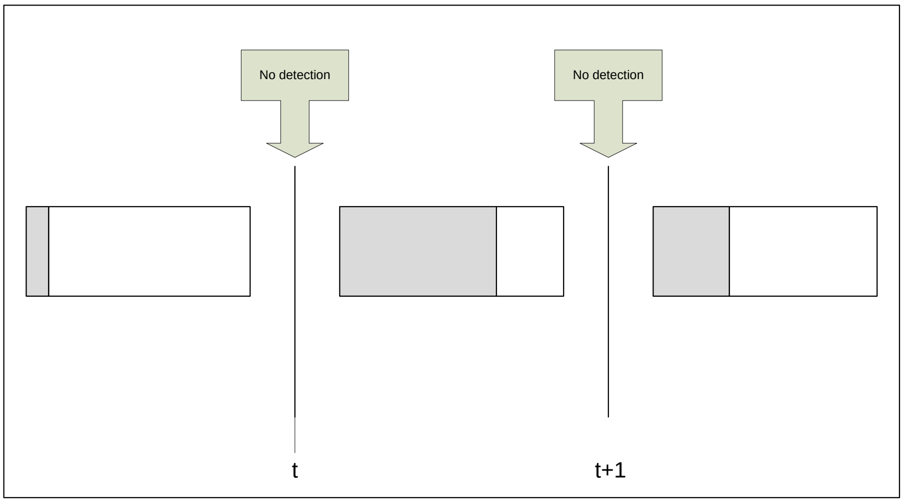


Figure 6-10: Changes to the probability for a presence of a track in two consecutive negative detections at times t and t+1.

In the above figure, the rectangular represents the belief about the track state, i.e. the grey area of the rectangular represents the probability for the track presence and the white area represents the probability for the track absence.

In conclusion, it is possible, through the application of the Bayes rule for subsequent timeframes, to incorporate the latest detector measurements and simultaneously update the belief about the continuation or ending of the human tracks. By the introduction of threshold probabilities for the birth and the death of a track, P_{birth} and P_{death}, a stochastic decision, when to end a track and start a new one, is made. If a decision for ending a track is reached, the time of death is selected to be the timeframe following the last detection. Similarly, when a decision to start a new track is made, the time of birth of the track is selected to be the time of the first detection from the sequence of detections that has led to the decision.

In summary, the proposed method uses three parameters, one characterising the human motion, i.e. the maximum walking velocity, to limit the range of possible detections for estimation of the possible data associations and two more parameters, containing the threshold probabilities for the ending and starting of a track. Also, the method is able to incorporate knowledge from several timeframes to infer when to start and end a track using Bayesian inference. Finally, the method, by expanding appropriately the search area, accounts for the individual's walking characteristics and reduces the risk of a wrong association with a clutter or a wrong detection, i.e. one belonging to a different track.

6.8 Numerical Approximation in the Tracking Algorithm

There are several groups of methods suitable for tracking a moving target. Each group is characterised by several specific advantages and disadvantages, making it suitable for different types of tracking scenarios. A short consideration of the main tracking methods is presented below regarding assessing their suitability for the context of multiple human tracking.

Filter suitability considerations

The Kalman filter (KF), due to its simplicity and low computational overheads, represents an efficient solution in many typical situations. However, there are restrictive limitations on its usage, i.e. the requirement that both transition and observation densities be Gaussian distributions with mean values expressed as linear transformations. As a result, the Kalman filter is not an optimal solution for a number of the real world tracking scenarios, as they do not meet the above criteria. Moreover, due to the need to handle the data associations between detection and multiple targets, which have to be embedded in the joint probability, the Kalman filter is also not suitable for application in multiple human tracking. Instead, in such cases, the extended Kalman filter (EKF) and the particle filters can be applied. However, despite the extended range of applications, these filters are still not universally applicable to every scenario. In particular, the EKF filter requires a local linearisation of the models, a process that sometimes could involve complicated mathematical techniques. In contrast, a particle filter can be applied directly for tracking of any non-linear system without the need of linearisation of the models. However, they have substantial computational overheads and are also not well suited for tracking in systems with high dimensionality.

In the context of multiple human tracking with variable data associations, it is considered that if a particle filter is used for exploring the data association space then the high number of possible combinations of tracks and associations will reduce the number of particles per combination significantly, potentially leaving some combinations without particles. Therefore, due to the above particle starvation problem it will be unreasonable to expect that the optimal candidate states will be explored. As a result, the algorithm will be very ineffective to infer the optimal state or, in the worst-case scenario, it may even be unable to reach convergence. One might expect that the above problem can be handled by substantially increasing the number of particles, which would guarantee a sufficient number of them per possible state. However, the resulting overhead will require massive computational resource or will slow the algorithm substantially because of the above high variance in the importance weights, poor particle diversity and poor tracking performance.

Several resampling methods have been developed, aimed at alleviating the particle starvation problem. Although a reasonable success rate has been reported by Doucet & Johansen (2011), the problem still remains when dealing with long tracking periods. The main benefit of the particle filters is their simple software implementation. Their main drawback is the exponential overhead growth with an increase of the dimensionality. Therefore, this results in the best practice recommending not using particle filters in problems with more than four dimensions. In contrast to the particle filters, the MCMC methods can cope much better with the problem of the high dimensionality at a lower computational load. However, they are not well suited for tracking dynamic states.

MCMC based numerical approximation

Due to the lack of a well suited filter to the needs of multiple human tracking, i.e. one that is able to handle efficiently both multiple state dimensions and a nonlinear system, a hybrid approach between a MCMC and a particle filter, i.e. MCMC-PF, proposed by Khan et al. (2005) is used. This approach takes advantage of the positive aspects of the two components of MCMC-PF, i.e. the multidimensional capabilities of the MCMC part and the dynamic state tracking capabilities of the particle filter part. This unique combination allows the above identified requirements for the multi-human tracking with a variable number of targets to be addressed simultaneously. It also enables use of variable detection associations as described below.

In particular, the key idea of the MCMC-PF based method for multiple human tracking is that the MCMC is used in the outer loop to explore the space of the number of tracks while the states of the individual tracks is estimated in the inner loop by the particle filter, using the fixed number of tracks, set by the outer loop. In particular, the state of the particle filter includes both mapping of the tracks to the measurements in the current timeframe and the dynamic coordinates of the people. Additionally, instead of sampling all states of tracks and data associations simultaneously, they are sampled sequentially in separate groups to accelerate the computation. In particular, the detection associations are sampled first and then the track states to reduce the computational overhead. Upon convergence, the particles with the most likely set of states are estimated using the previously discovered distribution of the most probable data associations and dynamic coordinates of tracked people over the whole duration of the tracking window.

As the MCMC-PF is a batch method, in contrast to the online nature of a pure particle filter, the measurements of the entire tracking period have to be available before the estimation of the state of the tracks in that period can start. This problem is mitigated in the method by the using a - sliding window, allowing the inclusion of the latest detections as they arrive. Processing a single window of measurement is sufficient for the purpose of track analysis for the computation of optimal navigation approach strategy.

MCMC move	Reverse Move	Description
Birth	Death	The total number of active tracks is increased by one. The new track is associated with the new detection.
Death	Birth	The total number of tracks is decreased by one. The associations of the track after time t (if any) are assigned to clutter.

Table 6-2: MCMC moves used by the method

A high-level overview of the algorithm is presented below. First, a track is randomly selected from all currently active tracks. Then, the algorithm proceeds to the MCMC part by sampling a move from all possible moves, given in Table 6-2 above. Next, it samples the detection associations and finally the state of the tracks. Within the step sampling of the states of the tracks, a Kalman filter is used to make a prediction of the track state based on the data association hypothesis made in the previous step, i.e. sampling the data associations. Finally, the algorithm calculates the acceptance ratio, using the proposed probabilities. Similarly to the standard MCMC, the acceptance ratio is utilised by the algorithm to decide whether to make the current MCMC move or skip it.

```
Input: Observations $Y_{1:t}$, time period $T$
Output: Estimated track states $X_{1:t}$ and data associations $\vartheta$
for $t$ = 1 to $T$ do
   for $iter = 2$ to $Num\_Iter$ do
      choose a track $i \in \{1,..,N_t\}$ randomly
      choose a $move\_type$ $\in \{1,..,4\}$ randomly
         choose proposal origin time frame $t^*$ depending on the $move\_type$
         copy a particle $p$ randomly from frame $t^*$ and track $i$
         create a new proposed particle ṗ from $p$
         propose new associations for ṗ and calculate proposal probability for them, $P(\theta_t)$
         propose new state for ṗ and calculate the proposal probability for it, $P(X_t, |X_{t-1})$
      calculate the posterior for ṗ
    calculate the acceptance ratio $\alpha$
   pick $rand$, uniformly distributed random number between 0 and 1,
   if $\alpha \geq rand$ then
      accept ṗ for the new particle
    else
        accept $p$ for the new particle
    end if
  end for
end for
```

Algorithm 6-1: MCMC-PF Multi Human Tracking

6.9 Human Group Segmentation

The next step in the process of generating a minimally intrusive navigation path for the service robot, after estimation of the individual tracks, is estimation of the group boundaries. The estimation is based on analysis of the tracks as described below. Subsequently, the group boundaries are used as a basis for computation of an optimal approach strategy of the service robot to navigate to an individual group member.

A group of people is a highly complex social formation. There is a range of interpersonal relations between the individual members. The proximetrics theory (Hall 1968), which has been developed using a substantial number of physiological experiments, states that the social relations among people are correlated to the physical distance between the individuals during interactions. This theory is the foundation of the group formation method proposed below. The distances between the group members are computed from the individual tracks, generated by the track position estimation algorithm as described in Section 6.8 above. Similarly to the single linkage clustering algorithm (Hartigan 1975), the Euclidean distance between the track positions in the timeframe, k, is compared with the clustering threshold, d_{cl_tr}, which is a parameter given in the algorithm. The result is a set of clusters $\Psi_k = \{\psi_i | i = 1, \dots, N_k\}$, illustrated in Figure 6-11, where N_k is the number of clusters in time k.

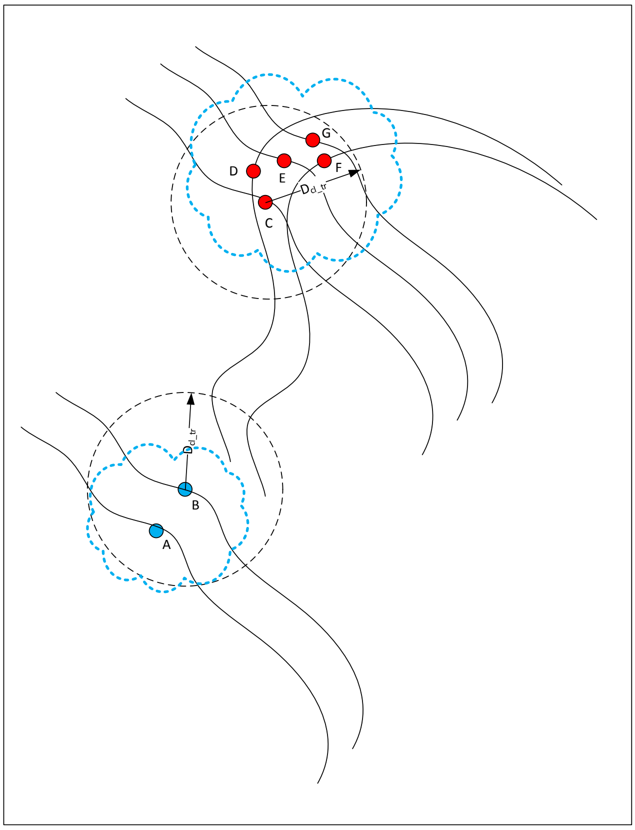


Figure 6-11: Proximetrics based clustering of states of the tracks

As demonstrated in Figure 6-11, it is possible that several track states can occur in close proximity without the respective people being part of the same group. It is possible that this is a random occurrence at a particular time. Therefore, the number of track state clusters does not directly correlate to the number of human groups. As a result, in order to simplify the operation of the algorithm it is necessary to filter out the tracks of people that do not belong to the group. This filtering is accomplished by analysis of the velocity components V_x, V_y of the track state vectors of the tracks within each cluster. In particular, the single linkage clustering algorithm, described above, is applied for a second time. This time, it is used to select a sub-set of the people, selected in the first phase of the method, by considering the velocity components of their estimated states. The distance threshold, d_{v_tr}, is used as the main parameter in the method, as illustrated in Figure 6-12 below.

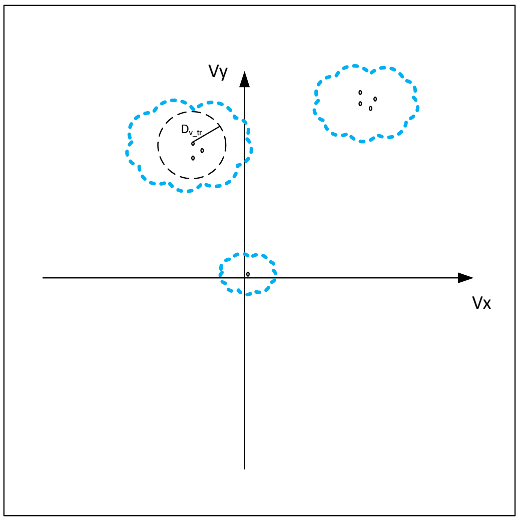


Figure 6-12: Clustering of the state of velocity components of tracks to form groups of people within each cluster of tracks

If the above method finds that there are several separate track groups within a single track cluster, the navigation planning of the robot is delayed until the individual human groups become clearly separated from one another. Finally, with human groups sufficiently separated, the computation of the navigation approach of the robot starts. For this purpose, the minimally intrusive navigation path is computed by the method, proposed below.

6.10 Minimally Intrusive Navigation in Crowded Spaces

Typically, people, who are members of a social group, interact with each other relatively frequently. This interaction represents a vital part of the group cohesion process (Dion 2001). Any interruption to the interaction, e.g. by a moving robot, is considered an unacceptable or undesirable intrusion. Also, such an action by a robot could be perceived by its users as violation of the established norms of acceptable social behaviour. Such a violation could potentially result in more negative attitudes towards the service robot and, consequently, hinder their successful adoption as helpers supporting the prolonged independent living of elderly people at home.

Occasionally, in practice, there might be an urgent need for an immediate and fast action by the robot, e.g. a delivery of a lifesaving medication to a person in urgent need. Such an emergency would necessitate the fastest possible navigation approach regardless of social norms. However, in the majority of the situations, such a high priority is not necessary. Therefore, the situational awareness in regard to people, allowing the decision-making module to differentiate between the need for urgent action and typical everyday operation, plays an important role in setting the optimisation priorities in navigation planning.

It is argued that a service robot that is aware of the situation and adheres to the socially accepted norms of behaviour, e.g. through human-aware navigation avoiding any unnecessary interruption to the human interactions, will be able to engage in more meaningful HRI. It will also have a considerably higher chance of acceptance as a personal helper in comparison with a robot that purely follows a strategy aimed at optimising its physical parameters of operation, e.g. energy use, time for delivery. Therefore, a method for minimally intrusive navigation path computation is proposed below. The method, residing on the third layer of the situational awareness paradigm, analyses the recent human tracks, generated by the algorithm proposed in 6.8, and the identified social groups, as described in 6.9, to gain a level of awareness of human interaction and generate a navigation path that avoids the spatial areas with high probability of social interaction.

The key idea of the proposed method is to build a probabilistic model of human interactions within the group, which represents the probability of new social interaction occurring between the group members. This model is then used to compute an appropriate navigation path that minimises the probability of interruption by the robot. In particular, the method utilises the proposed pairwise affinity measure to evaluate the probability of interaction between each pair of people within the group. Subsequently, from the affinity measures computed for all pairs within the group, a probabilistic spatial group interaction map is built. Finally, the above interaction map is used to compute the minimally intrusive approach path to the target person within the group.

6.10.1 Area of Active Interaction

When people interact with other members of their group they prefer to maintain a certain distance range depending on the type of interaction. This phenomenon has cultural, sociological and behavioural aspects that have been extensively studied in depth by the proxemics theory (Hall 1968). The spatial area, where such interactions occur, referred to further as the Area of Active Interactions (AAI), typically spans across the close and far ranges of the personal space as identified by the proxemics theory. In the proposed method, the AAI is used to identify the most probable interaction candidates among the people in the group. Subsequently, the behaviour of the interacting human pair is further analysed by computing the proposed affinity measure. This determines the extent of interaction based on analysis of historical track data. Moreover, the AAI around a person is also used by the robot as a target destination point, which guarantees an optimal navigation destination for natural, close interaction. Such humanistic behaviour mimics the established social human behaviour patterns and further facilitates the natural human-robot interaction.

The active interaction area of a person depends on their motion pattern. Typically, when people are stationary, their active interaction area is located mainly in front of them, as this allows face to face conversations to take place and maximum dexterity in using hands. However, when people start moving, the interaction area is considered to be shifting to both sides of the person. The change of shape and position of the AAI is also confirmed by the typical position and distance that people involved in the interaction have when they are stationary or walking side-by-side.

In similarity to the above optimal navigation destination in static situations, AAI can be used by the robot to engage in interaction with a moving person, e.g. delivering an object without obstructing the path of the person. Using the dynamic AAI as a target destination for an approaching robot, gives an option to the individual to choose whether to engage in interaction with the robot or to ignore the robot. In this way, the risk of violation of the socially accepted behaviour norms is minimised by utilising a passive interaction pattern. Indeed, blocking the path of a moving person without the person signalling willingness to engage in an interaction could be perceived by the person as an unfriendly, unintelligent or socially unacceptable action. Therefore, in situations when the conceptual information is not available or not understood sufficiently by the robot, it is considered desirable that passive interaction patterns are used when interacting with people.

The AAI in relation to the position of the person is modelled as the following probabilistic distribution function, specifying the probability of interaction for points in the surrounding space:

$$P_{AAI}(v,x,y) = A_C e^{-A_A\left(\frac{x^2}{A_B+|v|}+(y-\frac{K}{e^{|v|}})^2(A_B+|v|)\right)}, \tag{6.44}$$

where v is the speed of the person; x, y are coordinates of the evaluated point in the local coordinate system; A_A and A_B are parameters that determine how the shape of the interaction area changes with change in the speed v; K is a parameter determining the most likely distance for a face-to-face interaction; A_C is a normalisation parameter that guarantees that the overall interaction probability in the whole space equals to 1. For illustration, the position and the shape of AAI is given in Figure 6-13 for fixed parameters, i.e. $A_C = 1$, $A_A = 5$, $A_B = 0.03$, K = 3 and two different speeds $V = 0km/h$ (top) and $V = 7km/h$, as heat maps.

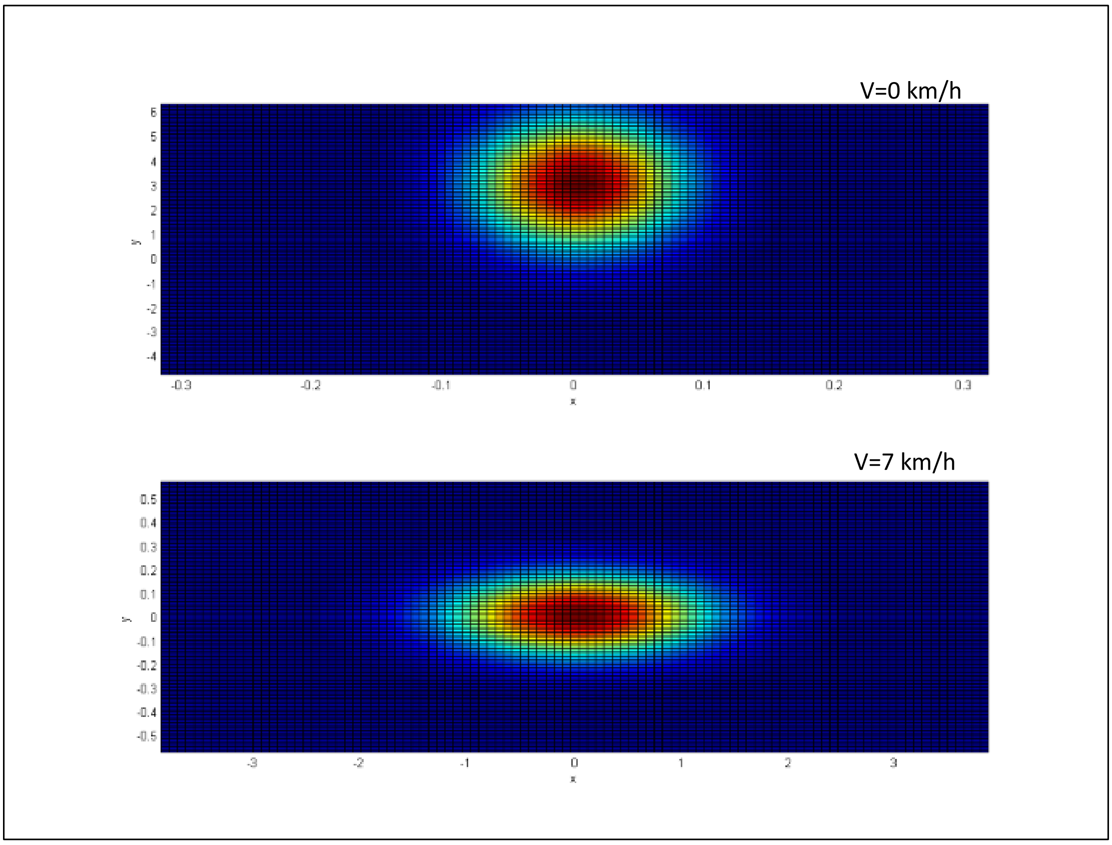


Figure 6-13: Active Area of Interaction at various speeds

In the figure, it can be observed that the AAI shifts from a position in front of a stationary person to a position to the side of a moving person.

6.10.2 Pairwise Affinity Measure

The Pairwise Affinity Measure (PAM) represents the probability of an interaction between two people. It is computed from the human position and orientation. The measure, inspired by the social force model (Helbing & Molnar 1995), is based on the Euclidean distance between two interacting people and the direction in which they are facing. In particular, the pairwise affinity measure is based on the AAI model (6.44). This provides dependence of PAM on the dynamically changing shape of the AAI and subsequently on the speed of the people. The PAM also depends on the directions in which people are facing, as an additional factor that determines the intensity of the interaction. In particular, when people are facing each other or walking next to each other, the interaction has maximum intensity. Otherwise, when the people are facing in opposite directions, PAM is minimal.

After the individual measure of interaction is computed separately for each person from the pair, the arithmetic mean for the pair is also calculated and used for generation of the group interaction map as explained below.

The Pairwise Affinity Measure (PAM) for the MN pair of people, shown in Figure 6-14, is given as:

$$Aff^{NM} = Aff^{MN} = \frac{Aff_M(N) + Aff_N(M)}{2}. \tag{6.45}$$

where $Aff_M(N)$ is the affinity measure of person M to person N and $Aff_N(M)$ is the affinity measure of person N to person M are given as:

$$Aff_M(N) = \\ = \begin{cases} A_C e^{-A_A\left(\frac{(d\sin(\varphi))^2}{A_B+|v_N|}+(d\cos(\varphi)-\frac{K}{e^{|v|}})^2(A_B+|v_N|)\right)}, & \varphi \leq \frac{\pi}{2} \\ 0 \qquad , & \varphi > \frac{\pi}{2} \end{cases} \tag{6.46}$$

Analogously:

$$Aff_N(M) = \\ = \begin{cases} A_C e^{-(A_A+K|v_N|)\left(\frac{(d\sin(\psi))^2}{A_B+|v_N|}+(d\cos(\psi))^2(A_B+|v_N|)\right)}, & \psi \leq \frac{\pi}{2}, \\ 0 \qquad , & \psi > \frac{\pi}{2} \end{cases} \tag{6.47}$$

where V_M, V_N are the speeds of person M and N respectively, φ and ψ are the angles between the M-N line and the directions that they are facing; d is the distance between person M and person N. The parameters A_A, K, A_B, A_C, inputs to the method, determine the shape and the relative position of the AAI, determining the pairwise function, in relation to each person.

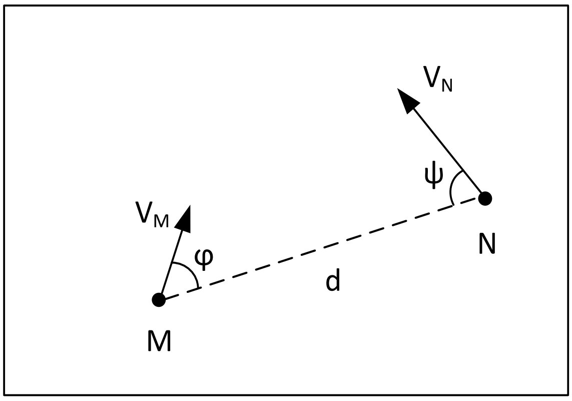


Figure 6-14: Pairwise affinity measure calculation

The resulting pairwise affinity measure, $Aff^{NM} \in [0..1]$, is a scalar that represents the current dynamic probability of interaction between the pair of people. However, for the purpose of human track analysis a measure evaluating a longer period is needed as proposed below.

The Historical Pairwise Affinity Measure (HPAM), provides a more stable measure of the pairwise affinity that filters out the transient fluctuations in PAM. The additional stability is achieved by using the exponential moving average function, also known as the exponentially weighted moving average (Croarkin & Tobias 2012). In particular, as the weighting for the factors for the previous values decrease exponentially with time the most recent values of PAM are given the highest importance as specified by the following recursive function:

$$AH_t^{NM} = \begin{cases} Aff_t^{NM} & ,for\ t = 1 \\ \alpha Aff_t^{NM} + (1 - \alpha) Aff_{t-1}^{NM} & ,for\ t > 1 \end{cases}, \tag{6.48}$$

where, $0 < \alpha < 1$, is a coefficient, provided as an input parameter, representing the degree of weighting decrease in HPAM for the pair N-M at time t.

The HPAMs, calculated in the above way for all pairs in the group from the estimated track states, are used in the following section to generate the map of interaction links between the human group members.

6.10.3 Group Interaction Connectedness

The interaction connection links, computed for the separate members of a group from the respective HPAMs, are used as a basis for analysis of group interaction connectedness. The links are stronger when the human pair interacts more frequently and weaker when no interaction has occurred over longer periods of time. The criteria is used to identify parts of the group that are not interacting as described below.

Initially, a graph of human interactions, which consists of vertices positioned at the current locations of the group members and edges representing the HPAMs between them is built. The edge position is computed by defining a line passing through the positions of person M, (x_M, y_M) and person N (x_N, y_N):

$$ax + by + c = 0\ , \tag{6.49}$$

where a, b, c are coefficients: $a = y_N - y_M$, $b = x_M - x_N$, $c = x_N y_M - x_M y_N$,

Then, a probabilistic distribution, referred to as a map of group interactions, is built from the separate interactions. In particular, the map is computed by defining a normal 2D distribution at each point of the edge of the graph, perpendicular to the edge. Then, for each point in the Euclidean space with coordinates (x, y), the shortest distance from the point to all the edges of the graph is computed using the following equation proposed by Weisstein (2010):

$$dist_{MN} = \frac{|ax + by + c|}{\sqrt{a^2 + b^2}}, \tag{6.50}$$

where a, b, c are the coefficients of the line MN from (6.49) and (x, y) are the coordinates of the point.

Finally, using the equation of a normal distribution and the above calculated distance $dist_{MN}$ as an offset from the mean, the spatial probability distribution of the interaction probability for the MN pair of people is computed as:

$$P_{MN}(x, y, \sigma) \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{dist_{MN}{}^2}{\sigma^2}} & , \quad x_M \leq x \leq x_N \; AND \; y_M \leq y \leq y_N \\ 0 & , \quad otherwise \end{cases} \tag{6.51}$$

An illustrative spatial probability distribution for two people, M$(-3.5, -6)$ and N(2.5,6) and $\sigma^2 = 1$, is shown in Figure 6-15 below.

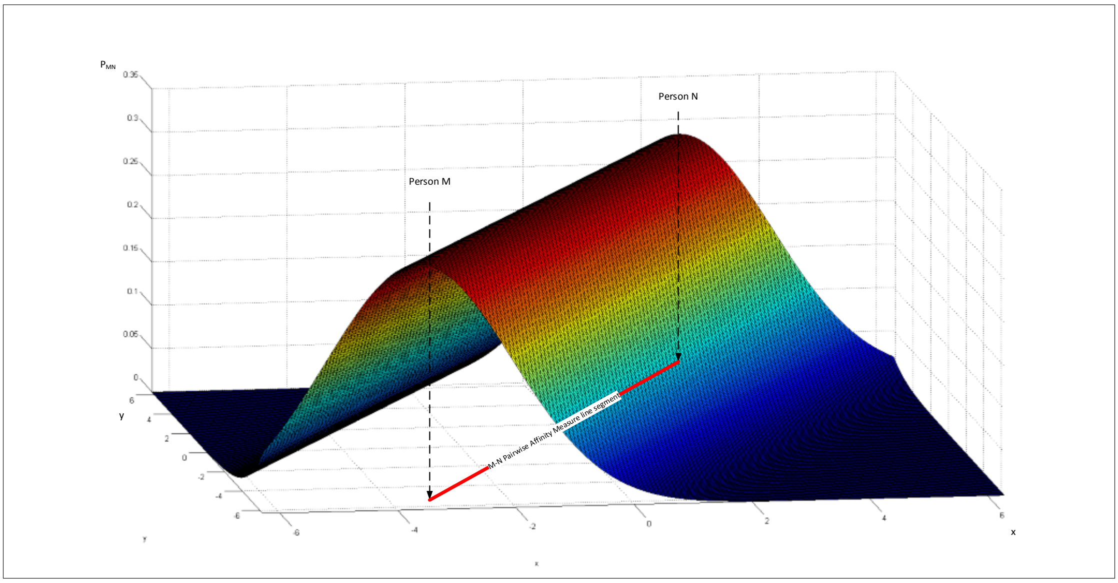


Figure 6-15: Pairwise Spatial Interaction Probability Distribution computed from the Pairwise Affinity Measure (PAM)

The purpose of the interaction map is to identify the areas with low probability of human interaction activities within the human group. These areas are then used for computation of an optimised navigation path to the targeted person. The optimisation in the path planning is intended to minimise the probability of disruption of the potential interaction. In particular, this is achieved by a minimisation of the accumulated Pairwise Spatial Probability Function score along the generated navigation path, e.g. by avoiding areas with high probability of interaction. The

Pairwise Spatial Probability Function, shown in Figure 6-15, is used as a heuristic function in the path-planning algorithm, described later.

In particular, the graph of interactions consists of vertices positioned at the current locations, i.e. estimated track states, of the group members and edges, representing the pairwise affinity measures between them. This graph forms the basis on which the map of group interaction connectedness is constructed. The process is accomplished by splitting the area into a grid of cells and calculating the probability of interaction for each cell. The calculation is based on the minimal distance of the cell's centre to the edges of the graph, which is then used as an offset in a normal distribution with its mean positioned on the edge of the graph.

After computation of the distance to all graph edges, the probability of interaction of each cell is given by the following mixture of Gaussians:

$$p(x) = \sum_{k=1}^{m} \alpha_k \; N\left(x \middle| \mu_k, \sigma_k^2\right), \tag{6.52}$$

where α_k is the mixture coefficient, which is selected to be equal to the pairwise affinity measure for the graph edge k, μ_k is the mean which lies on the nearest point from the edge k, σ_k^2 is the variance of the normal distribution associated with the edge k.

The value of the variance σ_k^2 depends on the composition of the interactions used in the computation of the HPAM. If $\mathrm{AH}_\mathrm{t}^\mathrm{NM}$ is a result of predominantly recent interactions then the certainty in its spatial position is relatively high. The high certainty results in a smaller variance σ_k^2. In contrast, if AH_t^{NM} is a result from older interactions, the certainty in its spatial position is lower, which results in an increased variance, σ_k^2, and a wider spreading of the distribution function.

In the next step, the variance of the normal distribution is set using the following procedure. Using the least squared distance minimisation, a straight line is fitted over the time series of the pairwise affinity measure for the pair, i.e. Aff^NM from (6.46), as illustrated in Figure 6-16. Then, the angle of the slope of that line, i.e. angle β in Figure 6-16, measured in radians, is used as value for the distribution of the variance for the corresponding edge in the graph.

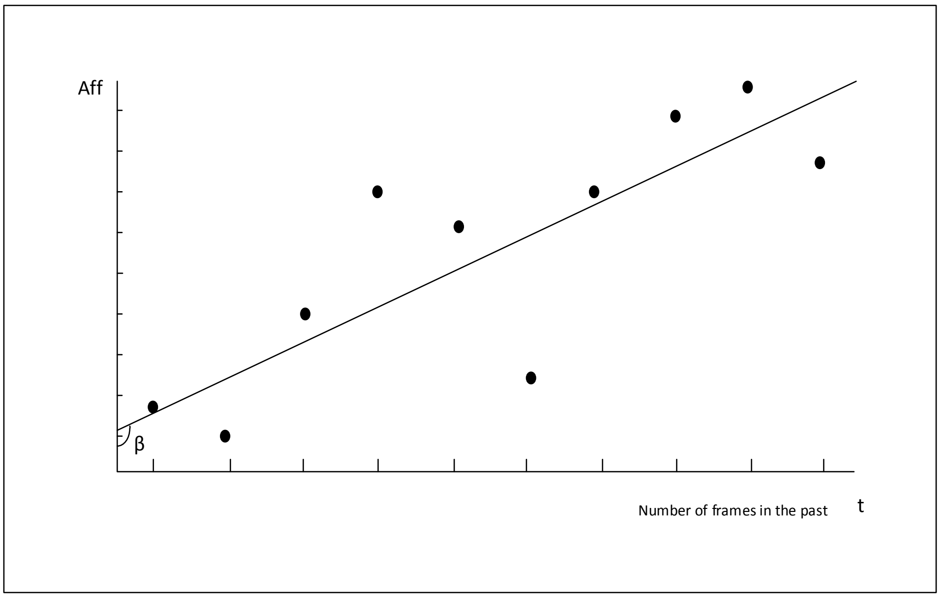


Figure 6-16: Estimation of the Composition of the Historical Variance used in the computation of the Map of Group Interaction Connectedness

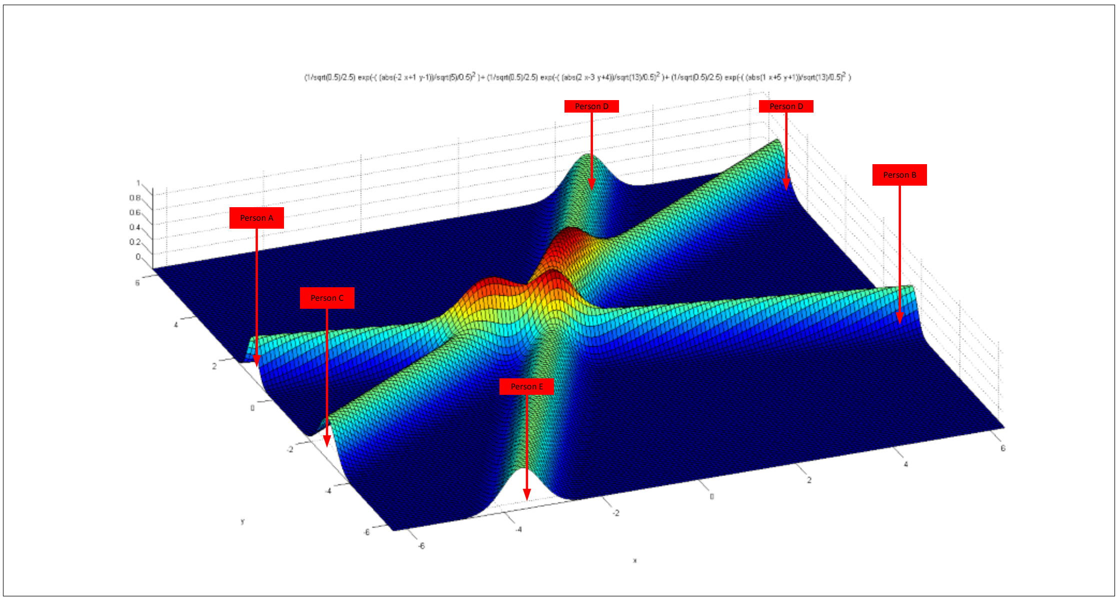


Figure 6-17: Typical Map of Group Interactions between 3 pairs of people

In the following step, the map of group interactions, illustrated in Figure 6-17, is generated using the proposed method. The map of interactions is then used by the robot's algorithms operating at the Human-aware planning phase of the situational awareness paradigm, Figure 1-3, to compute a human-aware navigation path as described below.

6.10.4 Minimally Intrusive Approach Path

Finding a minimally intrusive navigation path for the robot to reach the target person, layer three of the situational paradigm presented in Figure 1-3, is formulated as an optimisation task with a dual goal consisting of: a) avoidance of interruption to human interactions and b) minimisation of the energy or time to reach target person.

Since avoiding areas with high interaction potential requires longer routes, the above two sub-goals contradict with each other and, generally, it is unlikely that both can be minimised simultaneously.

The dual optimisation task is approached by creating a complex cost function that combines both sub-goals. The combination is achieved by assigning a cost per unit travelled distance and a cost for crossing an area with high interaction probability. The former cost is directly linked to the Euclidean distance of the navigation path, approximated by the accumulative distance between the cells from the map included by the navigation path. The latter cost is proportional to the interaction probability from the group interactions map. In particular, the combined cost function, C, is constructed as:

$$C = AC_{distance} + BC_{climb} \ , \tag{6.53}$$

where $C_{distance}$ is the unit cost for moving from one cell of the grid into the next, C_{climb} is the cost for a unit increase in the interactions probability, A and B are parameters that are provided to the algorithm as input. It should be noted that only an increase in the interaction potential, not decrease, is taken into account since once the human interaction is disrupted the harm cannot be undone by removal of the disruption. Therefore, C_{climb} is given as:

$$C_{climb} = \begin{cases} \Delta p & , \Delta p \geq 0 \\ 0 & , \Delta p < 0 \end{cases} , \tag{6.54}$$

where Δp is the change in the interaction potential from (6.52).

An additional restriction in finding the optimal path is the requirement that the robot should not use paths that run through or very close to obstacles and people. Therefore, a minimal distance between a robot and object or people is enforced at the final stage of the navigation path planning. The enforcement is achieved by increasing the cost function of the cells surrounding the obstacle through the introduction of a third cost function, p_p. The function p_p is modelled as a bivariate Gaussian function, with its mean centred on the location of the person and a variance proportional to the estimated speed of the person:

$$p_p(x) = N\left(x \middle| \mu_p, \sigma_p^2\right) , \tag{6.55}$$

where μ_p is the mean, i.e. location of the person and σ_p^2 is the variance of the Gaussian.

The dependence of the variance on the speed σ_p^2, is given as a linear function with coefficients C, D that models the requirement that a moving person needs a bigger safety zone:

$$\sigma_p^2(v) = Cv + D, \tag{6.56}$$

The classic A* planning algorithm (Hart et al. 1968) is then applied for solving the navigation planning problem given the above combined cost function. In the algorithm, the Euclidian distance to the goal position is used as the heuristic function. As required by the A* algorithm, beginning from the start position, each node is expanded in sequence according to the combined cost and heuristic values until the goal position is reached. Finally, the resulting search path, representing the least intrusive approach to the targeted person that takes into account the interactions within the social group, is smoothed and used for navigation of the robot.

6.11 Experiments

Two groups of experiments were carried out in order to evaluate the proposed algorithms. The first group targeted the method for multi-human track generation from noisy detections. The second group evaluated the generation of a minimally intrusive navigation path from the human tracks estimated in the first experiment.

For the purpose of the experiments, a synthetic detection dataset was generated automatically using typical values for the parameters of human motion, e.g. the maximum speed. The dataset contained a variable number of tracks, from one to fifteen. The tracks, referred to hereon as the correct tracks, were saved and then sampled randomly to generate the synthetic detection dataset, shown in Figure 6-18. Moreover, a large number of random false positives were added to the above dataset to simulate typical sensor noise, i.e. the red dots in Figure 6-19. Then, the resulting detection dataset was used by the multi-human tracking algorithm to evaluate its performance by comparing the results with the recorded correct tracks, illustrated in Figure 6-20. A standardised methodology for evaluation of the results was applied as explained below. It is considered that using synthetic data instead of real world detection data allows greater control over the variables in the experiments, e.g. sensor noise, specificity and selectivity parameters of the sensors.

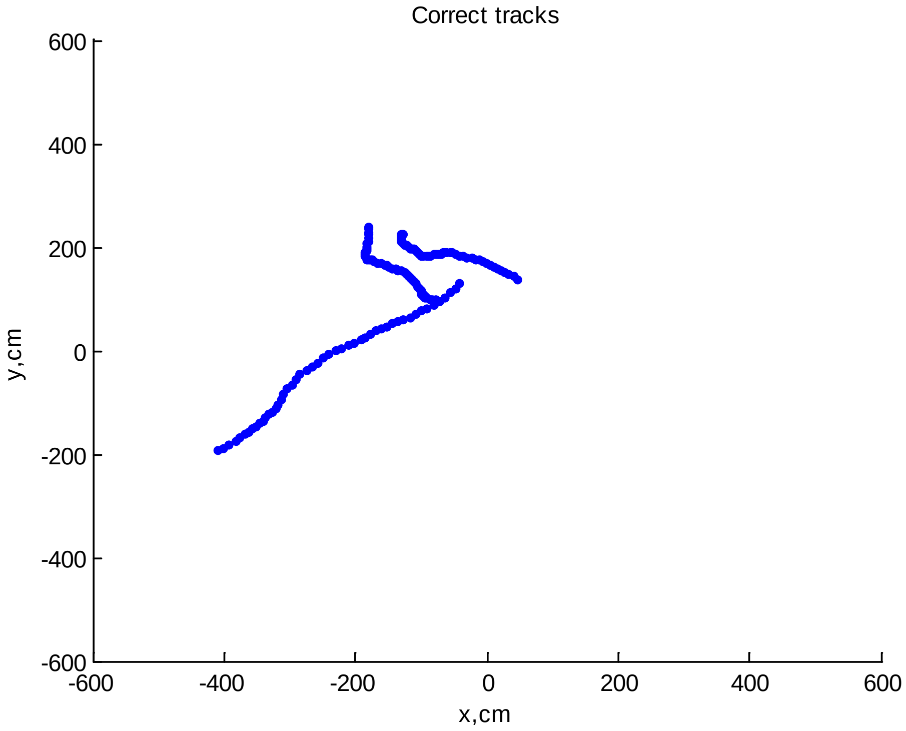


Figure 6-18: Correct tracks

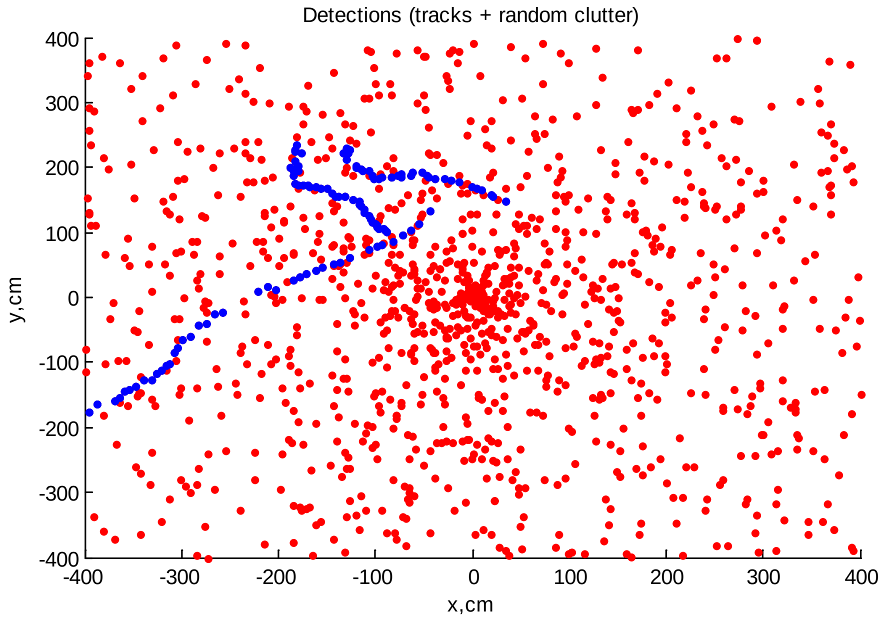


Figure 6-19: Detections used as input to the algorithm

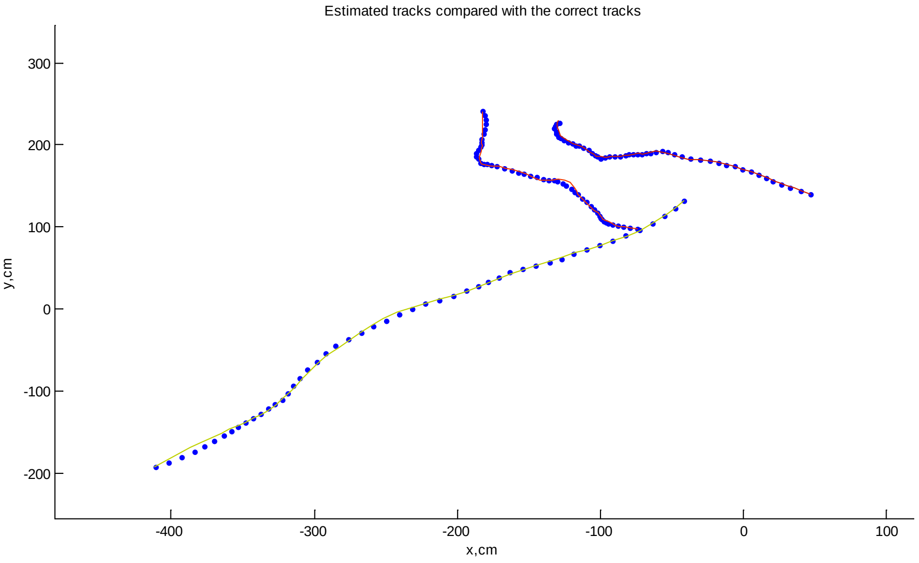


Figure 6-20: Estimated tracks in comparison with the correct tracks

A standardised metric, CLEAR Multiple Object Tracking metric (Bernardin & Stiefelhagen 2008), allowing comparison of the performance of the multi-human tracking algorithm against other tracking frameworks, was used to quantitatively evaluate the performance of the multi–human track algorithm. In particular, the metric consists of two components: Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Position (MOTP). The first component, MOTA, measures all object configuration errors made by the tracker over the time span of the tracking, e.g. false positives, false negatives, mismatches. MOTP measures the precision of the track state estimates. It operates by establishing a mapping at each time frame between the proposed-by-tracking framework target state and the ground truth. Based on the mapping, it calculates the position error and the number of false detections, detection errors and mismatches. Initially, the metric initialises with an empty mapping between the targets and the hypotheses at time $t = 1$. Then the algorithm is executed with the parameter d_{max} , which measures the maximum distance between a true and the estimated target. At the end of the observation period, the mean false detection rate, the detection failure rate, the mismatch rate and the position error rate are averaged as follows:

$$\overline{f_p} = \frac{\sum_t f_{p_t}}{\sum_t g_t} \tag{6.57}$$

$$\overline{f_n} = \frac{\sum_t f_{n_t}}{\sum_t g_t} \tag{6.58}$$

$$\overline{mm} = \frac{\sum_t mm_t}{\sum_t g_t} \tag{6.59}$$

$$\text{MOTA} = 1 - \overline{f_n} - \overline{f_p} - \overline{mm} \tag{6.60}$$

$$\text{MOTP} = \frac{\sum_t d_t}{\sum_t c_t} \tag{6.61}$$

where f_{p_t} is the number of the false positives, f_{n_t} is the number of false negative (misses), g_t is the number of people present in the frame, mm_t is the number of mismatches , d_t is the distance between the correct position of the person and the corresponding hypothesis, and c_t is the number of matches found for time t. As specified by the methodology, a false positive tracking result in a particular timeframe occurs when the tracker is in a state outside the circle defined by a centre which is the correct position of the track, taken from the synthetically generated track data, and a radius which is equal to a given threshold distance, $Dist_T$ as shown in Figure 6-21 below.

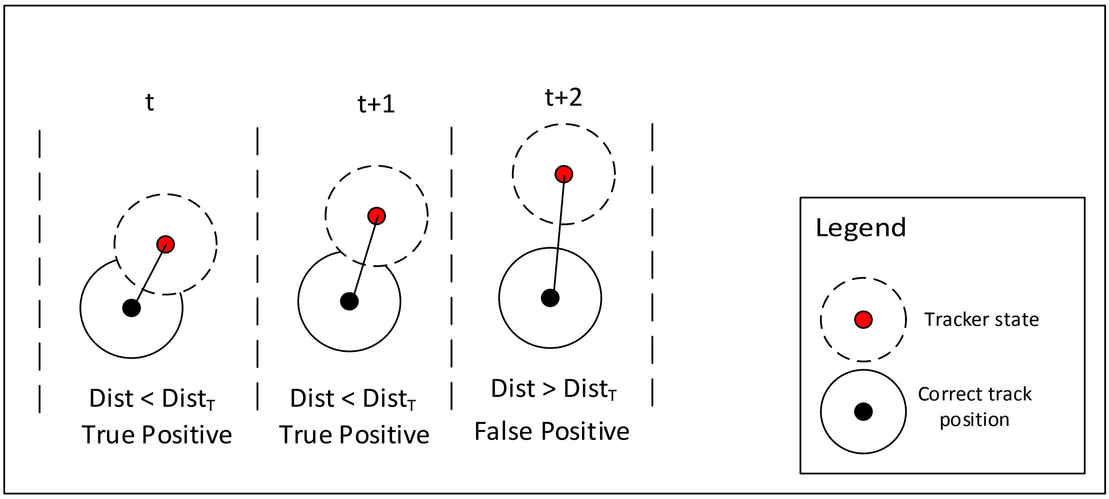


Figure 6-21: Determining true positive and false positive tracking results

In the experiments, the number of track mismatches was counted manually by comparing the estimated tracks with the correct tracks - a mismatch occurs when estimated tracks incorrectly swap places, e.g. due to wrong data associations.

As observed from (6.60) and (6.61), an ideal tracking algorithm would achieve a score of 1.0 for MOTA and 0.0 for MOTP. Also, it can be concluded that higher readings for MOTA and lower readings for MOTP represent better results.

Overall, five different experiments, with a various number of tracks, varying from 2 to 6 were conducted. The resulting estimated tracks in each of the experiments were compared with the correct tracks and the CLEAR Multiple Object Tracking metric parameters, presented in Table 6-3, were calculated.

	Experiment 3-tracks	Experiment 4-tracks	Experiment 5-tracks	Experiment 6-tracks	Experiment 7-tracks

$\sum_t f_{p_t}$	8	2	6	11	22
$\overline{f_p}$ $(Dist_T > 13\ cm)$	0.053	0.01	0.024	0.037	0.062
$\sum_t f_{n_t}$	0	0	0	0	24
$\overline{f_n}$	0	0	0	0	0.068
$\sum_t mm_t$	0	0	0	0	1
$\overline{mm}$	0	0	0	0	0.003
MOTA,%	94.7	99	97	96.3	86.7
MOTP, cm	6.44	5.14	9.17	8.77	10.17

Table 6-3: Results of the multiple human track experiments

In the table, the MOTA metric, the tracking accuracy, is expressed in % and the tracking precision is expressed in cm. The average MOTA for all experiments is 94.74% and the average MOTP is 7.93cm

From the table and the results presented in Appendix H, it can be observed that in most experiments the algorithm was able to identify correctly and robustly the number of tracks and the data association. However, with the increase in the number of tracks, especially when they occur in close proximity to each other, it was observed that the probability of a wrong data association increases. The ability of the algorithm to self-heal after more than one track is assigned to a detection is a positive feature that helps recovering in cases of an incorrect data association. An example of such a recovery can be observed in the experiment with seven tracks. In this experiment, a track was incorrectly associated with a nearby detection leading to a mismatch as shown in Figure 6-22. However, after discovering that two tracks were related to the same measurements, the algorithm was able to recover by finding the correct association and continued tracking the people. In cases when two tracks swap places, currently the algorithm is unable to recover due to lack of identity information. An extention of the algorithm that relies on identity information to initiate correction of swapped tracks will be investigated in future work.

Benchmarking

The CLEAR metric allows comparison with other tracking methods. A single person tracking method (Potamianos 2007) has reported results as MOTA 85.5% and MOTP 8.8 cm. Although the tracking is not executed on the same dataset and the proposed method tracks multiple people instead a single person, improvement in accuracy of the proposed method can be observed. In another work (Dehghan 2014), tracking multiple people in videos, the reported MOTA is 90% and MOTP is 6.9 cm., which are slightly worse than the result of the proposed method in this experiment.

In conclusion, as demonstrated, the proposed multiple human tracking method is able to achieve relatively good results. Moreover, in the unlikely event of a wrong data association, while a mismatch between tracks and detections can occur over a short period of time, the algorithm is likely to recover in most cases when two tracks are associated with the same measurement.

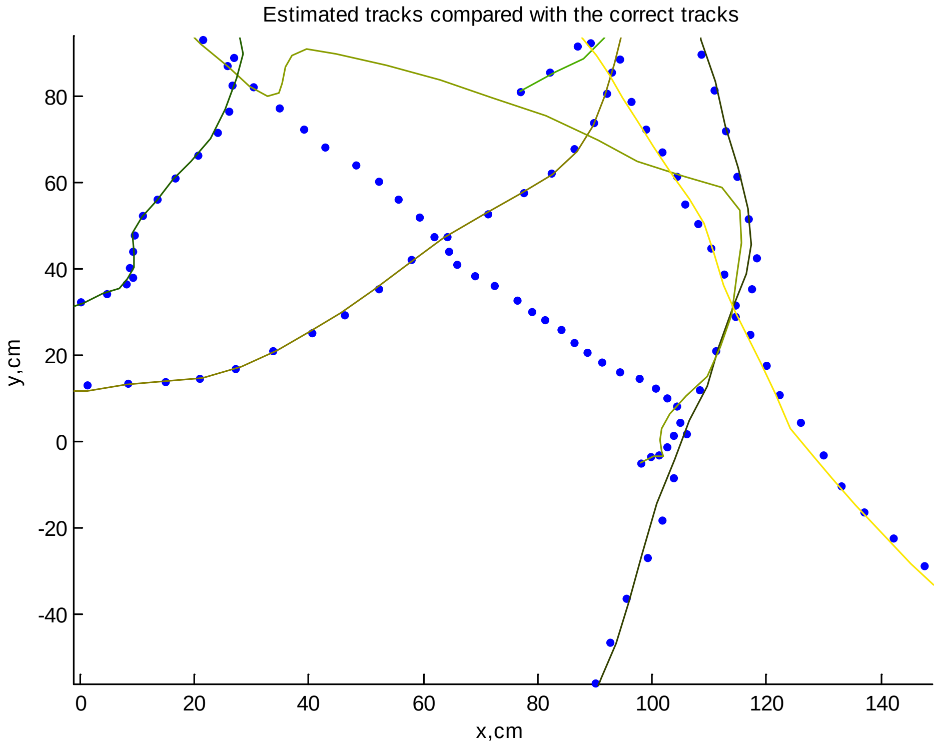


Figure 6-22: The worst problem detected in the experiments with the method for identification of multiple human tracks

User Study

The second group of experiments was carried out to verify and evaluate the acceptance by people of the generated minimally intrusive navigation paths. The evaluation of the results was carried out by a group of ten volunteers each ranking the acceptability of the generated navigation paths according to their personal opinion. In this group of experiments, the human tracks, generated by the first set of experiments, were entered into the Minimally Intrusive Navigation method, described in 6.10, to generate a navigation path for the robot. In particular, the navigation path was generated from a fixed starting position for the robot, i.e. the bottom left corner of the observable space, to a number of human targets, as shown in Figure 6-23. This setup was designed to simulate situations where a service robot is tasked to deliver an object to person involved in interactions with other people in the environment.

Further experiments, evaluating the behaviour of the method in different situations with a variable number of human tracks, i.e. 3 to 15 tracks, were carried out to evaluate the deterioration in performance in relation to the number of people, i.e. the decrease in accuracy and the increase in errors in crowded enviroments. The detailed results from these experiments are presented as Appendix H. As seen from the graphs, the navigation path, generated by the algorithm, reaches the target person while avoiding areas with high probabilty of interaction between people.

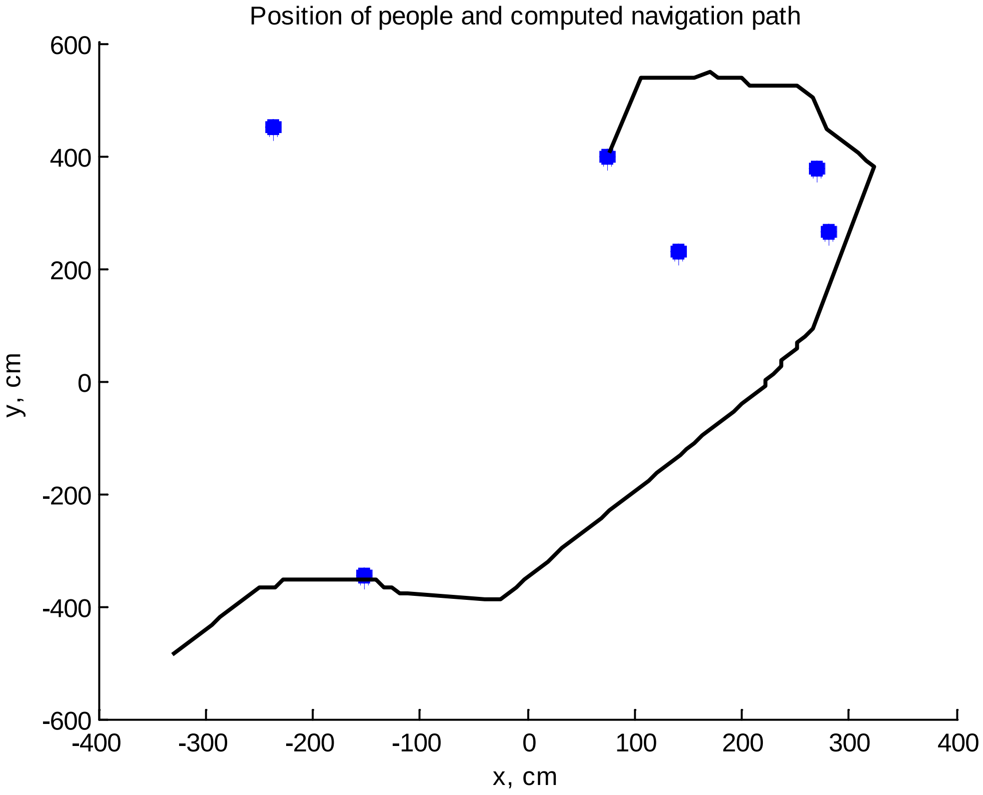


Figure 6-23: A typical navigation path computed by the algorithm

In the experiments, the function for keeping minimal distance between the navigation path and people was temporarily disabled to allow a direct observation of the effects of the minimally intrusive robot path planning method.

Evaluation of the generated navigation paths was carried out in a questionnaire type of exercise in which the generated paths were rated by a number of people. The participants were asked to use a scale from 1 to 10 to rate the acceptability of the generated navigation path in terms of compliance with the social norms. A score of one indicates the least intrusive navigation path and ten is the most intrusive navigation path. The responses are summarised in Table 6-4 below.

	Exp. 3-people	Exp. 4-people	Exp. 5-people	Exp. 6-people	Exp. 7-people	Exp. 15-people
evaluator 1	8	9	10	9	8	10
evaluator 2	10	9	8	10	10	8
evaluator 3	7	9	10	9	8	10
evaluator 4	9	9	8	8	10	10

evaluator 5	7	8	8	10	8	9
evaluator 6	7	8	9	9	10	10
evaluator 7	9	8	6	10	9	8
evaluator 8	8	9	10	9	9	10
evaluator 9	8	7	9	8	10	9
*evaluator*10	8	9	9	8	7	10
*evaluator*11	8	9	7	10	10	9
average score	8.09	8.54	8.54	9.09	9	9.4
st. deviation	0.94	0.68	1.29	0.83	1.09	0.84

Table 6-4: Results of the rating of the intrusiveness of the generated navigation path

The average score across all evaluators is 8.76 out of a possible 10, with a standard deviation of 1.02, which is considered a satisfactory result. It is also interesting to note that the average rating increases with the increase of the number of people in the scene. The above observation suggests that the algorithm performance does not saturate with an increase of complexity.

In conclusion, the result from the human evaluation confirms that human evaluators found the generated navigation paths to be acceptable, in terms of causing minimal disruption to the interaction in a human group.

6.12 Summary

In this chapter, a novel method for computing a minimally invasive navigation path by a service robot is proposed, developed and tested. The method, based on the proximetrics theory (Hall 1968), analyses the human tracks, estimated from a dataset of noisy detections, to compute the intensity of the human interactions between the people within a group. In the process of track estimation, the number of people in the scene is constantly estimated as well as the data association between detections and tracks. Additionally, knowledge about typical human walking patterns is embedded into the estimation model to allow the elimination of false positive detections that have occurred in unlikely positions. Subsequently, in the second part of the method, the generated human tracks are analysed to determine the number of human groups and their boundaries as well as the intensity of the interpersonal interactions within each of the groups. Finally, a map of human interaction probability, resulting from the track analysis, is used to generate minimally intrusive navigation paths for the robot. The method was evaluated in experiments with synthetically generated detection data. Then, the appropriateness of the generated navigation paths by the algorithm has been rated manually by the participants in the experiment. The evaluation of the generated paths has confirmed that the generated navigation paths are found to be a socially acceptable navigation path for the service robot in situations when human interactions occur.

It is a known that MCMC based methods are resource intensive. Therefore, it is no surprise that the track analysis part of the proposed method could not achieve real time execution in the experiments given the current hardware. However, as the method is considered suitable for parallelisation, an optimised GPU implementation, could be able to achieve a real time performance in future work.

Currently, the method takes into account past and current human detection information to evaluate the interaction between them. It is feasible, by using the estimated speed and the direction of movement of each individual, reported by the associated filter, to compute their future predicted positions. Then, the generated navigation path can be based on the predicted future state of the human interaction.

If two people are walking towards each other, clearly there is an increased probability that they are going interact. This probability is further increased by additional cues, e.g. if they are looking at each other, wave or give other signs that they are ready to engage in interaction. Therefore, as future work, prediction of future interaction based on detected human non-verbal cues is considered to be a potentially valuable research direction for future development of the proposed method. Moreover, as the robot speed is limited and in a dynamic scene the human interaction can evolve rapidly, a feature for additional dynamic partial re-planning of the planned socially aware navigation path is considered a potential direction for further development of the algorithm.

Contributions

Three contributions are made in this chapter:

- A probabilistic method for estimation of multiple human tracks from noisy measurements;
- A method for estimation of human groups and the map of human interaction;
- An optimisation method for generation of a minimally intrusive navigation path in situations where social interaction occurs.

7 Conclusions

7.1 Contributions

The main focus of this work has been defined as the investigation of reliable and accurate methods for human sensing that enable closer and more natural human-robot interactions in service robotics. Overall, the task has been approached through proposing a robot's situational awareness paradigm in regard to people, shown in Figure 1-3. The paradigm defines four layers, each with an increasing level of abstraction. The layer stack interconnects the low level sensing to high level control tasks that control and coordinate the robot's ability to achieve optimal HRI. The main emphasis of the work has been placed on the lower levels of the paradigm, i.e. the Perception level. However, in order to illustrate the viability of the connection to higher levels, through the Cognition and Judgement actions of the robot, some research investigation has been carried out on the Comprehension and Planning levels. In particular, the following problems, identified as hinderance factors in the optimal paradigm implementation have been addressed:

In Chapter 4, at Perception level, a set of methods for detection of human presence and location in the environment, using information from the typical on-board sensors of the robot, has been investigated. Additionally, a method for classifier fusion, augmenting the human detection information, has been proposed. The method uses both previous knowledge and dynamic information about the detectors' performance to improve the reliability of the human detectors. The method also addresses a variety of issues in real-life service robot scenarios such as occlusions and noisy measurements.

In Chapter 5, a method for improved human pose tracking from a low-cost 3D sensor, operating also at the Perception level of the paradigm, is proposed. The method combines the ICP local pose optimisation algorithm, commonly used in pose tracking, with a novel global pose configuration guidance method that uses the displacement between identified matching key point pairs from the point cloud. The result of the combination is an improvement in the characteristics of human pose tracking. In particular, the method is based on achieving alignment of the key point pairs, identified by matching the proposed local feature signatures in the point-cloud data with pre-recorded signatures from the human body surface. Then, the computed pose configuration is used to guide the pose configuration convergence. As a result, the robustness of the pose tracking is increased due to reducing the effect of the local minima. Additionally, as the global guidance to the local pose optimisation accelerates the initial stages of the convergence process, more time is left for the latter stages of the process. This enables more precise convergence to the correct human pose and a higher tracking precision to be achieved.

In Chapter 6, a method for human track analysis and human interaction aware navigation path generation is proposed. The first part of the method, operating at the Comprehension level of the paradigm, uses the generated human location information, i.e. the output of the combined classifier described in Chapter 4, to estimate the number of people in the scene and their most probable tracks. Assigning detections to tracks is considered to be a simple task only in situations when people are positioned sufficiently apart from each other. However, when people are in close proximity to one another, the problem becomes more challenging due to the increased ambiguity in assignments between tracks and measurement data. In such cases, the standard distance threshold separation methods fail and a single wrong detection-track association could lead to a tracking failure. Subsequently, the proposed method analyses the generated human tracks in order to cluster the people at the scene into groups and later estimate the active interactions between members of each group. Finally, the second part of the method, operating at the Planning layer of the paradigm, generates a navigation path that minimises the interruption to human interaction. The generated navigation path is then made available to the higher Decision making level of the paradigm for enactment.

Overall, this work investigates a number of improvements spanning a wide area of human perception capabilities. Several methods are proposed, addressing issues that are identified as hindering the optimal HRI. As demonstrated in experiments, the information, generated by the proposed methods, about the human presence, position and pose tracking, represents a foundation on which higher levels from the robot's control stack can operate reliably by taking into account the human information. The increased availability of the reliable human information at Perception level of the proposed paradigm is considered to be a very important factor for optimal operation of the higher level control algorithms. Such an improvement is envisioned to enable a more meaningful HRI and to contribute to increased acceptance of the service robots and thus help extend the period of independent living at home.

This thesis proposes a novel methodology for improving the human-awareness of HRI in service robotics. The following main contributions support the above methodology:

1. A novel paradigm defining the robot's situational awareness in regard to people;
2. A conceptual model of a framework for human perception of a service robotic system;
3. A novel method for classifier fusion in human detection and localisation. The method accumulates knowledge about the performance of the individual detectors that is taken into account when combining the output of several human detectors, according to the situation;
4. A novel human pose tracking method that combines both global and local human pose optimisation to improve the robustness of the human pose tracking;
5. A novel method for tracking of multiple people in noisy environments that associates probable human detections with human tracks to infer the number and the position of the most probable human tracks;
6. A novel method for minimally intrusive robot navigation path planning that minimises the interruption of human group interaction.

Alongside the above main contributions, a number of additional contributions have also been made as follows:

1. In Chapter 4, a method for compensation of the distortion in the images of moving targets in sequential scanning devices, e.g. a laser range finder;
2. In Chapter 4, a method for estimation of human leg speed from the distortion of the human laser range finder scans of the human legs;
3. In Chapter 4, A Random Forest based method for shape classification in laser range finder scan images;
4. In Chapter 4, A novel point cloud descriptor, i.e. HISP, which is generic in nature and can be used in a wide range of local signature -based applications;

Evaluation of the Contributions

The contributions made in this work represent an advance in enabling today's service robots to engage in closer interaction with people and provide more user-friendly service to their users. In particular, the contributions, listed above, are the result of the methods operating on the Perception, Comprehension and Human-Aware Planning layers of the situation awareness in regard to people paradigm, proposed in 1.8. The paradigm serves as a high-level underlying logical structure interlinking algorithms operating at different levels of the robot control, i.e. from the low level sensor data to the high decision making level. The assumption that more accurate and reliable information about the human presence, location and actions enables closer physical interactions between a service robot and a person has led to the investigation of a number of methods proposed for improving the human perception capabilities of the service robots. Additionally, through the proposed method for estimation of the human interaction by analysis of their tracks, operating at comprehension and planning level of the paradigm, the navigation planning of the robot is enhanced to take into account the interaction within the human group and compute as a result a less intrusive navigation path for the robot. The resulting interaction aware navigation is considered to represent progress in making the robot's behaviour more socially compliant and lead to improved acceptance as domestic helpers.

Although the proposed set of methods do not represent a complete methodology that is able to address all possible sensor modality combinations, they cover the most common sensors in service robotics today. Further sensor modalities can be added because of the modular nature of the framework. The key operational principle of all proposed methods is based on their ability to extract additional information cues about the people in the environment from already existing data. Such data reuse enables increased hardware cost optimisation as it keeps the number of the sensors low and saves energy. In addition to the proposed method in the lower layers of the paradigm, a method regarding the above integration of the extracted information cues into the higher-level decision making modules of the robot is also investigated. This method for human interaction aware path planning provides an important link between the low level human sensing and high-level control of the robot. The improved affordability and the operational characteristics of the robots is envisioned to be a contributing factor for a wider use of the robots for elderly care at home.

Human sensing is a very broad area with many applications outside robotics. Therefore, the contributions of this work are also applicable, with minor modifications, to a wider range of problems in industry and life. Typically, in an industrial safety context, where automated machinery is operated, when a human approaches the danger zone the machine has to be stopped. In this context, the control algorithms of the industrial machines can benefit directly from the proposed methods for improving the human-sensing capabilities of machines. An earlier human detection would enable a planned response, e.g. a warning or machine slowdown, resulting in overall improved safety. Similarly, human interaction analysis from human tracks has potential applications in numerous domains, including automated crowd monitoring for security purposes or customised services in retail industry.

The significance of each individual contribution is evaluated below:

The first main contribution, related to the proposed context awareness in regard to people paradigm, is considered to be of fundamental importance for improving the HRI as it provides a logical framework supporting the integration of sensor data into different levels of control. This enables a structured modular approach to the investigation of new algorithms for human sensing as well as easier integration with the already existing algorithms in the robot's control structure. Additionally, application of the above structured approach through the proposed paradigm enables new algorithms to be added to the framework in the future to enable new sensor modalities to be included, further increasing the robustness and precision of the information about the people.

The second contribution is about the proposed concept for implementation of the human perception framework. Definition of the main functional modules and relations enables further development and extension of the overall functional structure. This enables easy implementation of the paradigm in real world control frameworks for service robot control.

The third contribution makes improvements to the overall reliability and accuracy of the human detection and localisation based on dynamic combination of different data sources. High reliability of detection is considered an essential factor in achieving optimal results for the HRI control of the robot as occlusions and variations in human appearance result in low confidence of the human location that prevent any close physical interactions.

The fourth contribution, an improvement to the reliability of the human pose tracking, addresses a significant challenge in service robotics as it enables improved safety in robot arm manipulation based on more reliable information about the current human pose configuration. Ultimately, this contribution can result in better collision avoidance and improved HRI in applications which require a close physical contact between a robot and a person.

Finally, the fifth contribution enables computation of a minimally intrusive navigation path for the robot in crowded environments. The ability of the robot to take into account the interactions within a group of people is considered an essential factor for enabling socially compatible action that is envisioned to lead to improved acceptance of the robot by its users.

Overall, the contributions made in this work are considered to represent a significant factor for achieving safer and more meaningful human robot interaction as a result of the improved human perception capabilities of the service robots.

Mapping between contributions and objectives

Mapping the above contributions onto the objectives, described in 1.2, allows evaluation of how successfully the aims were achieved. In particular, the main contributions 1-3 and additional contributions 1-4 relate to the first objective as they improve the reliability of human presence and localisation by relying only on information from typical sensors of a service robot. In fact, by processing the information from the sensors in a new way and subsequently fusing the output of the detectors through dynamic weighting, a significant improvement in the overall reliability of human detection and localisation is achieved. The fourth main contribution can be mapped to the second objective as it is related to improving the reliability of the current methods for human pose tracking. By introducing global human pose guidance, the human pose tracking process is made significantly more robust to disturbance from external factors. Finally, the fifth and sixth contributions relate to the third objective as they provide means for analysis of the human tracks and computation of the resulting minimally intrusive navigation path. As demonstrated by the user survey the generated navigation path is considered by the users to be compliant with their understanding of the social norms of behaviour. Therefore, as all three objectives are met it is considered that this work achieves its aim by enabling personal service robots to engage in more natural and socially acceptable interaction with the people they serve.

7.2 Future Work

Due to the broad research area, covered in this work, there are numerous possibilities for further investigation and adaptation of the proposed methods. Some of the more prominent research directions include:

- Development of a stochastic model for human pose tracking that is able to utilise one or more additional point cloud features parallel to the proposed HISP descriptor. This represents an interesting future research direction that can investigate probabilistic arbitration between multiple conflicting key-point target pairs.

Development of a method for a stochastic prediction of human interactions within a social group. The method is expected to infer likely future human interaction from various cues about the people in the enviroment, e.g. human pose dynamics, gaze direction and environmental context. Such an investigation is considered to be very useful for enabling dynamic re-planning of the minimally intrusive navigation path in dynamic environments.

- Additionally, similarly to the human track analysis, described earlier, the human pose dynamics could be analysed to extract predictive cues allowing optimised arm navigation path planning of the robot. This will ensure safer and more socially compliant physical interaction, e.g. direct passing of an object from the robot to the human.

Future research, relying on the human information made available from proposed human perception methodology, can focus at the higher levels of the proposed paradigm. Similarly to the methods proposed in Chapter 6, it is considered that an investigation, aimed at improving the cognition and judgement capabilities of the robot, i.e. layers two and three of the paradigm, has significant potential to further improve the ability of the robot to engage in efficient HRI.

In conclusion, the research carried out in this work is considered to be a significant step forward in improving the human perception capabilities of today’s service robots. It also represents a foundation for further improvement of the cognition, judgement and decision making abilities of the service robots of tomorrow, which is envisioned to lead to more meaningful and user-friendly HRI.

8 References

Abdel-Malek, K. et al., 2004. Human performance measures:mathematics. In Proccedings of the ASME Design Engineering Technical Conferences (DAC 2004), Salt Lake Scty.

Agarwal, A. & Triggs, B., 2004. 3D human pose from silhouettes by relevance vector regression. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., 2.

Agarwal, A. & Triggs, B., 2006. Recovering 3D human pose from monocular images. IEEE transactions on pattern analysis and machine intelligence, 28(1), pp.44–58.

Aggarwal, J.K. & Cai, Q., 1999. Human Motion Analysis: A Review. Computer Vision and Image Understanding, 73, pp.428–440. Available at: http://www.sciencedirect.com/science/article/pii/S1077314298907445.

Aguiar, E. de et al., 2007. Marker-less Deformable Mesh Tracking for Human Shape and Motion Capture. 2007 IEEE Conference on Computer Vision and Pattern Recognition.

Althaus, P. et al., 2004. Navigation for human-robot interaction tasks. IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, 2.

Andriluka, M., Roth, S. & Schiele, B., 2008. People-tracking-by-detection and people-detection-by-tracking. 2008 IEEE Conference on Computer Vision and Pattern Recognition.

Anguelov, D. et al., 2005. Discriminative learning of Markov random fields for segmentation of 3D scan data. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2.

Ankur, A. & Bill, T., 2006. A local basis representation for estimating human pose from cluttered images. In Proceedings of the 7th Asian conference on Computer Vision ACCV. pp. 3851:50–59.

Ankur, A. & Bill, T., 2004. Learning to track 3D human motion from silhouettes. In Proceedings of the twenty-first international conference on Machine learning.

Arabo, A., Brown, I. & El-Moussa, F., 2012. Privacy in the Age of Mobility and Smart Devices in Smart Homes. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom. pp. 819–826. Available at:

Arras, K. & Cerqui, D., 2005. Do we want to share our lives and bodies with robots? A 2000-people survey. Autonomous Systems Lab (ASL), Swiss Federal …, (0605), pp.1–41. Available at:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.577&rep=rep1&type=pdf.

Arras, K.O. et al., 2008. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. 2008 IEEE International Conference on Robotics and Automation.

Arun, K.S., Huang, T.S. & Blostein, S.D., 1987. Least Squares Fitting of two 3D Point Sets. IEEE transactions on Pattern Analysis and Machine Intelligence, PAMI-9(5), pp.698–700.

Ayis, S. et al., 2007. Determinants of reduced walking speed in people with musculoskeletal pain. Journal of Rheumatology, 34, pp.1905 –1910.

Bailey, T. & Durrant-whyte, H., 2006. Simultaneous Localisation and Mapping (SLAM): Part II State of the Art. Computational Complexity, 13(3), pp.1–10. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1678144.

Balan, A.O. & Black, M.J., 2006. An Adaptive Appearance Model Approach for Model-based Articulated Object Tracking. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 1.

Bandouch, J., Engstler, F. & Beetz, M., 2008. Evaluation of Hierarchical Sampling Strategies in 3D Human Pose Estimation. In Proceedings of the 19th British Machine Vision Conference (BMVC).

Bekey, G.A., 2008. Springer Handbook of Robotics (B. Siciliano and O. Khatib; 2008) [Book Review]. IEEE Robotics & Automation Magazine, 15(3).

Bellotto, N. & Hu, H., 2009. Multisensor-based human detection and tracking for mobile service robots. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, 39(1), pp.167–181.

Ben-chen, M. & Gotsman, C., 2008. Characterizing Shape Using Conformal Factors. Proceedings of the 1st Eurographics conference on 3D Object Retrieval, pp.1–8.

Ben-Chen, M., Gotsman, C. & Bunin, G., 2008. Conformal Flattening by Curvature Prescription and Metric Scaling. Computer Graphics Forum, 27(2), pp.449–458. Available at: http://doi.wiley.com/10.1111/j.1467-8659.2008.01142.x.

Benjemaa, R. & Schmitt, F., 1998. A Solution for the Registration of Multiple 3D Point Sets Using Unit Quaternions. Lecture Notes In Computer Science, 1407, pp.34–50.

Bentley, J.L., 1975. Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9), pp.509–517.

Berclaz, J., Fleuret, F. & Fua, P., 2006. Robust People Tracking with Global Trajectory Optimization. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 1.

Bergman, N., 1999. Recursive Bayesian Estimation: Navigation and Tracking Applications. Linkoping university.

Berkmann, J. & Caelli, T., 1994. Computation of surface geometry and segmentation using covariance techniques, Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=334391.

Bernardin, K. & Stiefelhagen, R., 2008. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. EURASIP Journal on Image and Video Processing, 2008, pp.1–10. Available at: http://jivp.eurasipjournals.com/content/2008/1/246309 [Accessed October 5, 2012].

Besl, P.J. & McKay, H.D., 1992. A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(2).

Bezdek, J. et al., 1999. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing The Handbo., Springer.

Bhattacharyya, A., 1943. On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society, 35, pp.99–109. Available at: http://www.ams.org/mathscinet-getitem?mr=0010358.

Biau, G., 2012. Analysis of a Random Forests Model. Journal ofMachine Learning Research, 13, pp.1063–1095.

Blackman, S.S., 2004. Multiple hypothesis tracking for multiple target tracking. Aerospace and Electronic Systems Magazine, IEEE, 19(1), pp.5–18.

Bloch, I., 1996. Information combination operators for data fusion: A comparative review with classification. IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans., 26(1), pp.52–67.

Bloom, D.E. et al., 2011. Population aging:facts, challengies, and responses. Benefits and Compensation International, 41(1), pp.8–22.

Borenstein, E. & Ullman, S., 2002. Class-specific, top-down segmentation. In A. Heyden et al., eds. Computer Vision—ECCV 2002. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 109–122. Available at: http://link.springer.com/chapter/10.1007/3-540-47967-8_8.

Borovik, A. & Katz, M.G., 2011. Who Gave you the Cauchy-Weierstrass Tale? The Dual History of Rigorous Calculus. Foundations of Science, 17(3), p.46. Available at: http://arxiv.org/abs/1108.2885.

Bourdev, L. & Brandt, J., 2005. Robust object detection via soft cascade. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 236–243.

Boykov, Y. & Funka-Lea, G., 2006. Graph Cuts and Efficient N-D Image Segmentation. International Journal of Computer Vision, 70(2), pp.109–131.

Breiman, L., 2001. Random Forests R. E. Schapire, ed. Machine Learning, 45(1), pp.5–32. Available at: http://www.springerlink.com/index/U0P06167N6173512.pdf.

Brubaker, M.A. & Fleet, D.J., 2008. The Kneed Walker for human pose tracking. 2008 IEEE Conference on Computer Vision and Pattern Recognition.

Brubaker, S.C. et al., 2008. On the design of cascades of boosted ensembles for face detection. International Journal of Computer Vision, 77(1-3), pp.65–86.

BSI, 2012. BS ISO/IEC/IEEE 42010:201. ISBN-13: 978-0580705212

Burgard, W. et al., 1999. Experiences with an interactive museum tour-guide robot. Artificial Intelligence, 114(1-2), pp.3–55.

Buys, K. et al., 2013. An adaptable system for RGB-D based human body detection and pose estimation. Journal of Visual Communication and Image Representation, online.

Caillette, F., Galata, A. & Howard, T., 2008. Real-time 3-D human body tracking using learnt models of behaviour. Computer Vision and Image Understanding, 109(2), pp.112–125.

Caillette, F. & Howard, T., 2006. Real-time markerless 3-d human body tracking. Citeseer. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.4571&amp;rep=rep1&amp;type=pdf.

Carballo, A., Ohya, A. & Yuta, S., 2009. Multiple people detection from a mobile robot using double layered laser range finders. Proceedings of the IEEE ICRA 2009 Workshop on People Detection and Tracking, (May), pp.1–7.

Carton, D. et al., 2013. Proactively approaching pedestrians with an autonomous mobile robot in urban environments. In Experimental Robotics Springer Tracts in Advanced Robotics. pp. 199–214. Available at: http://link.springer.com/chapter/10.1007/978-3-319-00065-7_15.

Caruana, R., Karampatziakis, N. & Yessenalina, A., 2008. An empirical evaluation of supervised learning in high dimensions. Proceedings of the 25th International Conference on Machine Learning (2008), pp.96–103. Available at: http://portal.acm.org/citation.cfm?doid=1390156.1390169.

Catani, F. et al., 1996. Position and orientation in space of bones during movement: experimental artefacts. Clinical Biomechanics, 11(2), pp.90–100.

Cedras, C. & Shah, M., 1995. Motion-based Recognition a Survey. Journal on Image and Vision Computing, 13(2), pp.122–155.

Chaffin, D.B., Anderson, B.J. & Martin, B.J., 2006. Occupational Biomechanics, Wiley, New York,NY.

Chai, Y. et al., 2011. Multi target tracking using multiple independent particle filters for video surveillance, IEEE.

Cho, H. et al., 2012. Real-time pedestrian detection with deformable part models. In IEEE Intelligent Vehicles Symposium, Proceedings. pp. 1035–1042.

Choquet, G., 1954. Theory of capacities. Annales de l'institut Fourier, 5, pp.131–295.

Chung, W. et al., 2009. Safe navigation of a mobile robot considering visibility of environment. IEEE Transactions on Industrial Electronics, 56(10), pp.3941–3950.

Cleveland, W.S. & Loader, C.R., 2001. Smoothing by Local Regression: Principles and Methods. ANNMEDPSYCHOL, Annales-Me(10), pp.10–49. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.4020.

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

Cover, T. & Hart, P., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), pp.21–27. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1053964.

Cox, I.J., 1993. A review of statistical data association for motion correspondence. Int. J. Comput. Vision, 10(1), pp.53–66. Available at: http://portal.acm.org/citation.cfm?id=173513.

Croarkin, C. (NIST) & Tobias, P., 2012. NIST/SEMATECH e-Handbook of Statistical Methods, NIST. Available at: http://www.itl.nist.gov/div898/handbook/.

Cui, J.C.J. et al., 2006. Laser-based Interacting People Tracking Using Multi-level Observations. 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems.

Dalal, N., Triggs, B. & Schmid, C., 2006. Human detection using oriented histograms of flow and appearance. In ECCV'06 Proceedings of the 9th European conference on Computer Vision - Volume Part II. pp. 428–441.

Dalal, N. & Triggs, W., 2004. Histograms of Oriented Gradients for Human Detection C. Schmid, S. Soatto, & C. Tomasi, eds. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR05, 1(3), pp.886–893. Available at: http://eprints.pascal-network.org/archive/00000802/.

Dautenhahn, K. et al., 2005. What is a robot companion - Friend, assistant or butler? In 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS. pp. 1488–1493.

Dehghan, A., Idrees, H., Zamir, A. R., & Shah, M. (2014). Automatic Detection and Tracking of Pedestrians in Videos with Various Crowd Densities. In Pedestrian and Evacuation Dynamics 2012 (pp. 3-19). Springer International Publishing.

Demirdjian, D., Ko, T. & Darrell, T., 2003. Constraining human body tracking. Proceedings Ninth IEEE International Conference on Computer Vision.

Dempster, A.P., 2008. Upper and lower probabilities induced by a multivalued mapping. Studies in Fuzziness and Soft Computing, 219, pp.57–72.

Deutscher, J. & Reid, I., 2005. Articulated Body Motion Capture by Stochastic Search. International Journal of Computer Vision, 61(2), pp.185–205. Available at: http://link.springer.com/10.1023/B:VISI.0000043757.18370.9c\nhttp://www.springerlink.com/openurl.asp?id=doi:10.1023/B:VISI.0000043757.18370.9c.

Dewaele, G. et al., 2006. The alignment between 3-d data and articulated shapes with bending surfaces. Computer Vision–ECCV 2006, pp.578–591. Available at: http://www.springerlink.com/index/a767882591189358.pdf.

Dion, K.L., 2001. Group cohesion: From “field of forces” to multidimensional construct: Erratum. Group Dynamics: Theory, Research, and Practice, 5(1), pp.2–2.

Dollar, P. et al., 2009. Pedestrian detection: A benchmark. 2009 IEEE Conference on Computer Vision and Pattern Recognition.

Dollar, P. et al., 2012. Pedestrian Detection: An Evaluation of the State of the Art. In IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 743–761.

Dollár, P. et al., 2012. Pedestrian detection: an evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence, 34(4), pp.743–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21808091.

Doucet, A., Freitas, N. de, et al., 2000. Rao-Blackwellised particle filtering for dynamic Bayesian networks. Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence, pp.176–183.

Doucet, A., Godsill, S. & Andrieu, C., 2000. On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. Statistics and Computing, 10(3), pp.197–208.

Doucet, A. & Johansen, A., 2011. A tutorial on particle filtering and smoothing: fifteen years later. In Handbook of Nonlinear Filtering. pp. 656–704. Available at: http://automatica.dei.unipd.it/tl_files/utenti/lucaschenato/Classes/PSC10_11/Tutorial_PF_doucet_johansen.pdf.

Douillard, B. et al., 2011. On the segmentation of 3D LIDAR point clouds. 2011 IEEE International Conference on Robotics and Automation, pp.2798–2805.

Ducourant, T. et al., 2005. Timing and distance characteristics of interpersonal coordination during locomotion. Neuroscience Letters, 389(1), pp.6–11.

Durrant-whyte, H. & Bailey, T., 2006. Simultaneous Localisation and Mapping (SLAM): Part I The Essential Algorithms. History, 13(2), pp.1–9. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.4195&amp;rep=rep1&amp;type=pdf.

Elbanhawi, M. & Simic, M., 2014. Sampling-Based Robot Motion Planning: A Review. IEEE Access, 2, pp.56–77. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6722915.

Elgammal, A. & Lee, C.S., 2007. Nonlinear manifold learning for dynamic shape and dynamic appearance. Computer Vision and Image Understanding, 106(1), pp.31–46.

Elgammal, A. & Lee, C.-S.L.C.-S., 2004. Inferring 3D body pose from silhouettes using activity manifold learning. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., 2.

Enzweiler, M. & Gavrila, D.M., 2009. Monocular pedestrian detection: survey and experiments. IEEE transactions on pattern analysis and machine intelligence, 31(12), pp.2179–2195.

Van Erp, M. & Schomaker, L., 2000. Variants Of The Borda Count Method For Combining Ranked Classifier Hypotheses. In Z. Bo, D. Dayong, & Z. Ling, eds. Seventh international workshop on frontiers in handwriting recognition. pp. 443–452.

Ess, A. et al., 2008. A mobile vision system for robust multi-person tracking. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4587581.

Ess, A. et al., 2009. Robust multiperson tracking from a mobile platform. IEEE transactions on pattern analysis and machine intelligence, 31(10), pp.1831–1846.

Ess, A., Leibe, B. & Gool, L. Van, 2007. Depth and Appearance for Mobile Scene Analysis. 2007 IEEE 11th International Conference on Computer Vision.

Farin, G., 1997. Curves and Surfaces for Computer Aided Geometric Design, A Practical Guide, 4th Editio., Academic Press, San Diego.

Faux, I.D. & Pratt, M.J., 1979. Computational Geometry for Design and Manufacture, Ellis Horwood Publishing, Chichester, UK.

Feil-Seifer, D. & Mataric, M., 2011. Automated detection and classification of positive vs. negative robot interactions with children with autism using distance-based features. In Proceedings of the 6th international conference on Human-robot interaction - HRI ’11. ACM Press, p. 323. Available at: http://dl.acm.org/citation.cfm?id=1957656.1957785.

Feil-Seifer, D. & Mataric, M.J., 2009. Human-robot interaction. In R. A. Meyers, ed. Invited contribution to Encyclopedia of Complexity and Systems Science. Springer New York, pp. 4643–4659. Available at: http://robotics.usc.edu/publications/585/.

Felzenszwalb, P., McAllester, D. & Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model. 2008 IEEE Conference on Computer Vision and Pattern Recognition.

Felzenszwalb, P.F. & Huttenlocher, D.P., 2004. Efficient Graph-Based Image Segmentation. International Journal of Computer Vision, 59(2), pp.167–181. Available at: http://link.springer.com/10.1023/B:VISI.0000022288.19776.77.

Fischler, M.A. & Bolles, R.C., 1981. Random sample consensus. Communications of the ACM, 24(6), pp.381–395. Available at: http://dl.acm.org/citation.cfm?id=358669.358692.

Fod, A., Howard, A. & Matari, M.J., 2002. Laser-Based People Tracking. Robotics, 3(May), pp.3024–3029. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.1183&amp;rep=rep1&amp;type=pdf.

Fong, T., Nourbakhsh, I. & Dautenhahn, K., 2003. A survey of socially interactive robots. In Robotics and Autonomous Systems. pp. 143–166.

Forsyth, D.A. et al., 2005. Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis. Foundations and Trends® in Computer Graphics and Vision, 1(2/3), pp.77–254.

Fortmann, T., Bar-Shalom, Y. & Scheffe, M., 1983. Sonar tracking of multiple targets using joint probabilistic data association. IEEE Journal of Oceanic Engineering, 8(3).

Freeman, B. et al., 2013. Depth mapping using projected patterns. Available at: https://www.google.com/patents/US8493496.

Freund, Y. & Schapire, R.E., 1997. A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. Journal of Computing Systems and Science, 55(1), pp.119–139.

Gamini Dissanayake, M.W.M. et al., 2001. A solution to the simultaneous localization and map building (SLAM) problem. IEEE Transactions on Robotics and Automation, 17(3), pp.229–241.

Gavrila, D.M. & Munder, S., 2006. Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle. International Journal of Computer Vision, 73(1), pp.41–59. Available at: http://www.springerlink.com/index/10.1007/s11263-006-9038-7.

Gavrila, D.M. & Philomin, V., 1999. Real-time object detection for “smart” vehicles. In Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE, pp. 87–93 vol.1. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=791202.

Gelb, A., 1996. Applied Optimal Estimation, MIT Press, MA.

Gerasimos P. & Zhenqiu Z. 2007 A joint system for single-person 2D-face and 3D-head tracking in CHIL Seminars, Multimodal Technologies for Perception of Humans, Lecture Notes in Computer Science, pp 105-118, Volume 4122, Springer, ISBN: 978-3-540-69567-7

Girshick, R. et al., 2011. Efficient regression of general-activity human poses from depth images. 2011 International Conference on Computer Vision, pp.415–422.

Gokberk, B., Salah, A.A. & Akarun, L., 2005. Rank-based decision fusion for 3D shape-based face recognition. Proceedings of the IEEE 13th Signal Processing and Communications Applications Conference, 2005.

Golovinskiy, A. & Funkhouser, T., 2009. Min-cut based segmentation of point clouds. In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops. IEEE, pp. 39–46.

Golub, G. & Kahan, W., 1965. Calculating the Singular Values and Pseudo-Inverse of a Matrix. Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis, 2(2), pp.205–224. Available at: http://www.jstor.org/stable/2949777.

Gómez, J. & Garrido, S., 2013. Social Path Planning: Generic Human-Robot Interaction Framework for Robotic Navigation Tasks, Cognitive Robotics Systems: Replicating Human Actions and Activities. In Workshop of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'13). Tokio.

Goodrich, M.A. & Schultz, A.C., 2007. Human-Robot Interaction: A Survey. Foundations and Trends® in Human-Computer Interaction, 1(3), pp.203–275.

Grabisch, M., 1996. The application of fuzzy integrals in multicriteria decision making. European Journal of Operational Research, 89(3), pp.445–456.

Grest, D., Woetzel, J. & Koch, R., 2005. Nonlinear Body Pose Estimation from Depth Images. In Pattern Recognition. pp. 285–292. Available at: http://www.springerlink.com/content/319v00jj6t4yj8pv.

Guzzi, J. et al., 2013. Human-friendly robot navigation in dynamic environments. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Karlsruhe, Germany.

H. Sidenbladh, J. Blanck, L.S., 2002. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. Computer Vision - ECCV 2002, 2350, pp.784–800.

Hahnel, D. et al., 2003. Map building with mobile robots in dynamic environments. Robotics and Automation, 2003. Proceedings. ICRA &apos;03. IEEE International Conference on, 2, pp.1557–1563. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1241816.

Hahnel, D., Thrun, S. & Burgard, W., 2003. An extension of the ICP algorithm for modeling nonrigid objects with mobile robots. In International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc., pp. 915–920. Available at: http://portal.acm.org/citation.cfm?id=1630791.

Hall, D.L.D.L., Member, S. & Llinas, J., 1997. An introduction to multisensor data fusion. Proceedings of the IEEE, 85(1), pp.6–23. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=554205.

Hall, E.T., 1968. Proxemics. Current Anthropology, 9(2/3), p.83.

Harper, S. & Hamblin, K., 2014. International Handbook on Ageing and Public Policy, Edward Elgar Publishing Ltd.

Hart, P.E., Nilsson, N.J. & Raphael, B., 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Transactions on Systems Science and Cybernetics, 4(2).

Hazewinkel, M., 2001. Lagrange interpolation formula, Encyclopaedia of Mathematics, Springer. Available at: http://www.encyclopediaofmath.org/index.php/Lagrange_interpolation_formula.

Helbing, D. & Molnar, P., 1995. Social force model for pedestrian dynamics. Physical Review E, 51(5), pp.4282–4286.

Henry, P. et al., 2010. Learning to navigate through crowded environments. In Proceedings - IEEE International Conference on Robotics and Automation. pp. 981–986.

Hildebrandt, K., Polthier, K. & Wardetzky, M., 2007. On the convergence of metric and geometric properties of polyhedral surfaces. Geometriae Dedicata, 123(1), pp.89–112.

Himmelsbach, M., 2010. Fast segmentation of 3D point clouds for ground vehicles. Intelligent Vehicles …, pp.560–565. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5548059.

Hofmann, M. & Gavrila, D.M., 2011. Multi-view 3D Human Pose Estimation in Complex Environment. International Journal of Computer Vision, 96(1), pp.103–124. Available at: http://link.springer.com/10.1007/s11263-011-0451-1.

Hoiem, D., Efros, A.A. & Hebert, M., 2006. Putting Objects in Perspective. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2.

Hou, C., Ai, H. & Lao, S., 2007. Multiview Pedestrian Detection Based on Vector Boosting. Asian Conference on Computer Vision, 4843, pp.210–219. Available at: http://www.springerlink.com/index/e312862wn76x5q06.pdf.

Howe, N.R., 2007. Silhouette lookup for monocular 3D pose tracking. Image and Vision Computing, 25(3), pp.331–341.

I T, J., 2002. "Principal Component Analysis,2nd ed," Springer Series in Statistics. Available at: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-95442-4.

Ioffe, S. & Forsyth, D., 2001. Human tracking with mixtures of trees. Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, 1.

IPA (Fraunhofer), 2014. Care-O-bot. Available at: http://www.care-o-bot.de/en/care-o-bot-3.html [Accessed September 1, 2014].

Isard, M. & Blake, A., 1998. Condensation — Conditional density propagation for visual tracking. International Journal of Computer Vision, 29(1), pp.5–28. Available at: http://www.springerlink.com/index/xl887466h454318k.pdf.

Jacobs, R.A. et al., 1991. Adaptive Mixtures of Local Experts. Neural Computation, 3(1), pp.79–87.

Jain, A. & Ross, A., 2002. Fingerprint mosaicking. 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 4, pp.IV–4064–IV–4067.

Jain, V. & Learned-Miller, E., 2010. Fddb: A benchmark for face detection in unconstrained settings. UMass Amherst Technical Report. Available at: http://works.bepress.com/erik_learned_miller/55/.

Jazwinski, A., 1970. Stchastic Process and Filtering Theory. Mathematics in Science and Engineering, Academic Press, 64.

Jensen, B. et al., 2005. Robots meet Humans-interaction in public spaces. IEEE Transactions on Industrial Electronics, 52(6).

Jordan, M.I. & Jacobs, R.A., 1993. Hierarchical mixtures of experts and the EM algorithm, MIT Press. Available at: http://www.mitpressjournals.org/doi/abs/10.1162/neco.1994.6.2.181.

Julier, S.J. & Uhlmann, J.K., 2004. Unscented filtering and nonlinear estimation. Proceedings of the IEEE, 92(3).

Kahn, P.H. et al., 2008. Design patterns for sociality in human-robot interaction. Proceedings of the 3rd international conference on Human robot interaction HRI 08, p.97. Available at: http://portal.acm.org/citation.cfm?doid=1349822.1349836.

Kalman, R.E., 1960. A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME-Journal of Basic Engineering, 82(Series D), pp.35–45. Available at: http://fluidsengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1430402

Kanda, T. et al., 2009. Abstracting peoples trajectories for social robots to proactively approach customers. IEEE Transactions on Robotics, 25(6), pp.1382–1396.

Kehl, R. & Van Gool, L., 2006. Markerless Tracking of Complex Human Motions from Multiple Views. Journal on Computer Vision and Image Understanding, 104(2), pp.190–209.

Kessler, J., Schröter, C. & Gross, H., 2004. Approaching a person in a socially acceptable manner using a fast marching planner. In ICRA. pp. 368–377.

Khaleghi, B. et al., 2013. Multisensor data fusion: A review of the state-of-the-art. Information Fusion, 14(1), pp.28–44.

Khan, Z., Balch, T. & Dellaert, F., 2005. MCMC-based particle filtering for tracking a variable number of interacting targets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(11), pp.1805–1819.

Kim, K.K.K. et al., 2004. Background modeling and subtraction by codebook construction. 2004 International Conference on Image Processing, 2004. ICIP ’04., 5.

Kirby, R., Simmons, R. & Forlizzi, J., 2009. COMPANION: A Constraint-Optimizing Method for Person-Acceptable Navigation. In Proceedings - IEEE International Workshop on Robot and Human Interactive Communication. pp. 607–612.

Klasing, K. et al., 2009. Comparison of surface normal estimation methods for range sensing applications, Ieee. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5152493.

Kleinehagenbrock, M. et al., 2002. Person tracking with a mobile robot based on multi-modal anchoring. Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication.

Kluge, B., Kohler, C. & Prassler, E., 2001. Fast and robust tracking of multiple moving objects with a laser range finder. Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164), 2.

Knoop, S., Vacek, S. & Dillmann, R., 2006. Sensor fusion for 3D human body tracking with an articulated 3D body model. Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.

Koay, K.L.K.K.L. et al., 2007. Living with Robots: Investigating the Habituation Effect in Participants' Preferences During a Longitudinal Human-Robot Interaction Study. RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication.

Kolsch, M. et al., 2007. Vision-based human motion analysis: An overview. Computer Vision and Image Understanding, 108(1), pp.4–18. Available at: http://www.sciencedirect.com/science/article/pii/S1077314206002293.

Kopf, J. et al., 2007. Joint bilateral upsampling. ACM Transactions on Graphics, 26(3), p.96. Available at: http://portal.acm.org/citation.cfm?doid=1276377.1276497.

Kruse, T. et al., 2010. Dynamic generation and execution of human-aware navigation plans. Proceedings of the International Conference on Autonomous Agents and Multiagent Systems, 1(c), pp.1523–1524. Available at: http://dl.acm.org/citation.cfm?id=1838462.

Kruse, T. et al., 2013. Human-aware robot navigation: A survey. Robotics and Autonomous Systems, 61(12), pp.1726–1743.

Kuncheva, L.I. & Whitaker, C.J., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 51(2), pp.181–207. Available at: http://www.springerlink.com/index/Q16465142221T2QV.pdf.

Kuo, A.D., 2001. A simple model of bipedal walking predicts the preferred speed-step length relationship. Journal of biomechanical engineering, 123(3), pp.264–269.

Kuo, A.D, et al., 2010. Dynamic Principles of Gait and Their Clinical Implications, Journal of American Physical Therapy Association, Phisical Therapy, 90(2), pp.157-174, doi 10.2522/20090125

Kushleyev, A. & Likhachev, M., 2009. Time-bounded lattice for efficient planning in dynamic environments. In Proceedings - IEEE International Conference on Robotics and Automation. pp. 1662–1668.

Laga, H., 2009. Modeling the Spatial Behavior of Virtual Agents in Groups for Non-verbal Communication in Virtual Worlds. ReCALL, (5), pp.154–159.

Lam, C.P. et al., 2011. Human-centered robot navigation-towards a harmoniously human-robot coexisting environment. IEEE Transactions on Robotics, 27(1), pp.99–112.

LaValle, S.M., 2006. Planning Algorithms. Methods, 2006, p.842. Available at: http://ebooks.cambridge.org/ref/id/CBO9780511546877.

Lee, D. & Nakamura, Y., 2007. Motion Capturing from Monocular Vision by Statistical Inference Based on Motion Database: Vector Field Approach. In International Conference on Intelligent Robots and Systems (IROS). pp. 617–623.

Leibe, B., Schindler, K., et al., 2008. Coupled object detection and tracking from static cameras and moving vehicles. IEEE transactions on pattern analysis and machine intelligence, 30(10), pp.1683–1698.

Leibe, B. et al., 2007. Dynamic 3d scene analysis from a moving vehicle. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. pp. 1–8. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4270171\npapers2://publication/uuid/BB049104-3C9D-4768-A797-A6282FF166A4\npapers2://publication/uuid/00DAB485-59C9-4377-A560-149E3E16D7A1\npapers2://publication/uuid/818FB2E4-4302-4BA2-8EC8-2D0B6A461FF9.

Leibe, B., Leonardis, A. & Schiele, B., 2004. Combined Object Categorization and Segmentation with an Implicit Shape Model. In ECCV'04 Workshop on Statistical Learning in Computer Vision. pp. 1–16.

Leibe, B., Leonardis, A. & Schiele, B., 2008. Robust Object Detection with Interleaved Categorization and Segmentation. International Journal of Computer Vision, 77(1-3), pp.259–289.

Leibe, B., Schindler, K. & Gool, L. Van, 2007. Coupled Detection and Trajectory Estimation for Multi-Object Tracking. 2007 IEEE 11th International Conference on Computer Vision.

Leibe, B., Seemann, E. & Schiele, B., 2005. Pedestrian detection in crowded scenes. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1.

De Leva, P., 1996. Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters. Journal of biomechanics, 29(9), pp.1223–1230.

Levine, J.M. & Moreland, R.L., 1998. Small Groups. In D. T. Gilbert, S. T. Fiske, & G. Lindzey, eds. The handbook of social psychology. McGraw-Hill, pp. 415–469. Available at: http://search.epnet.com/login.aspx?direct=true&AuthType=cookie,ip,url,uid&db=psyh&an=1998-07091-026.

Li, S.Z. et al., 2002. Statistical learning of multi-view face detection. In Proc. European Conf. on Computer Vision. Springer, pp. 67–81. Available at: http://www.springerlink.com/index/GK3JLV5GWPF282JY.pdf.

Li, T. et al., 2012. Future Control and Automation W. Deng, ed., Berlin, Heidelberg: Springer Berlin Heidelberg. Available at: http://www.springerlink.com/index/10.1007/978-3-642-31003-4 [Accessed October 25, 2014].

Lienhart, R., Kuranov, A. & Pisarevsky, V., 2003. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In Proceedings of the 25th DAGM Pattern Recognition Symposium. pp. 297–304. Available at: http://www.springerlink.com/index/WYP0XR0WMTMYY4RK.pdf\nhttp://link.springer.com/chapter/10.1007/978-3-540-45243-0_39.

Lopez-Rubio, E. & Luque-Baena, R.M., 2011. Stochastic approximation for background modelling. Computer Vision and Image Understanding, 115(6), pp.735–749.

Lord, J. et al., 2005. GUIDE FOR EVALUATING THE PREDICTIVE CAPABILITIES OF COMPUTER EGRESS MODELS. NIST GCR 06-886.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60, pp.91–110. Available at: http://portal.acm.org/citation.cfm?id=996342.

Lowe, D.G., 1999. Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision, 2.

Luber, M. et al., 2012. Socially-aware robot navigation: A learning approach. In IEEE International Conference on Intelligent Robots and Systems. pp. 902–907.

Luber, M., Spinello, L. & Arras, K.O., 2011. People tracking in RGB-D data with on-line boosted target models. 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.3844–3849.

Luo, H.L.H., 2005. Optimization design of cascaded classifiers. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), 1.

Luo, R.C. et al., 2007. Mobile robot based human detection and tracking using range and intensity data fusion. In Advanced Robotics and Its Social Impacts, 2007. ARSO 2007. IEEE Workshop on. pp. 1–6.

MakeHuman.org, 2014. MakeHuman. Available at: http://www.makehuman.org/ [Accessed November 29, 2014].

Marr, D. & Nishihara, H.K., 1978. Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain), 200(1140), pp.269–294.

Marsh, K.L. et al., 2006. Contrasting Approaches to Perceiving and Acting With Others. Ecological Psychology, 18(1), pp.1–38.

Mello, G.R. et al., 2012. Applications of wavefront technology. Journal of Cataract and Refractive Surgery, 38(9), pp.1671–1683.

Meng, L.M.L., Grimm, W. & Donne, J., 2002. Radar detection improvement by integration of multi-object tracking,

Meyer, M. et al., 2002. Discrete Differential-Geometry Operators for Triangulated 2-Manifolds. In H. C. Hege & K. Polthier, eds. Visualization and Mathematics III. pp. 113–134.

Mikić, I. et al., 2003. Human Body Model Acquisition and Tracking Using Voxel Data. International Journal of Computer Vision, 53(3), pp.199–223. Available at: http://www.springerlink.com/content/q25728g2q821627j.

Mikolajczyk, K., Schmid, C. & Zisserman, A., 2004. Human detection based on a probabilistic assembly of robust part detectors. In T. Pajdla & J. Matas, eds. Proc. European Conference on Computer Vision (ECCV). Springer Berlin Heidelberg, pp. 69–82. Available at: http://www.springerlink.com/content/j576cjbqmc4dqyug/\nhttp://www.cis.pku.edu.cn/faculty/vision/zhangchao/Summer2007/Zisserman_ECCV_2004.pdf\nhttp://www.springerlink.com/content/j576cjbqmc4dqyug\nhttp://www.springerlink.com/content/j576cjbqmc4dqyug/fulltext.pdf.

Moeslund, T.B. & Granum, E., 2001. A Survey of Computer Vision-Based Human Motion Capture. Computer Vision and Image Understanding, 81(3), pp.231–268. Available at: http://linkinghub.elsevier.com/retrieve/pii/S107731420090897X\nhttp://www.sciencedirect.com/science/article/pii/S107731420090897X.

Moeslund, T.B., Hilton, A. & Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 104(2-3), pp.90–126.

Moeslung, T. et al., 2011. Visual Analysis of Humans, Looking at people, Springer.

Moll, G.P. & Rosenhahn, B., 2009. Ball joints for Marker-less human Motion Capture. Applications of Computer Vision (WACV), 2009 Workshop on.

Moosmann, F., Pink, O. & Stiller, C., 2009. Segmentation of 3D lidar data in non-flat urban environments using a local convexity criterion. 2009 IEEE Intelligent Vehicles Symposium.

Mori, G. et al., 2004. Recovering human body configurations: combining segmentation and recognition. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., 2.

Mori, G. & Malik, J., 2002. Estimating human body configurations using shape context matching. Proceedings of the 7th European Conference on Computer Vision (ECCV '02) Part III, pp.666–680. Available at: http://www.springerlink.com/index/l9hb1wf2dw62wqdw.pdf.

Mori, G. & Malik, J., 2006. Recovering 3D human body configurations using shape contexts. IEEE transactions on pattern analysis and machine intelligence, 28(7), pp.1052–1062.

Mucientes, M. & Burgard, W., 2006. Multiple hypothesis tracking of clusters of people. In IEEE International Conference on Intelligent Robots and Systems. pp. 692–697.

Mündermann, L., Corazza, S. & Andriacchi, T.P., 2006. The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications. Journal of neuroengineering and rehabilitation, 3, p.6.

Murphy, K., Torralba, A. & Freeman, W., 2003. Using the forest to see the trees: a graphical model relating features, objects and scenes. Advances in Neural Information …, 53(3), pp.107–114. Available at: http://doi.acm.org/10.1145/1666420.1666446.

Murray, R.M., Li, Z. & Sastry, S.S., 1994. A Mathematical Introduction to Robotic Manipulation, Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Mathematical+Introduction+to+Robotic+Manipulation#0.

Nedevschi, S., Bota, S. & Tomiuc, C., 2009. Stereo-Based Pedestrian Detection for Collision-Avoidance Applications. IEEE Transactions on Intelligent Transportation Systems, 10(3).

Nguyen, A. & Le, B., 2013. 3D Point Cloud Segmentation: A survey. In IEEE International Conference on Robotics, Automation and Mechatronics. IEEE.

Oehler, B. et al., 2011. Efficient multi-resolution plane segmentation of 3D point clouds. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 145–156.

Ohki, T., Nagatani, K. & Yoshida, K., 2010. Collision avoidance method for mobile robot considering motion and personal spaces of evacuees. In IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings. pp. 1819–1824.

Okuma, K. et al., 2004. A Boosted Particle Filter : Multitarget Detection and Tracking. Proceedings of the 8th European Conference on Computer Vision - ECCV 2004, pp.28–39. Available at: http://link.springer.com/chapter/10.1007/978-3-540-24670-1_3\nhttp://www.cs.ubc.ca/~little/links/linked-papers/kenji-eccv2004.pdf.

Ommer, B. & Buhmann, J., 2005. Object Categorization by Compositional Graphical Models. In Energy Minimization Methods in Computer Vision and Pattern Recognition. pp. 235–250. Available at: http://dx.doi.org/10.1007/11585978_16.

Ong, E.J. et al., 2006. Viewpoint invariant exemplar-based 3D human tracking. Computer Vision and Image Understanding, 104(2-3 SPEC. ISS.), pp.178–189.

Papageorgiou, C. & Poggio, T., 2000. Trainable system for object detection. International Journal of Computer Vision, 38(1), pp.15–33.

Papageorgiou, C.P., Oren, M. & Poggio, T., 1998. A general framework for object detection. In Computer Vision, 1998. Sixth International Conference on. pp. 555–562.

Pavlovic, V., Rehg, J.M. & Murphy, K.P., 1999. A dynamic Bayesian network approach to figure tracking using learned dynamic models, IEEE. Available at: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=791203.

Pearson, K. 1901. "On Lines and Planes of Closest Fit to Systems of Points in Space" (PDF). Philosophical Magazine 2(11): 559-572. doi:10.1080/14786440109462720.

Pellegrini, S., Schindler, K. & Nardi, D., 2008. A Generalisation of the ICP Algorithm for Articulated Bodies. *Proceedings of the British Machine Vision Conference*. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.165.4314.

Peng, C.-Y. & Hsu, S., 2012. A note on a Wiener process with measurement error. *Applied Mathematics Letters*, 25(4), pp.729–732. Available at: http://dx.doi.org/10.1016/j.aml.2011.10.010.

Penrose, R. & Todd, J.A., 1956. On best approximate solutions of linear matrix equations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(01), pp.17–19. Available at: http://journals.cambridge.org/abstract_S0305004100030929.

Pham, M.T. & Cham, T.J., 2007. Online learning asymmetric boosted classifiers for object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Phasespace, 2014. PhaseSpace. *Pahsespace website*. Available at: http://www.phasespace.com/impulse-motion-capture.html [Accessed October 19, 2014].

Piccardi, M., 2004. Background subtraction techniques: a review. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, 4.

Plagemann, C. et al., 2010. Real-time identification and localization of body parts from depth images. *Robotics and Automation (ICRA), 2010 IEEE International Conference on*.

Plankers, R. & Fua, P., 2003. Articulated soft objects for multiview shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9).

Platt, J., 2000. Probabilities for support vector machines A. Smola et al., eds. *Advances in Large Margin Classifiers*. Available at: http://research.microsoft.com/~jplatt/abstracts/SVprob.html.

Poppe, R., 2010. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6), pp.976–990.

Poppe, R., 2007. Evaluating Example-based Pose Estimation : Experiments on the HumanEva Sets. *CVPR 2nd Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM2)*.

Poppe, R. & Poel, M., 2006. Comparison of silhouette shape descriptors for example-based human pose recovery. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*.

R. Bro, et. al., 2007. Resolving the Sign Ambiguity in the Singular Value Decomposition. *SANDIA*.

Raghavendra, R. et al., 2011. Particle swarm optimization based fusion of near infrared and visible images for improved face verification. *Pattern Recognition*, 44(2), pp.401–411.

Ralescu, D. & Adams, G., 1980. The fuzzy integral. *Journal of Mathematical Analysis and Applications*, 75(2), pp.562–570.

Reid, D., 1979. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6).

Renze, K.J. & Oliver, J.H., 1996. *Generalized unstructured decimation [computer graphics]*,

Richard, P., 1981. *Robot manipulators:matematics, programming, and control: the computer control of robot manipulators*, MIT Press, MA.

Rios-Martinez, J. et al., 2012. Navigating between people: A stochastic optimization approach. In *Proceedings - IEEE International Conference on Robotics and Automation*. pp. 2880–2885.

Rogez, G. et al., 2008. Randomized trees for human pose detection. *2008 IEEE Conference on Computer Vision and Pattern Recognition*.

Rosales, R. & Sclaroff, S., 2002. Algorithms for inference in Specialized Maps for recovering 3D hand pose. *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*.

Rosales, R. & Sclaroff, S., 2000. Specialized mappings and the estimation of human body pose from a single image. *Proceedings Workshop on Human Motion*.

Ross, A. & Govindarajan, R., 2005. Feature level fusion using hand and face biometrics. In *Proceedings of SPIE - The International Society for Optical Engineering*. pp. 196–204. Available at: http://dx.doi.org/10.1117/12.606093.

Rother, C., Kolmogorov, V. & Blake, A., 2004. “GrabCut.” *ACM Transactions on Graphics*, 23(3), p.309.

Rusinkiewicz, S. & Levoy, M., 2001. Efficient variants of the ICP algorithm. In *Proceedings Third International Conference on 3D Digital Imaging and Modeling*. IEEE Comput. Soc, pp. 145–152. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=924423.

Rusu, R.B., 2009a. Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments. *Representations*, 24(4), pp.345–348. Available at: http://www.springerlink.com/index/10.1007/s13218-010-0059-6.

Rusu, R.B., 2009b. *Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments, PhD thesis*. Technische Universitaet Muenchen.

Rusu, R.B. & Cousins, S., 2011. 3D is here: Point Cloud Library (PCL). *2011 IEEE International Conference on Robotics and Automation*, 29(2), pp.1–4. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5980567.

Scandolo, L. & Fraichard, T., 2011. An anthropomorphic navigation scheme for dynamic scenarios. In *Proceedings - IEEE International Conference on Robotics and Automation*. pp. 809–814.

Schoenberg, J.R., Nathan, A. & Campbell, M., 2010. Segmentation of dense range information in complex urban scenes. *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*.

Schulz, D., 2006. A Probabilistic Exemplar Approach to Combine Laser and Vision for Person Tracking. In *Proc. of Robotics: Science and Systems*.

Schulz, D. et al., 2003. People Tracking with Mobile Robots Using Sample-Based Joint Probabilistic Data Association Filters. *The International Journal of Robotics Research*, 22(2), pp.99–116.

Schumitsch, B. et al., 2006. Identity management Kalman filter (IMKF). In *In Robotics: Science and Systems (RSS)*.

Seemann, E., Leibe, B. & Schiele, B., 2006. Multi-aspect detection of articulated objects. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 1582–1588.

Segal, A. V., Haehnel, D. & Thrun, S., 2009. Generalized-ICP. *Robotics: Science and Systems*. Available at: http://www.stanford.edu/~avsegal/resources/papers/Generalized_ICP.pdf.

Sehestedt, S., Kodagoda, S. & Dissanayake, G., 2010. Robot path planning in a social context. *Robotics Automation and Mechatronics (RAM), 2010 IEEE Conference on*.

Shakarji, C.M., 1998. Least-Squares Fitting Algorithms of the NIST Algorithm Testing System. *Journal of Research of the National Institute of Standards and Techn ology*, 103(6), pp.633–641.

Shan, Y. et al., 2006. Learning exemplar-based categorization for the detection of multi-view multi-pose objects. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 1431–1438.

Sharma, V. & Davis, J.W., 2007. Integrating Appearance and Motion Cues for Simultaneous Detection and Segmentation of Pedestrians. *2007 IEEE 11th International Conference on Computer Vision*.

Shen, C., Lin, X. & Shi, Y., 2009. Human pose estimation from corrupted silhouettes using a sub-manifold voting strategy in latent variable space. *Pattern Recognition Letters*, 30(4), pp.421–431.

Shi, D. et al., 2008. Human-aware robot motion planning with velocity constraints. In *2008 International Symposium on Collaborative Technologies and Systems, CTS'08*. pp. 490–497.

Shi, J.S.J. & Malik, J., 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8).

Shotton, J. et al., 2011. *Real-time human pose recognition in parts from single depth images*, IEEE. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5995316.

Sisbot, E.A. et al., 2007. A Human Aware Mobile Robot Motion Planner. *IEEE Transactions on Robotics*, 23(5).

Sminchisescu, C., 2006. 3D Human Motion Analysis in Monocular Video Techniques and Challenges. *2006 IEEE International Conference on Video and Signal Based Surveillance*.

Sminchisescu, C. et al., 2005. Discriminative density propagation for 3D human motion estimation. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1.

Sminchisescu, C., Kanaujia, A. & Metaxas, D., 2006. Learning Joint Top-Down and Bottom-up Processes for 3D Visual Inference. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2.

Smith, D. & Singh, S., 2006. Approaches to multisensor data fusion in target tracking: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 18(12), pp.1696–1710.

Smith, R. a., 1995. Density, velocity and flow relationships for closely packed crowds. *Safety Science*, 18(4), pp.321–327. Available at: http://linkinghub.elsevier.com/retrieve/pii/0925753594000514.

Šochman, J. & Matas, J., 2005. WaldBoost - Learning for time constrained sequential detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 150–156.

Spearpoint, M. & MacLennan, H. a., 2012. The effect of an ageing and less fit population on the ability of people to egress buildings. *Safety Science*, 50(8), pp.1675–1684. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0925753511003389 [Accessed September 3, 2012].

Stauffer, C. & Grimson, W.E.L., 1999. Adaptive background mixture models for real-time tracking. *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, 2.

Steinhauser, D., Ruepp, O. & Burschka, D., 2008. Motion segmentation and scene classification from 3D LIDAR data. *2008 IEEE Intelligent Vehicles Symposium*.

Strom, J., Richardson, A. & Olson, E., 2010. Graph-based segmentation for colored 3D laser point clouds. *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*.

Stuart, A. & Ord, J.K., 1987. *Kendall's advanced theory of statistics*,

Sudderth, E.B. et al., 2005. Learning hierarchical models of scenes, objects, and parts. *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 2.

Svenstrup, M. et al., 2009. Pose estimation and adaptive robot behaviour for human-robot interaction. In *Proceedings - IEEE International Conference on Robotics and Automation*. pp. 3571–3576.

Svenstrup, M., Bak, T. & Andersen, H.J., 2010. Trajectory planning for robots in dynamic human environments. In *IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010 - Conference Proceedings*. pp. 4293–4298.

Sviestins, E. et al., 2007. Speed adaptation for a robot walking with a human. In *Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on*. pp. 349–356.

Taylor, G. & Kleeman, L., 2004. A multiple hypothesis walking person tracker with switched dynamic model. *Proc. of the Australasian Conference on Robotics …*. Available at: http://www.araa.asn.au/acra/acra2004/papers/taylor.pdf.

Thrun, S. et al., 1999. MINERVA: a second-generation museum tour-guide robot. *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No.99CH36288C)*, 3.

Thrun, S., 2005. *Probabilistic Robotics* S. Thrun, W. Burgard, & D. Fox, eds., MIT Press. Available at: http://portal.acm.org/citation.cfm?doid=504729.504754.

Tipping, M.E., 2000. The Relevance Vector Machine. In *Advances in Neural Information Processing Systems 12*. MIT Press.

Topp, E.A. & Christensen, H.I., 2005. Tracking for following and passing persons. *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Torralba, A., 2003. Contextual Priming for Object Detection. *International Journal of Computer Vision*, 53(2), pp.169–191. Available at: http://dx.doi.org/10.1023/A:1023052124951.

Toyama, K. & Blake, A., 2002. Probabilistic Tracking with Exemplars in a Metric Space. *International Journal of Computer Vision*, 48(1), pp.9–19. Available at: http://www.springerlink.com/index/YDP91308TBRR173N.pdf.

Tsokas, N.A. & Kyriakopoulos, K.J., 2010. A multiple hypothesis people tracker for teams of mobile robots. *Robotics and Automation (ICRA), 2010 IEEE International Conference on*.

Umbach, D. & Jones, K.N., 2003. *A few methods for fitting circles to data*, New York: Institute of Electrical and Electronics Engineers. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1246564.

Urtasun, R., Fleet, D.J. & Fua, P., 2006. 3D People Tracking with Gaussian Process Dynamical Models. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 1.

Vapnik, V.N., 1995. *The Nature of Statistical Learning Theory* M. Jordan & S. L. Lauritzen, eds., Springer. Available at: http://portal.acm.org/citation.cfm?id=211359.

Velodyne Lidar, I., 2010. High Definition Lidar HDL-64E. *Datasheet*. Available at: http://velodynelidar.com/lidar/products/brochure/HDL-64E S2 datasheet_2010_lowres.pdf [Accessed April 30, 2013].

Viola, P. a & Jones, M.J., 2001. Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade. *proc. NIPS*, (December), pp.1311–1318.

Viola, P. & Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1.

Viola, P. & Jones, M., 2001. *Robust real-time face detection*, Springer. Available at: http://www.springerlink.com/openurl.asp?id=doi:10.1023/B:VISI.0000013087.49260.fb.

Viola, P., Jones, M.J. & Snow, D., 2003. Detecting pedestrians using patterns of motion and appearance. *Proceedings Ninth IEEE International Conference on Computer Vision*.

Walk, S. et al., 2010. Disparity statistics for pedestrian detection: combining appearance, motion and stereo. In *ECCV'10: Proceedings of the 11th European conference on Computer vision: Part VI*. Springer-Verlag, pp. 1030–1037.

Walters, M.L., Koay, K.L. & Woods, S.N., 2007. Robot to Human Approaches: Preliminary Results on Comfortable Distances and Preferences. In *AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics*. p. 203–. Available at: http://www.aaai.org/Papers/Symposia/Spring/2007/SS-07-07/SS07-07-022.pdf.

Wang, J., Fleet, D. & Hertzmann, A., 2006. Gaussian process dynamical models. In *Advances in Neural Information Processing Systems*. pp. 1441–1448. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.4340&rep=rep1&type=pdf.

Wang, J. & Yagi, Y., 2009. Adaptive mean-shift tracking with auxiliary particles. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 39(6), pp.1578–1589.

Wang, W.W.W., Zhang, J.Z.J. & Shen, C.S.C., 2010. Improved human detection and classification in thermal images. *Image Processing (ICIP), 2010 17th IEEE International Conference on*.

Weisstein, E.W., 2010. Point-Line Distance --2-Dimensional. *From MathWorld--A Wolfram Web Resource*. Available at: http://mathworld.wolfram.com/Point-LineDistance2-Dimensional.html [Accessed January 24, 2015].

White, F.E., 1991. Data Fusion Lexicon. In *The Data Fusion Subpanel of the Joint Directors of Laboratories, Technical Panel for C3*. p. 15. Available at: http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA529661\nhttp://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA529661.

Wu, B.W.B. et al., 2004. Fast rotation invariant multi-view face detection based on real Adaboost. *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*

Wu, B.W.B. & Nevatia, R., 2005. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 1.

Wu, B.W.B. & Nevatia, R., 2007. Simultaneous Object Detection and Segmentation by Boosting Local Shape Feature based Classifier. *2007 IEEE Conference on Computer Vision and Pattern Recognition*.

Wu, J. et al., 2008. Fast asymmetric learning for cascade face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3), pp.369–382.

Wu, J., Rehg, J. & Mullin, M., 2003. Learning a Rare Event Detection Cascade by Direct Feature Selection. *NIPS*, pp.1–17. Available at: http://smartech.gatech.edu/handle/1853/3228\nhttps://papers.nips.cc/paper/2353-learning-a-rare-event-detection-cascade-by-direct-feature-selection.pdf.

Wu, Z. & Leahy, R., 1993. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11).

Xavier, J. et al., 2005. Fast line, arc/circle and leg detection from laser scan data in a player driver. In *Proceedings - IEEE International Conference on Robotics and Automation*. pp. 3930–3935.

Xiao, R. et al., 2007. Dynamic cascades for face detection. In *Proceedings of the IEEE International Conference on Computer Vision*.

Xiao, R.X.R., Zhu, L.Z.L. & Zhang, H.-J.Z.H.-J., 2003. Boosting chain learning for object detection. *Proceedings Ninth IEEE International Conference on Computer Vision*.

Xie, D.X.D. et al., 2004. *A multi-object tracking system for surveillance video analysis*, Ieee. Available at: http://www.springerlink.com/content/v5hm0r9u243u108h/fulltext.pdf.

Yager, R.R., 1988. ORDERED WEIGHTED AVERAGING AGGREGATION OPERATORS IN MULTICRITERIA DECISIONMAKING. *IEEE Transactions on Systems, Man and Cybernetics*, 18(1), pp.183–190.

Yali, A. & Donald, G., 1997. Shape Quantization and Recognition with Randomized Trees. In *Calcuta Math. Society*. pp. 1545–1588.

Yan, F.Y.F. et al., 2006. A Novel Data Association Algorithm for Object Tracking in Clutter with Application to Tennis Video Analysis. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 1.

Yilmaz, A., Javed, O. & Shah, M., 2006. Object tracking: A survey. *ACM Computing Surveys*, 38(4), p.13. Available at: http://dx.doi.org/10.1145/1177352.1177355.

Yiu, Y.K. & Li, Z.X., 2003. Optimal forward kinematics map for a parallel manipulator with sensor redundancy. *Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation. Computational Intelligence in Robotics and Automation for the New Millennium (Cat. No.03EX694)*, 1.

Zajdel, W., Zivkovic, Z. & Krose, B.J.A., 2005. Keeping Track of Humans: Have I Seen This Person Before? *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*.

Zhang, B.Z.B., 2010. Computer vision vs. human vision. *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on*.

Zhang, C. & Zhang, Z., 2010. A Survey of Recent Advances in Face Detection. *Learning*, (June), p.17. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.167.5270&amp;rep=rep1&amp;type=pdf.

Ziebart, B.D. et al., 2009. Planning-based prediction for pedestrians. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*. pp. 3931–3936.

Zivkovic, Z. & Krose, B., 2007. Part based people detection using 2D range data and images. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*. pp. 214–219.

9 Appendices

Appendix A: The Jacobian of the forward kinematics of a kinematic chain

Here, the derivation of the Jacobian for an example forward kinematic chain from the parameterised human skeleton is presented.

The forward kinematic chain is given as:

$$p(\xi_0, \xi_1, \xi_2, \xi_3) = e^{\hat{\xi}_0} e^{\hat{\xi}_1} e^{\hat{\xi}_2} e^{\hat{\xi}_3} p \tag{9.1}$$

Moreover, the following assumptions are made:

$$\begin{aligned} \xi_0 &= (\omega_0, T_0) \\ \xi_1 &= (\omega_{1,}(l_1, 0{,}0)) \\ \xi_2 &= (\omega_{2,}(l_2, 0{,}0)) \\ \xi_3 &= \big(0, (l_2, 0{,}0)\big) \quad , \end{aligned} \tag{9.2}$$

where l_1, l_2, l_3 and T_0 are fixed and axes of rotation are given by:

$$\omega_i = (\mu_i^x, \mu_i^y, \mu_i^z) \tag{9.3}$$

From (9.1) the definition of Jacobian the following can be written:

$$J_p = \left[\frac{\delta p}{\delta\theta_0^x} \; \frac{\delta p}{\delta\theta_0^y} \; \frac{\delta p}{\delta\theta_0^z} \; \frac{\delta p}{\delta\theta_1^x} \; \frac{\delta p}{\delta\theta_1^y} \; \frac{\delta p}{\delta\theta_1^z} \; \frac{\delta p}{\delta\theta_2^x} \; \frac{\delta p}{\delta\theta_2^y} \; \frac{\delta p}{\delta\theta_2^z}\right], \tag{9.4}$$

where each of the partial derivatives $\frac{\delta p}{\delta\theta_i^x}, \frac{\delta p}{\delta\theta_i^y}, \frac{\delta p}{\delta\theta_i^z}, i = 0, \ldots 3$, is 1x3 column vector

For example, $\frac{\delta p}{\delta\theta_2^y} (\xi_0, \xi_1, \xi_2, \xi_3)$ after differentiation becomes:

$$\frac{\delta p}{\delta\theta_2^y} (\xi_0, \xi_1, \xi_2, \xi_3) = e^{\hat{\xi}_0} e^{\hat{\xi}_1} \frac{\delta e^{\hat{\xi}_2}}{\delta\theta_2^y} e^{\hat{\xi}_3} p \tag{9.5}$$

then, developing it further after replacing $\hat{\xi}_2$:

$$\frac{\delta p}{\delta\theta_2^y} (\xi_0, \xi_1, \xi_2, \xi_3) = e^{\hat{\xi}_0} e^{\hat{\xi}_1} \begin{bmatrix} \frac{\delta e^{\hat{\omega}_2}}{\delta\theta_2^y} & 0 \\ 0 & 0 \end{bmatrix} e^{\hat{\xi}_3} p \tag{9.6}$$

Equation (9.6) is equivalent to:

$$\frac{\delta p}{\delta \theta_2^y}\,(\xi_0, \xi_1, \xi_2, \xi_3) = e^{\hat{\xi}_0} e^{\hat{\xi}_1} \begin{bmatrix} \frac{\delta \widehat{\omega}_2}{\delta \theta_2^y} e^{\widehat{\omega}_2} & 0 \\ 0 & 0 \end{bmatrix} e^{\hat{\xi}_3} p \tag{9.7}$$

From (9.3) the following equation can be written for $\widehat{\omega}_2$:

$$\widehat{\omega}_2 = \begin{bmatrix} 0 & -\theta_2^z & \theta_2^y \\ \theta_2^z & 0 & -\theta_2^x \\ -\theta_2^y & \theta_2^x & 0 \end{bmatrix} \tag{9.8}$$

Then from (9.8), the derivative of $\widehat{\omega}_2$ with respect to μ_2^y is given as:

$$\frac{\delta \widehat{\omega}_2}{\delta \theta_2^y} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \tag{9.9}$$

Substitution of (9.9) into (9.7) gives:

$$\frac{\delta p}{\delta \theta_2^y}\,(\xi_0, \xi_1, \xi_2, \xi_3) = e^{\hat{\xi}_0} e^{\hat{\xi}_1} \begin{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} e^{\widehat{\omega}_2} & 0 \\ 0 & 0 \end{bmatrix} e^{\hat{\xi}_3} p \tag{9.10}$$

The other derivatives in (9.4), i.e. $\frac{\delta p}{\delta \theta_0^x}, \frac{\delta p}{\delta \theta_0^y}, \frac{\delta p}{\delta \theta_0^z}, \frac{\delta p}{\delta \theta_1^x}, \frac{\delta p}{\delta \theta_1^y}, \frac{\delta p}{\delta \theta_1^z}, \frac{\delta p}{\delta \theta_2^x}$ and $\frac{\delta p}{\delta \theta_2^z}$, are calculated in similar way. This leads to an expression for the Jacobian that is needed in the reverse kinematics part of the method.

Appendix B: Geometric features used in the leg detector

- Number of points: n
- Standard deviation, calculated as:

$$\sigma_{\,s} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\|x_{k-}\tilde{x}\|^2} \tag{9.11}$$

where $\tilde{x}$ denotes the centre of mass of the segment and n is the number of points in the segment.

- Width, calculated as the Euclidian distance between the last and the first points of the segment, $w(S_j)$:

$$w(S_j) = \sqrt{(x_n - x_1)^2 + (y_n - y_1)^2} \tag{9.12}$$

- Boundary length
This feature is the length of the poli-line made by the points of the segment. It is defined as:

$$l(S_j) = \sum_{i=1}^{n}\sqrt{(x_{i-1} - x_i)^2 + (y_{i-1} - y_i)^2} \tag{9.13}$$

- Linearity: This feature measures how straight the segment is. It is calculated as a ratio between the width of the segment, $w(S_j)$, and the Boundary length of the segment $l(S_j)$:

$$lin_s = \frac{l(S_j)}{w(S_j)} \tag{9.14}$$

- Circularity:
The circularity feature measures how close to a circle the segment shape is. In this work the circularity is defined as the ratio boundary length of the segment and the length of the arc of the best fitting circle as shown on Figure 9-1 below. In this work the circularity measure $C_{circ}(S_j)$ is defined as:

$$C_{circ}(S_j) = \frac{\sum_{i=1}^{n-1}\sqrt{(x_{i-1} - x_i)^2 + (y_{i-1} - y_i)^2}}{2\pi R\,(\frac{\alpha_{arc}}{360})} \tag{9.15}$$

where α_{arc} is the angle of the arc of the best fitting circle between the start point and end point of the segment. The angle α_{arc} is measured in degrees in degrees, n is the number of measurement points in the segment and x_i, y_i are the Cartesian coordinates of the i^{th} point in the j^{th} segment S_j .

It is easy to check that the following natural geometric features requirements, for such a defined circularity measure, hold:

a) $C_{circ}(S_j) \in (0,1]$;
b) $C_{circ}(S_j) = 1$ if and only if all points of the segment S_j is part of the circle
c) $C_{circ}(S_j)$ is invariant with respect to similarity transformations, e.g. translations, rotations an scaling

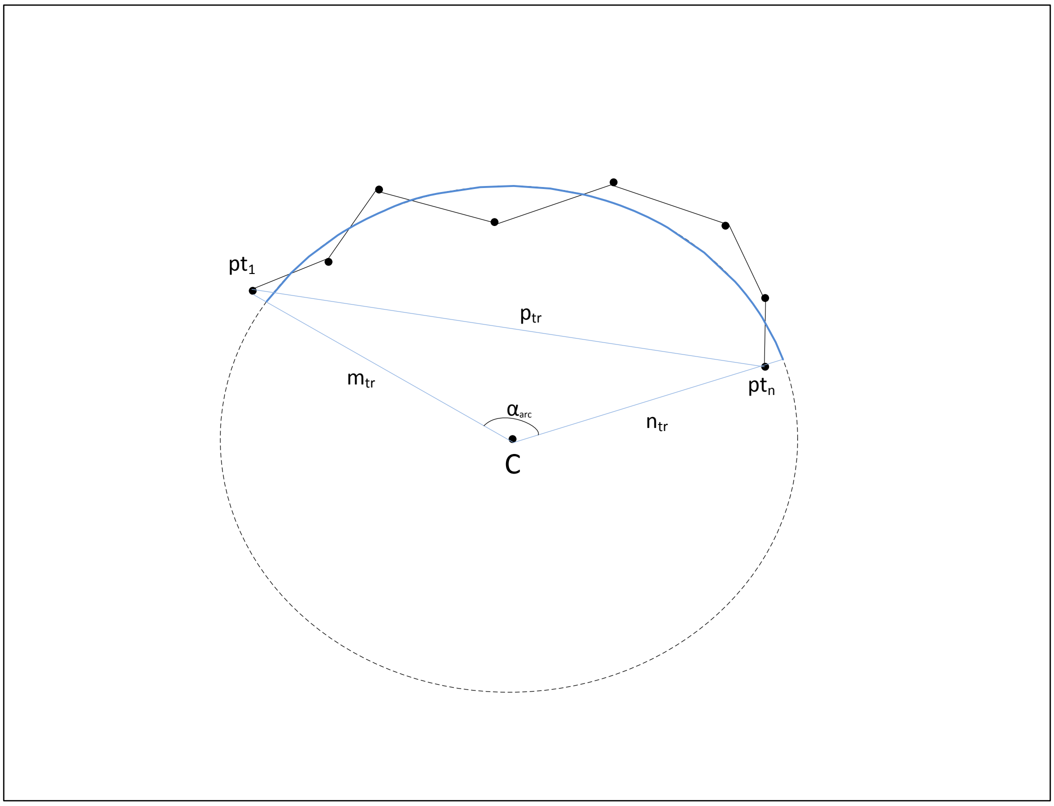


Figure 9-1: Circularity feature of a segment

Fitting the least squares fit (LSF) of a circle to the set of points in the segment is based on minimizing the mean square distance from the circle to the data points, i.e. finding the parameters a, b, r that minimise the objective function:

$$\mathfrak{C} = \sum_{i=1}^{n} d_i^2 \tag{9.16}$$

where d_i is the Euclidean distance from a point (x_i, y_i) to the circle as shown on Figure 9-2 below .

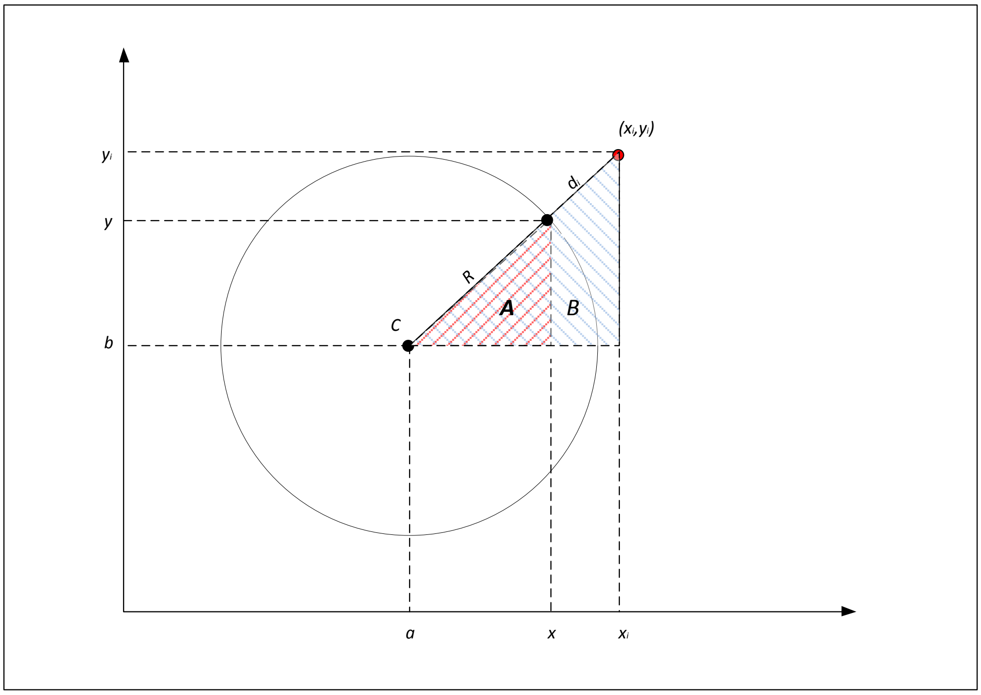


Figure 9-2: Least square fitting of a circle

If radius of the circle is R and the centre of the circle is defined by the point C with coordinates (a, b) then the fitting circle is defined by the equation (9.17). This equation follows from the Pythagorean theorem applied for the three sides of the right angled triangle A, made by the R and the projected distances on the axes between a point from the circle and the centre of the circle C, as shown on the Figure 9-2:

$$(x - a)^2 + (y - b)^2 = R^2 \tag{9.17}$$

The parameters R, a and b fully define the circle. The task of LSF becomes finding those R, a and b that minimise the objective function $\mathfrak{C}$.

When considering the point (x_i, y_i), its Euclidean distance to the circle, d_i, can be found by drawing a line between the point (x_i, y_i) and the centre of the circle,C; then calculating the Euclidean distance between (x_i, y_i) and (a, b), and finally subtracting R, known from (9.17).

The Euclidean distance between (x_i, y_i) and (a, b) can be found by applying the Pythagorean theorem for the triangle B, shown on Figure 9-2 . This gives the following:

$$(R + d_i)^2 = (x_i - a)^2 + (y_i - b)^2 \tag{9.18}$$

Rearranging (9.18) gives:

$$d_i = \sqrt[2]{(x_i - a)^2 + (y_i - b)^2} - R \qquad (9.19)$$

And finally from (9.16), minimizing $\mathfrak{C}$ is reduced to finding the values of the parameters :

$$(a, b, R) = argmin \sum_{i=1}^{n} (\sqrt[2]{(x_i - a)^2 + (y_i - b)^2} - R) \qquad (9.20)$$

However, as seen from (9.20) the minimisation of objective is a nonlinear problem for which no closed form solution exists. Although a number of iterative or approximate methods for computing the minimum of $\mathfrak{C}$ have been proposed (Umbach & Jones 2003), their high computational cost or approximate nature makes them unsuitable candidates for calculation of the circularity feature of the segments.

However, it is possible that the process is transformed into a set of linear equations through an appropriate change of variables and solved through pseudoinverse method (Penrose & Todd 1956) as follows below.

Equation (9.17), through application of the basic arithmetic rule for the squared subtraction, i.e. $(\mathrm{k} - \mathrm{l})^2 = \mathrm{k}^2 - 2\mathrm{kl} + \mathrm{l}^2$, and re-arrangement, can be transformed as:

$$x^2 + y^2 = 2xa + 2yb + (R^2 - a^2 - b^2) \qquad (9.21)$$

If a new variable γ is introduced to replace the term $(R^2 - a^2 - b^2)$ in (9.21) :

$$\gamma = (R^2 - a^2 - b^2) \qquad (9.22)$$

After the substitution equation (9.21) is transformed to:

$$x^2 + y^2 = 2xa + 2yb + \gamma \qquad (9.23)$$

This representation hides the parameter nonlinearities thought the introduction of the new variable. The motivation behind this change in the model is that the linear model which is linear can be solved relatively by relatively trivial means. Thus the problem encountered earlier has been reduced to a series of linear equations which are generated by applying the new model for every data point in the cluster $(x_1, y_2), \ldots, (x_n, y_n)$. As a result the following system of equations can be written:

$$\left| \begin{array}{c} {x_1}^2 + {y_1}^2 = 2x_1a + 2y_1b + \gamma \\ \dots \\ {x_i}^2 + {y_i}^2 = 2x_ia + 2y_ib + \gamma \\ \dots \\ {x_n}^2 + {y_n}^2 = 2x_na + 2y_nb + \gamma \end{array} \right. \tag{9.24}$$

This allows the construction of the following matrix equation in the form: $q = A.p$ as follows.

$$\begin{bmatrix} {x_1}^2 + {y_1}^2 \\ {x_2}^2 + {y_2}^2 \\ {x_3}^2 + {y_3}^2 \\ \dots \\ {x_n}^2 + {y_n}^2 \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \\ & \dots & \\ x_n & y_n & 1 \end{bmatrix} . \begin{bmatrix} a \\ b \\ \gamma \end{bmatrix} \tag{9.25}$$

where $p = [a\ b\ \gamma]^T$ is the vector of the unknown variables, finding of which is the intermediate target before computing the parameters of the circle.

Since the points forming the clusters are presumably noisy and most probably are not exactly on the circle, the system is overdetermined, i.e. more independent equations than unknown parameters. Moreover, because of noise the result of any subset of random equations will differ from any other subset of equations with the real possibility of a significant noise in any single measurement to skew the resulting equation.

Therefore, to reduce the influence of the noise the error is defined to be the squared distance between the point and the circle, i.e. the L_2 norm. This is the same measure $\mathfrak{C}$ defined earlier in (9.19).

The matrix equation (9.25) is solved using pseudoinverse:

$$p = (A^TA)^{-1}A^Tq \tag{9.26}$$

After $[a\ b\ \mu]^T$ is computed, the radius of the best fitting circle R is calculated from the definition of μ, which was introduced in (9.23) as:

$$R = \sqrt{\gamma + a^2 + b^2} \tag{9.27}$$

After the centre of the best fitting centre is known, i.e. the point C with coordinates (a, b) , applying the law of cosines from Euclidean geometry results in following equation that can be written for the triangle with sides m_{tr}, n_{tr} and p_{tr}, that are defined by the centre of the circle C, the first point of the segment pt_1 and the last point of the segment pt_n, as shown in Figure 9-1 above:

$$p_{tr}^{\ 2} = m_{tr}^{\ 2} + n_{tr}^{\ 2} - 2m_{tr}n_{tr}\cos(\alpha_{arc}) \,, \tag{9.28}$$

where:

$$p_{tr} = \sqrt{(x_n - x_1)^2 + (y_n - y_1)^2} \tag{9.29}$$

$$m_{tr} = \sqrt{(a - x_1)^2 + (y_1 - b)^2} \tag{9.30}$$

$$n_{tr} = \sqrt{(x_n - a)^2 + (y_n - b)^2} \tag{9.31}$$

After rearranging (9.28) for the angle of the best fitting circle arch of the segment S_j the following equation can be written:

$$\alpha_{arc} = \arccos(\frac{p_{tr}^{\ 2} - m_{tr}^{\ 2} + n_{tr}^{\ 2}}{2\sqrt{m_{tr}}\sqrt{n_{tr}}}) \tag{9.32}$$

Finally after substituting (9.32) and (9.27) in (9.15) the circularity measure $C_{circ}(S_j)$ can be computed.

- Radius

After solving the matrix equation (9.25) and finding the vector $p = [a\ b\ \gamma]^T$ of unknown values, the parameters of the best fitting circle, including its radius, R, are computer according the equation (9.29).

- Mean curvature

  The mean curvature of the segment is calculated as the average of the curvatures at the measurement points along the segment. For a curvature at a point of the segment, the centre of the curvature is defined by Cauchy as the insertion point of two infinitely close normal to the curve, the radius of the curvature as the distance from the point to the centre of the curvature, and the curvature itself as the inverse of the radius of the curvature (Borovik & Katz 2011).

  An illustrative example of curvature measure is shown in Figure 9-3 below. If the length of a small sub-segment at point A is L and this sub-segment is expanded along the normal vector field, the expanded sub-segment will have length change ΔL by a factor of $(1 + \epsilon k)$, where k is the curvature.

The method used for curvature estimation in the leg detector is based on the idea of building a parameterised curve from the points in the segment. Then, the parameterised curve is used to calculate curvature at points.

The curvature of a plane parametric curve $y = p(x)$, as defined in (Faux & Pratt 1979, p28) , is:

$$k(x) = \frac{p''(x)}{[1 + \{p'(x)\}]^{\frac{3}{2}}} \tag{9.33}$$

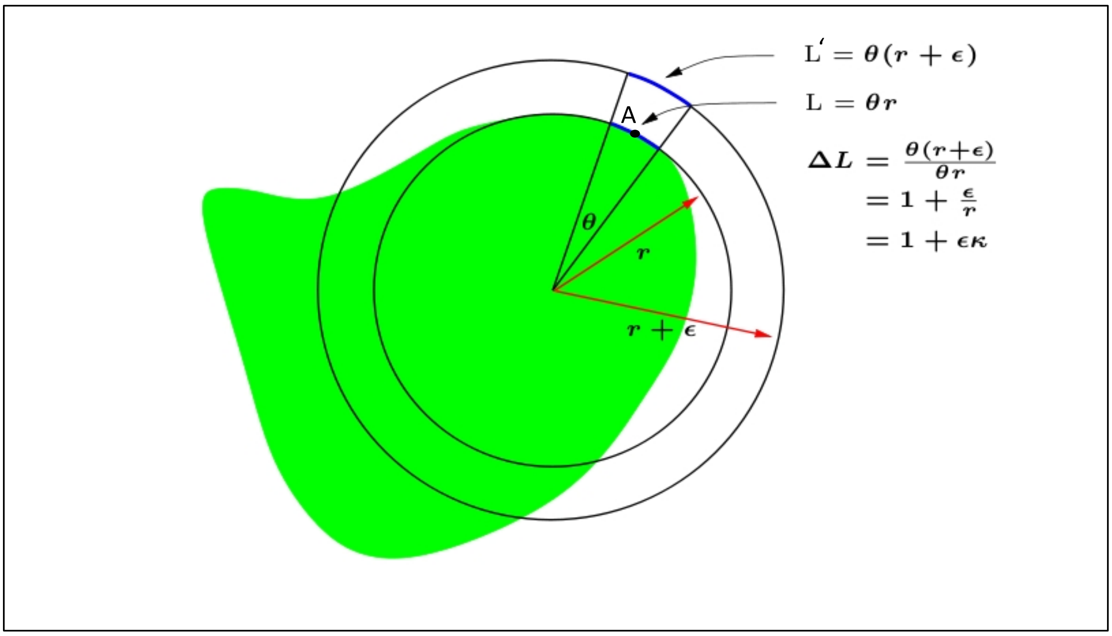


Figure 9-3: Curvature at a point of a curve

The parameterised curve is computed by approximating of the curve of the leg segment locally at point q_i by second order polynomials using also a two neighbouring points with indexes q_{i-k} and q_{i+k}.

The reasoning behind selecting only two neighbouring points is related to one of the fundamental properties of the polynomials stating that: in given $d + 1$ pairs (x_1, y_1), ... , (x_{d+1}, y_{d+1}), with all x_i distinct, there is unique polynomial $p(x)$ of degree at most d such that $p(x_i) = y_i$ for $1 \leq i\ d + 1$.

The approximating polynomial is defined by the coefficients a_2, a_1, a_0 as follows:

$$y = p(x) = a_2x^2 + a_1x + a_0 \tag{9.34}$$

Given the three points, $\{q_{i-k}, q_i, q_{i+k}\}$, computation of the coefficients needed for (9.34) is done by Lagrange interpolation method (Hazewinkel 2001). According to the method, if $\Delta_i(x)$ denotes the degree d-polynomial that goes through these $d+1$ points, then:

$$\Delta_i(x) = \frac{\prod_{j\neq i}(x - x_j)}{\prod_{j\neq i}(x_i - x_j)} \tag{9.35}$$

and

$$p(x) = \sum_{i=1}^{d+1} y_i\, \Delta_i(x) \tag{9.36}$$

For an illustration, supposing that a second order polynomial $p(x)$ that passes through the three points (1,1), (2,2) and (3,4) has to be found. Then applying Lagrange interpolation method and (9.35) for $d = 2$ and $i = \{1,2,3\}$ gives the following equations:

$$\begin{aligned} \Delta_1(x) = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} &= \frac{(x-2)(x-3)}{2} \\ &= \frac{1}{2}x^2 - \frac{5}{2}x + 3 \end{aligned} \tag{9.37}$$

$$\begin{aligned} \Delta_2(x) = \frac{(x - x_3)(x - x_1)}{(x_2 - x_3)(x_2 - x_1)} &= \frac{(x-3)(x-1)}{-1} \\ &= -x^2 + 4x - 3 \end{aligned} \tag{9.38}$$

$$\begin{aligned} \Delta_3(x) = \frac{(x - x_2)(x - x_1)}{(x_3 - x_2)(x_3 - x_1)} &= \frac{(x-2)(x-1)}{2} \\ &= \frac{1}{2}x^2 - \frac{3}{2}x + 3 \end{aligned} \tag{9.39}$$

The polynomial $p(x)$ is therefore given by:

$$p(x) = 1.\Delta_1(x) + 2.\Delta_2(x) + 4.\Delta_3(x) = \frac{1}{2}x^2 - \frac{1}{2}x - 1 \tag{9.40}$$

Indeed, upon verification with the point coordinates, the polynomial in (9.40) passes through the three given points.

Then, the next step is calculating the derivatives of $p(x)$ in (9.34), represented in Bézier form as defined in (Farin 1997) :

$$p'(x) = 2a_2x + a_1 \tag{9.41}$$

and

$$p''(x) = 2a_2 \tag{9.42}$$

Finally, the estimation of the curvature at that point can be computed by substitution of derivatives $p'(x)$ and $p''(x)$ in equation (9.33).

Appendix C: List of developed software modules

The following pieces of software were developed in this work to evaluate and demonstrate the work of the proposed methods:

- Classifier fusion algorithm, including human leg tracker from laser range finder and HISP based human body parts classification algorithm from depth data;
- GLAICP based human body pose tracking algorithm;
- Algorithm for human track generation from noisy data, analysis of the human; interactions and minimally intrusive path navigation planning.

Appendix D: Leg Distortion Appearance Model - Modelling of Distortion in Detecting a Moving Human Leg with a Sequential Laser Range Finder

In addition to the explanation of the underlying principles causing the distortion the Leg Appearance Model (LAM) is developed to facilitate the computation of the distortion in scanning with a laser range finder. This is achieved by considering the tangential, V_{leg_x}, and radial,V_{leg_y}, components of the leg's speed, V_{leg}, separately as follows.

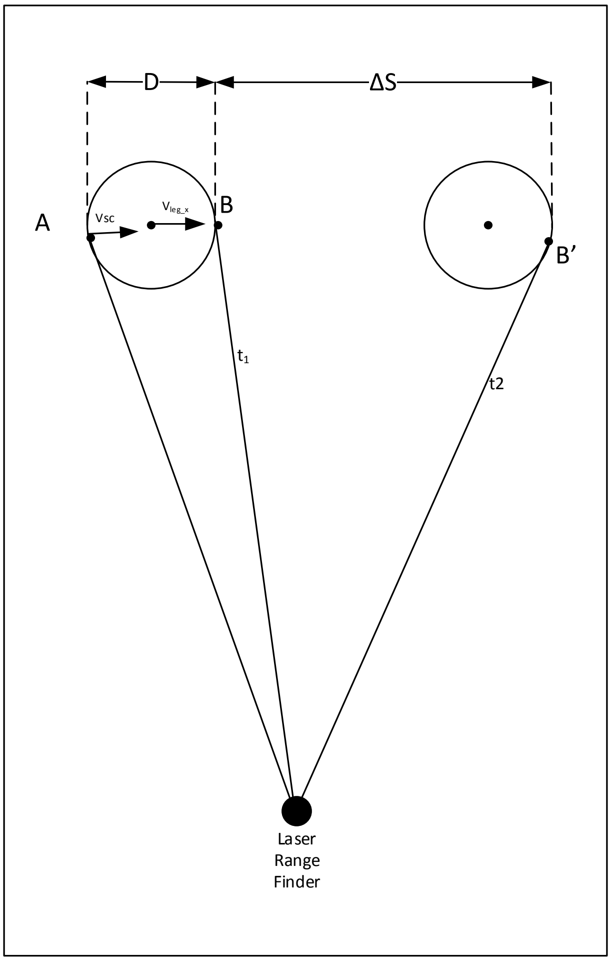


Figure 9-4: Calculation of the distortion resulting from the tangential leg motion component

From Figure 9-4, after approximations, e.g. that the distance to the leg does not change and V_{leg_x} is constant for the duration of the scan of the leg, the following equations can be written for the tangential motion component:

$$\left| \begin{aligned} w_{sc} d_1 (t_1 + t_1) &= D + \Delta S \\ \Delta S &= V_{leg\_x}(t_1 + t_2) \end{aligned} \right. \tag{9.43}$$

Which after some rearrangement gives:

$$\Delta S = D \frac{V_{leg\_x}}{w_{sc} d_1 - V_{leg\_x}} \tag{9.44}$$

and

$$V_{leg\_x} = V_{sc} \frac{1}{\frac{D}{\Delta S} + 1} \tag{9.45}$$

The radial motion component case is considered below and is illustrated in the Figure 9-5.

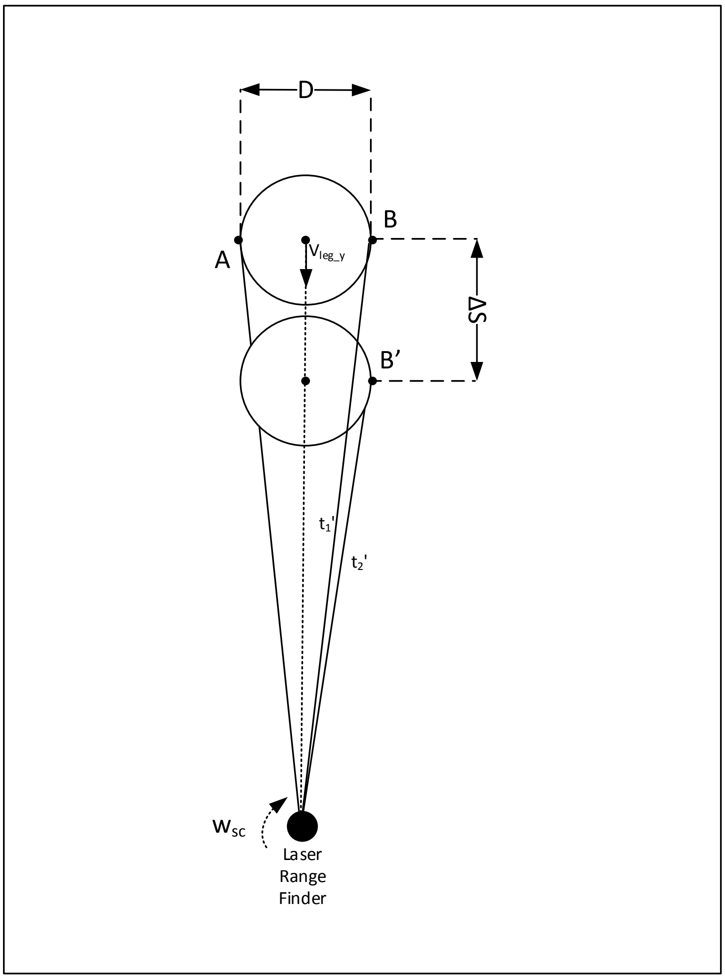


Figure 9-5: Computation of the distortion resulting from the radial leg motion component

The following equations can be given for the Figure 9-5 above.

$$\Delta S = a - b \tag{9.46}$$

and assuming constant V_{leg} for the duration of the scanning of the leg

$$\Delta S = V_{leg}(t_1 + t_2) \tag{9.47}$$

Then, using the Pythagoras theorem, (9.46) can be written as:

$$\Delta S = \sqrt{{d_1}^2 + R^2} - \sqrt{{d_2}^2 + R^2} \tag{9.48}$$

As the times t_1, t_2 can be approximately found using the angular scan speed of the scanner w_{sc} as $t_1 = \frac{R}{w_{sc}d_1}$ and $t_2 = \frac{R}{w_{sc}d_2}$ then (9.47) can be written as:

$$\Delta S = \frac{V_{leg}R}{w_{sc}d_1}\frac{(d_1 + d_2)}{d_2} \tag{9.49}$$

As typically the radius of a human leg R is much smaller in comparison to measurement distances, d_1 and d_2, then (9.48) can be given as:

$$\Delta S = d_1 - d_2 \tag{9.50}$$

After substitution of (9.50) into (9.49) and rearrangement the following reduced quadratic equation can be written:

$$\Delta S^2 - \left(d_1 + \frac{V_{leg}R}{w_{sc}d_1}\right)\Delta S + \frac{2V_{leg}R}{w_{sc}} = 0 \tag{9.51}$$

The solution of (9.51), which follows from the quadratic formula through completing the square, is then given as:

$$\Delta S = \frac{1}{2}\left(d_1 + \frac{V_{leg}R}{w_{sc}d_1}\right) \pm \sqrt{\frac{\left(d_1 + \frac{V_{leg}R}{w_{sc}d_1}\right)^2}{4} - \frac{2V_{leg}R}{w_{sc}}} \tag{9.52}$$

The equation (9.52), which represents the radial distortion of the image ΔS after the scan as a dependence of the parameters known or estimated at the beginning of the scan, allows prediction of the shape of the leg and the subsequent filtering of spurious detections.

Finally, from (9.49) and (9.50) with re-arrangement it follows that:

$$V_{leg\_y} = \frac{w_{sc}d_1(d_1\Delta S - \Delta S^2)}{R(2d_1 - \Delta S)} \tag{9.53}$$

Equation (9.53) above allows estimation of the radial component of the speed of the leg, V_{leg_y},based on the distortion of the image, ΔS.

Appendix E: Fixed Association Model

From the HMM, shown in Figure 6-5, of the fixed data associations the following equations can be given.

If the detection vector of the complete detection history can be represented in a stacked form by:

$$Y_{1:t} = \begin{bmatrix} Y_1 \\ \cdot \\ \cdot \\ Y_t \end{bmatrix} \tag{9.54}$$

Temporarily ignoring θ and applying the Bayes' rule to the last element of the detection vector gives:

$$P(X_t|Y_{1:t}) = \frac{P(Y_t|X_t, Y_{1:t-1})P(X_t|Y_{1:t-1})}{P(Y_t|Y_{1:t-1})} \tag{9.55}$$

From the Markov property assumption it follows that the detection observed at time t is conditionally independent of previously detection values given the current state value X_t. Therefore:

$$P(Y_t|X_t, Y_{1:t-1}) = P(Y_t|X_t) \tag{9.56}$$

After substitution we have the following equation which is in the update step of the optimal filter:

$$P(X_t|Y_{1:t}) = \frac{P(Y_t|X_t)P(X_t|Y_{1:t-1})}{P(Y_t|Y_{1:t-1})} \tag{9.57}$$

Where the normalisation constant is given as:

$$Z_t = P(Y_t|Y_{1:t-1}) = \int_{-\infty}^{\infty} P(Y_t|X_t)P(X_t|Y_{1:t-1})dX_t \tag{9.58}$$

The effect of a time step is obtained by observing that:

$$P(X_{t+1}, X_t|Y_{1:t}) = P(X_{t+1}|X_t, Y_{1:t})P(X_t|Y_{1:t}) \tag{9.59}$$

And then from (9.59) follows that:

$$P(X_{t+1}, X_t | Y_{1:t}) = P(X_{t+1}|X_t)P(X_t|Y_{1:t}) \tag{9.60}$$

which follows from the assumption that the process $\{x_t\}$ is Markovian, and that x_{t+1} is independent of Y_t when x_t is given. Integrating both sides to respect to X_t gives the time update equation:

$$P(X_{t+1}|Y_{1:t}) = \int_{-\infty}^{\infty} P\,(X_{t+1}|X_t)P(X_t|Y_{1:t})dX_t \tag{9.61}$$

which is also referred as Chapman-Kolmogorov equation (Jazwinski 1970) and is used in the prediction step of a filter.

The equation (9.61) written for the previous time step, $t - 1$, then will become:

$$P(X_t|Y_{1:t-1}) = \int_{-\infty}^{\infty} P\,(X_t|X_{t-1})P(X_{t-1}|Y_{1:t-1})dX_{t-1} \tag{9.62}$$

Finally, the following equation can be given:

$$P(X_t|Y_{1:t}) = \frac{P(Y_t|X_t)P(X_t|X_{t-1})P(X_{t-1}|Y_{1:t-1})}{P(Y_t|Y_{1:t-1})} \tag{9.63}$$

Equation (9.63) above gives the recursive filtering calculation between the previous time frame at t-1 and the current one.

Appendix F: Additional experiments of the leg tracker

Additional four experiments were carried with variable number of people. The localisation error was averaged for all participants.

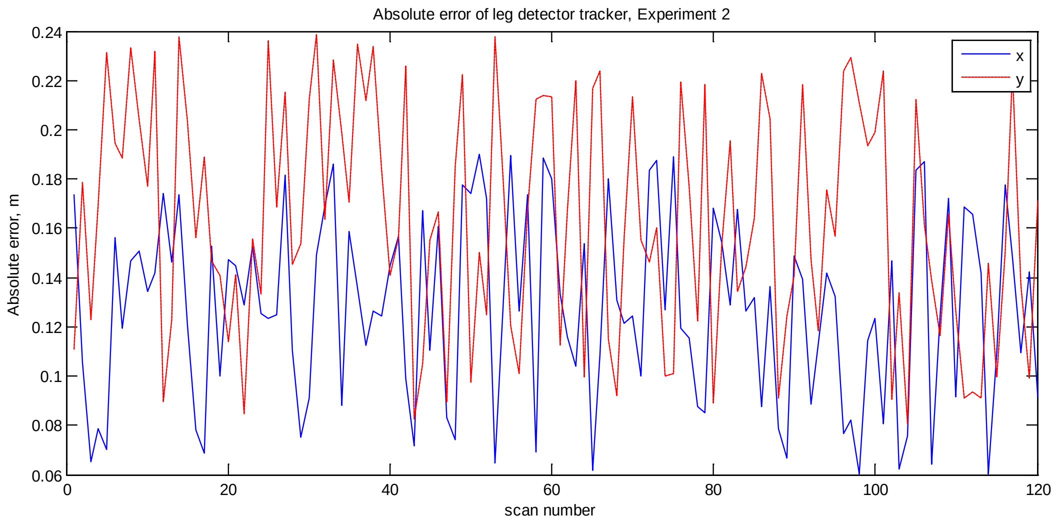


Figure 9-6: Leg detector experiment with two people at the opposite end of the room

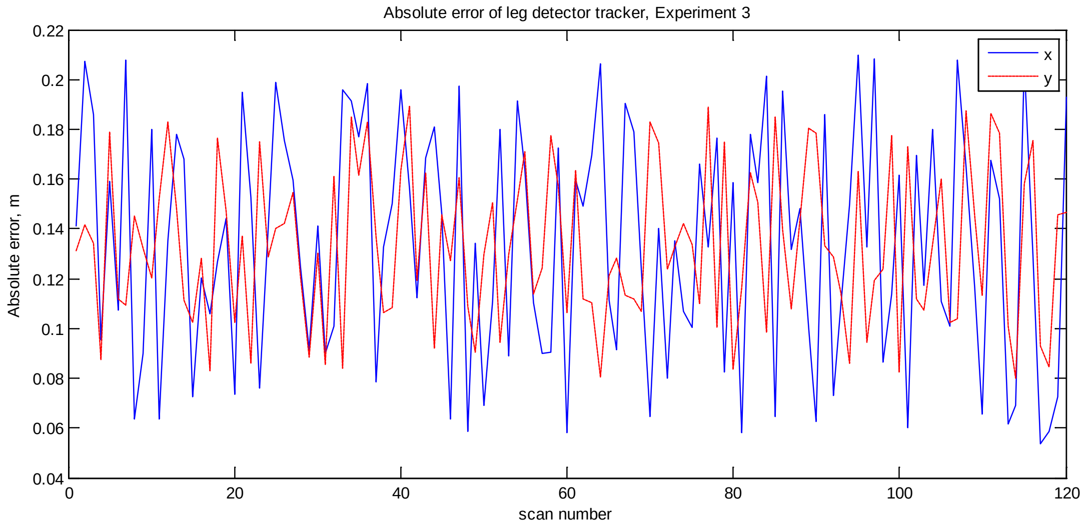


Figure 9-7: Leg detector with two people near the centre of the room

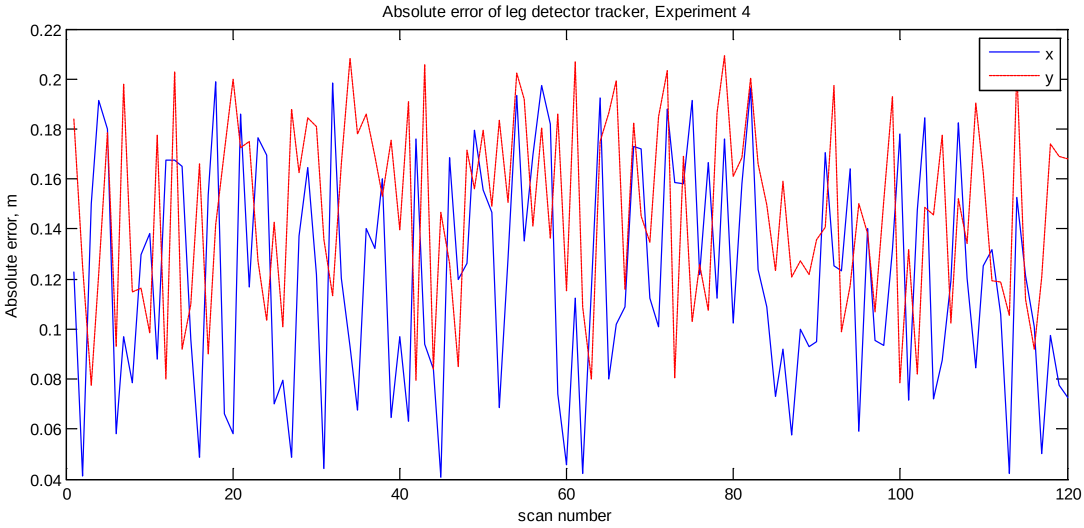


Figure 9-8: Leg detector experiment with three people at the edges of the room

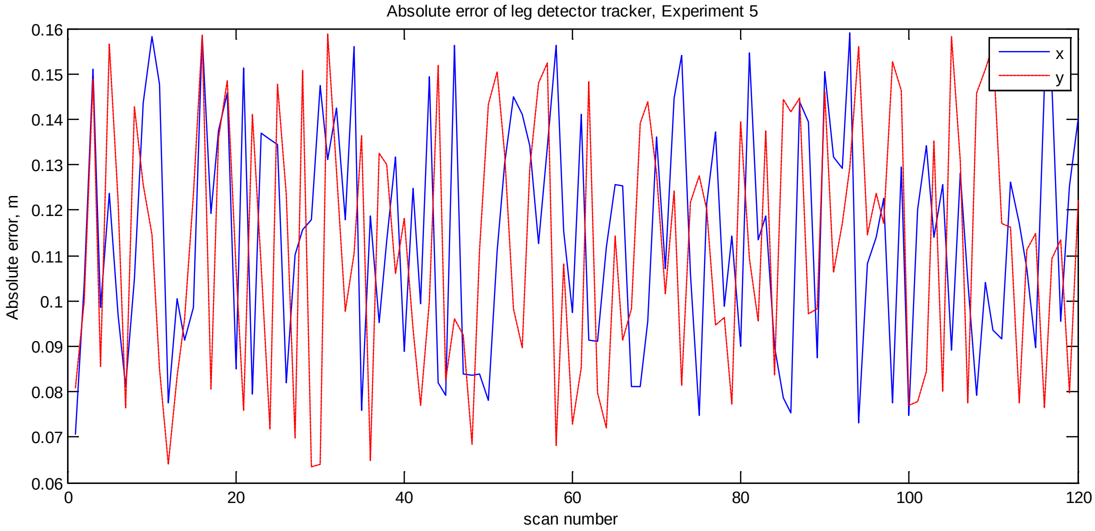
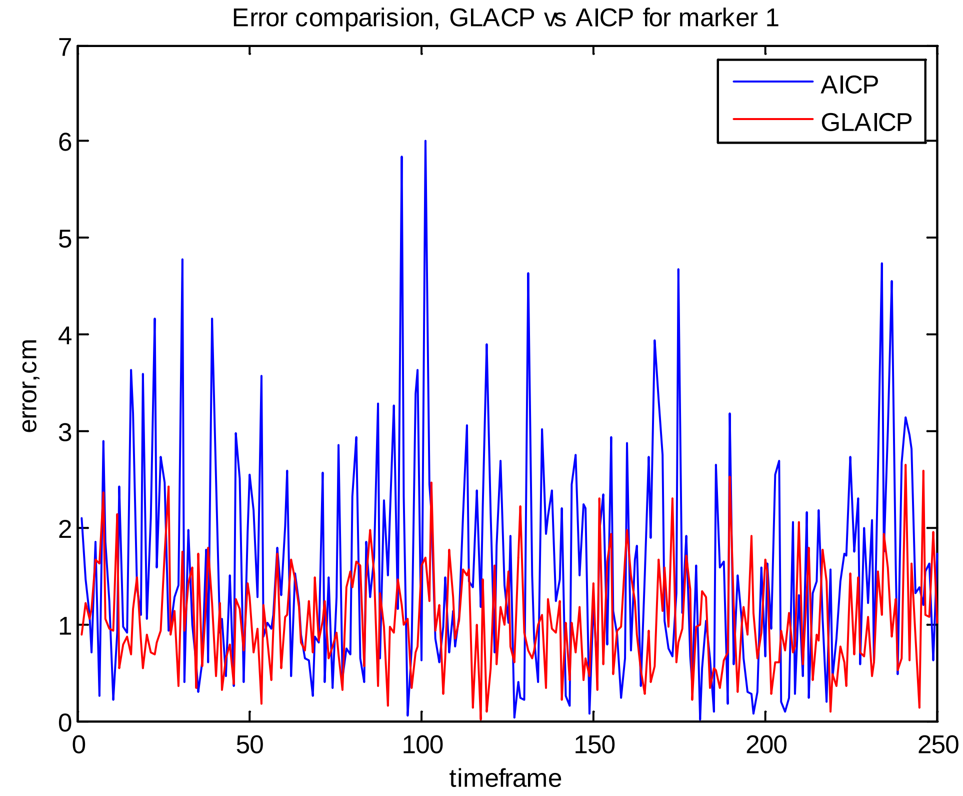
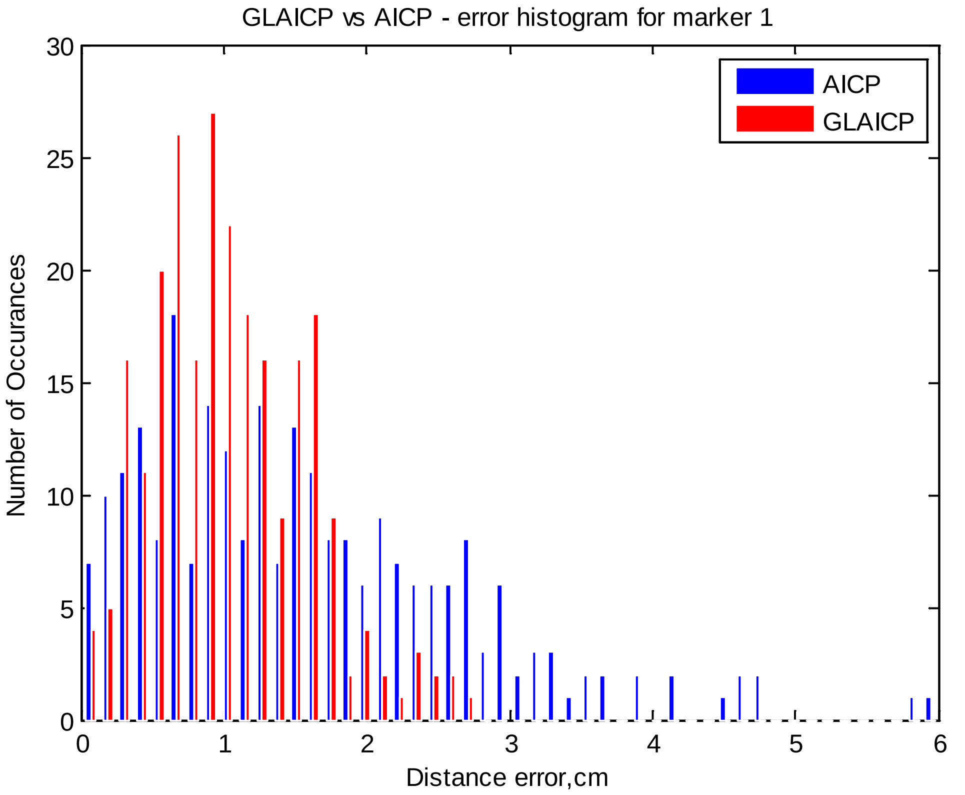
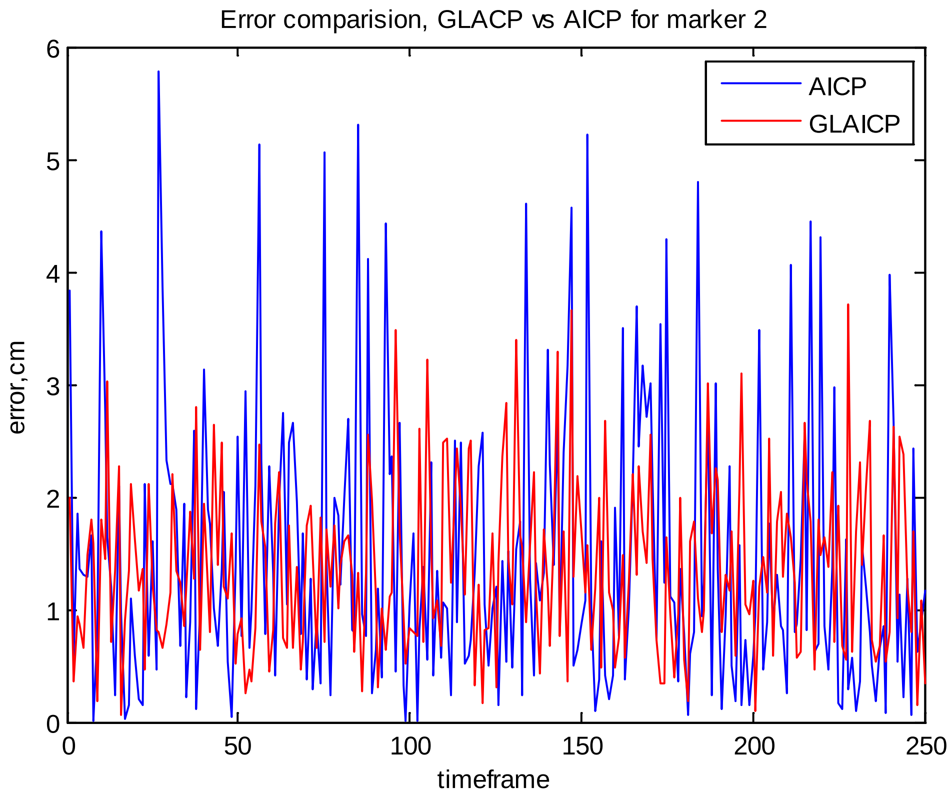


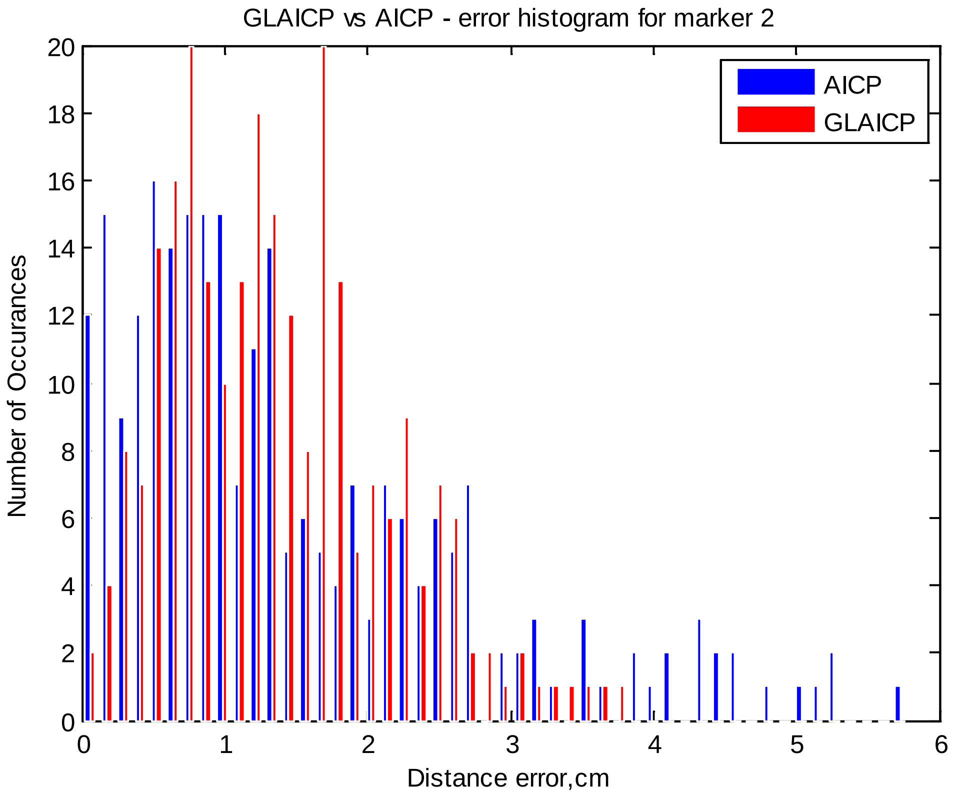
Figure 9-9: Leg detector experiment with three people near the centre of the room

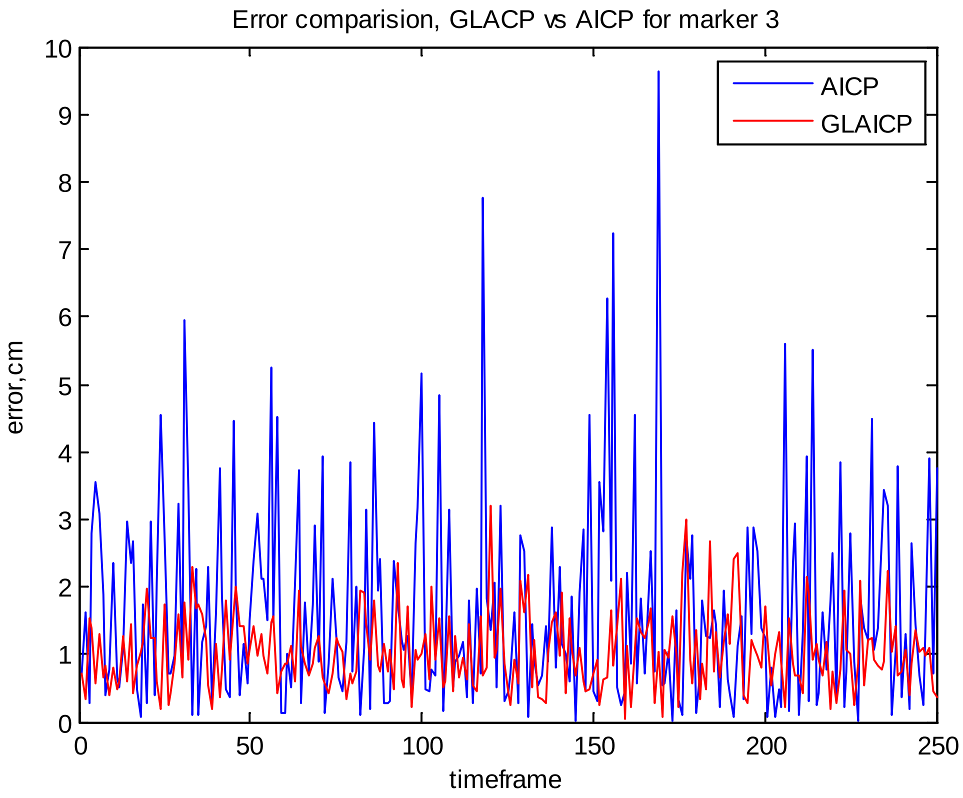
Appendix G: Additional experiments of the GLAICP algorithm

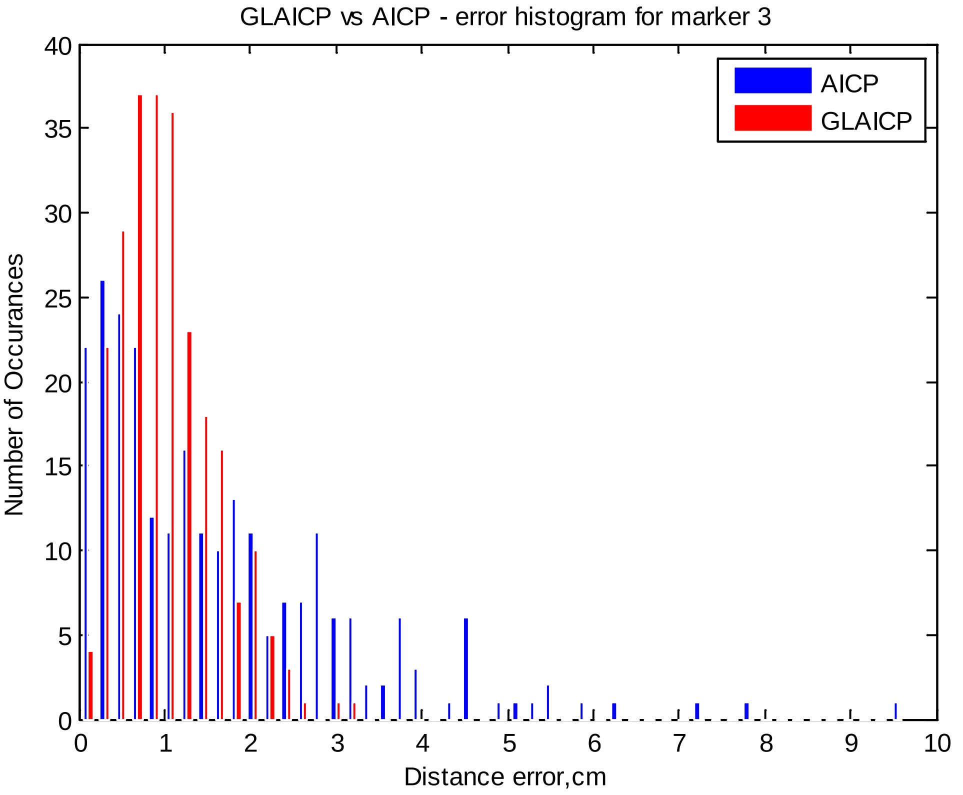


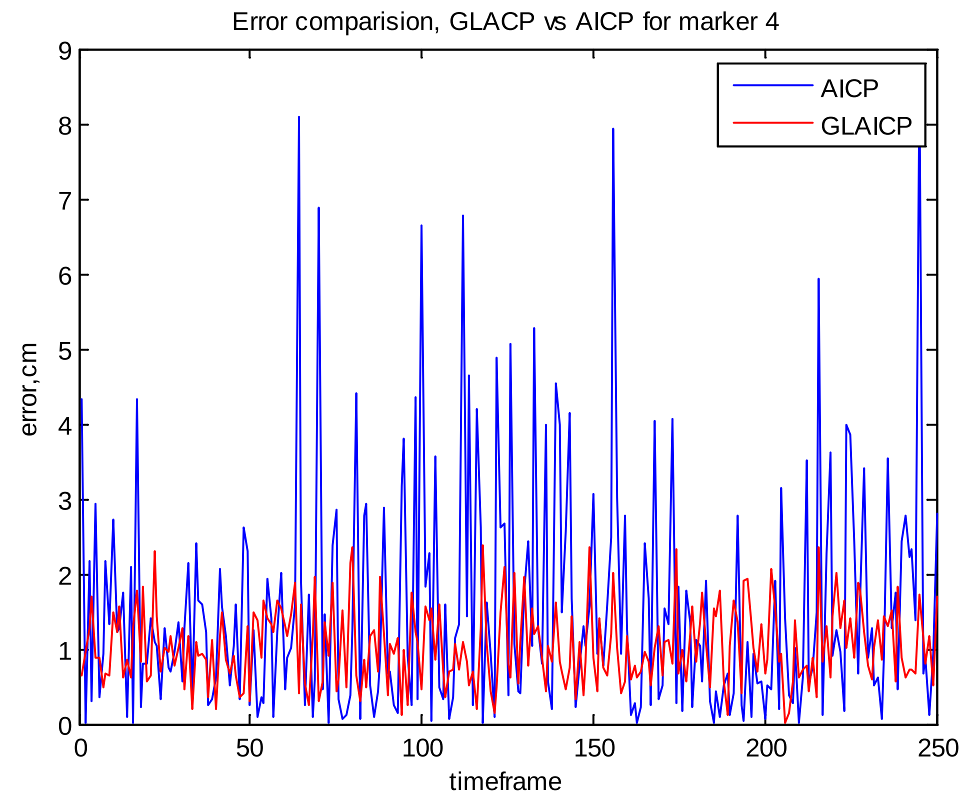


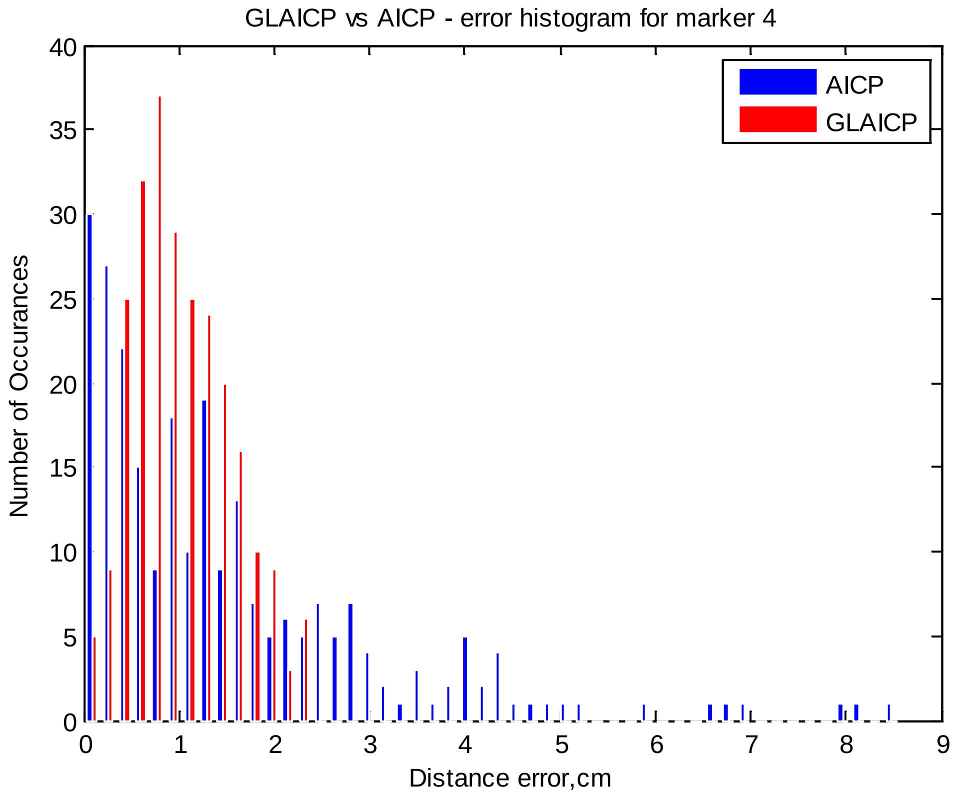

Error comparision, GLACP vs AICP for marker 2
AICP
GLAICP
error,cm
0
1
2
3
4
5
6
0
50
100
150
200
250
timeframe

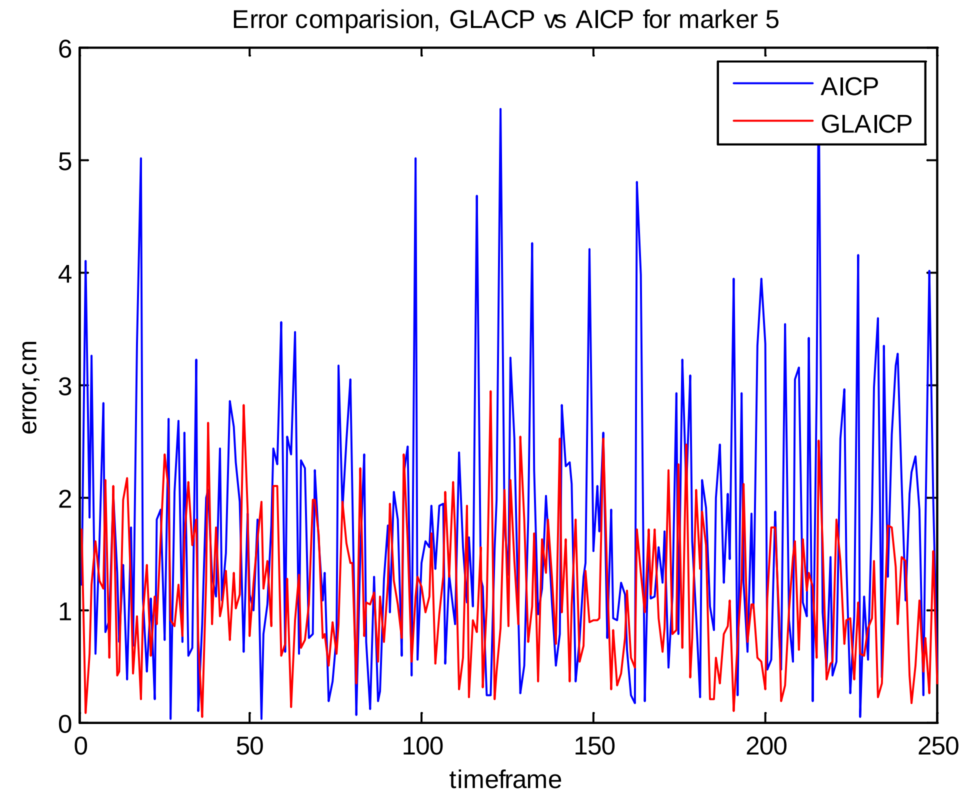

GLAICP vs AICP - error histogram for marker 2
AICP
GLAICP
Number of Occurances
0
2
4
6
8
10
12
14
16
18
20
0
1
2
3
4
5
6
Distance error,cm

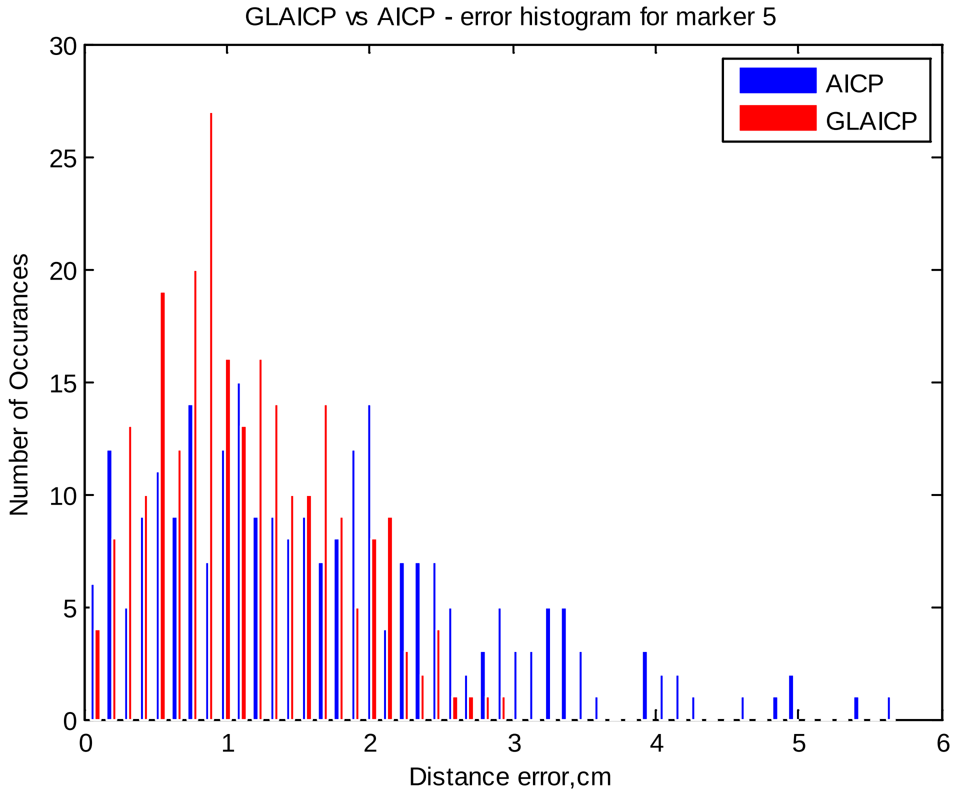

Error comparision, GLACP vs AICP for marker 3
AICP
GLAICP
error,cm
timeframe
0
1
2
3
4
5
6
7
8
9
10
0
50
100
150
200
250

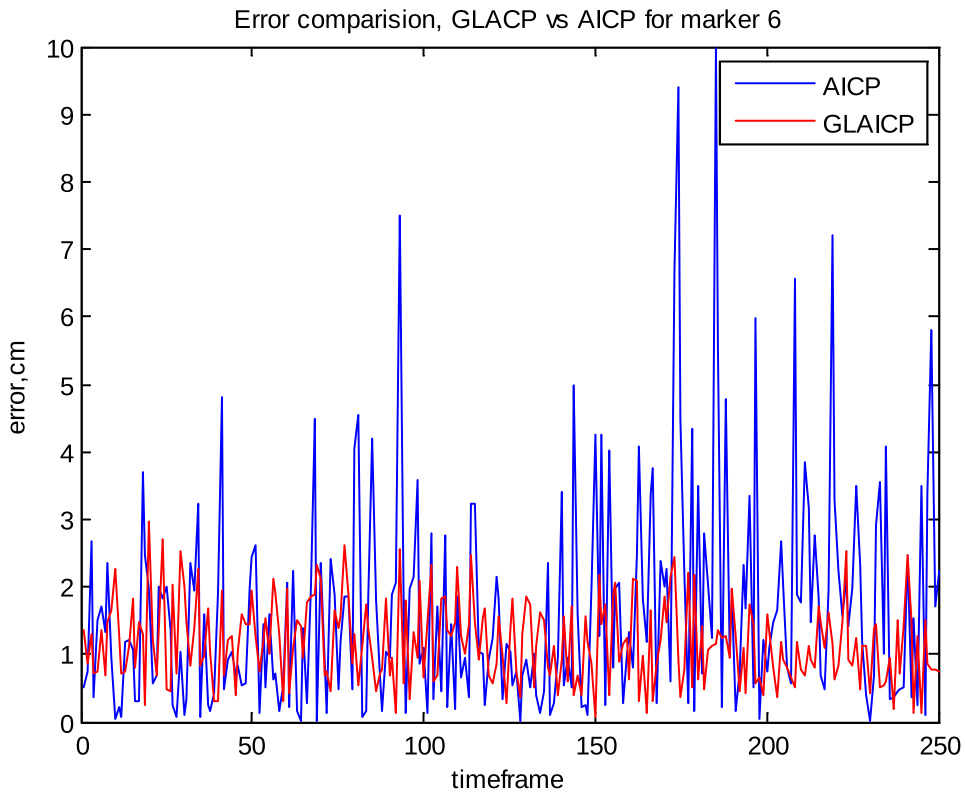

GLAICP vs AICP - error histogram for marker 3
AICP
GLAICP
Number of Occurances
Distance error,cm
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5
6
7
8
9
10

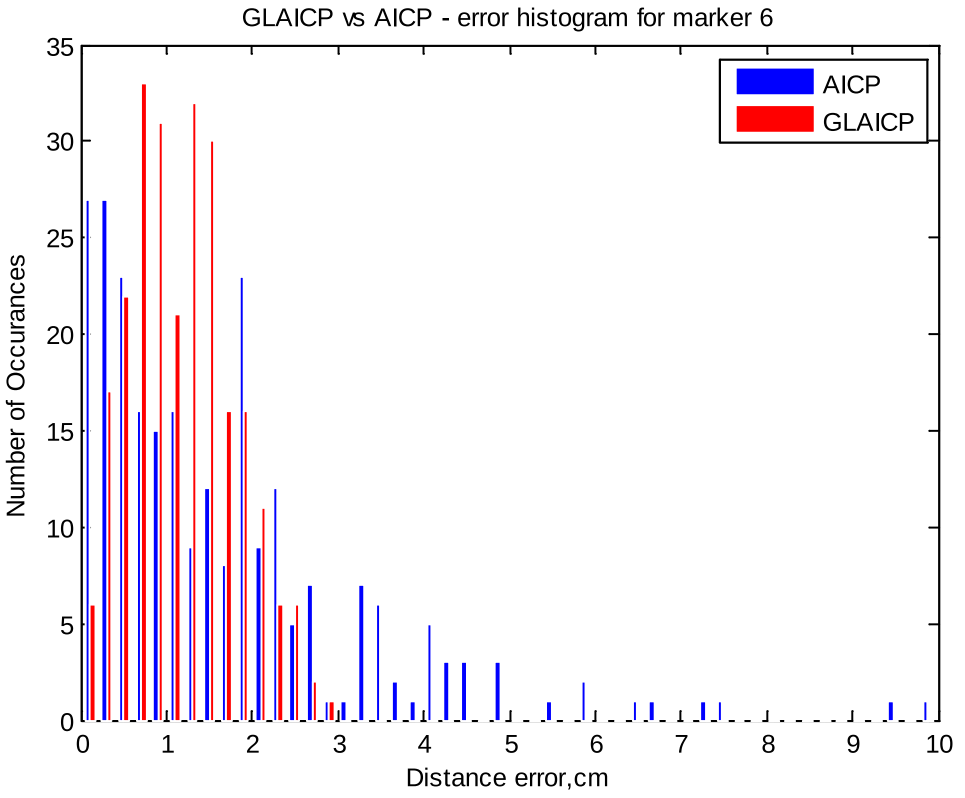

Error comparision, GLACP vs AICP for marker 4
error,cm
timeframe
AICP
GLAICP
0
1
2
3
4
5
6
7
8
9
0
50
100
150
200
250

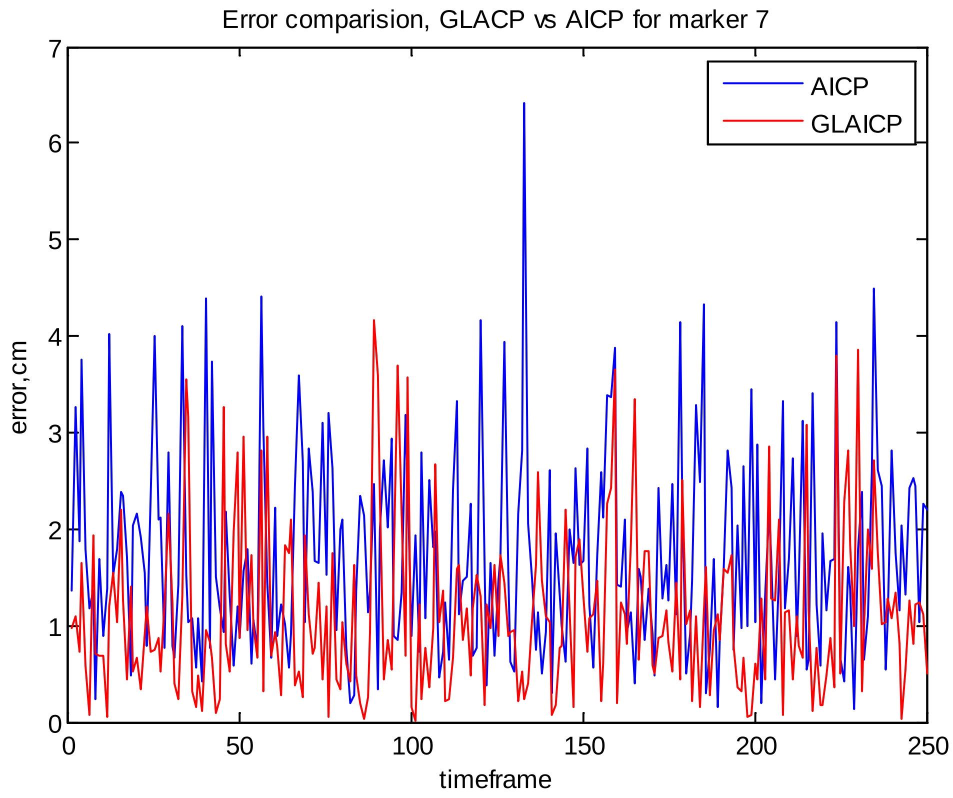

GLAICP vs AICP - error histogram for marker 4
Number of Occurances
Distance error,cm
AICP
GLAICP
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5
6
7
8
9

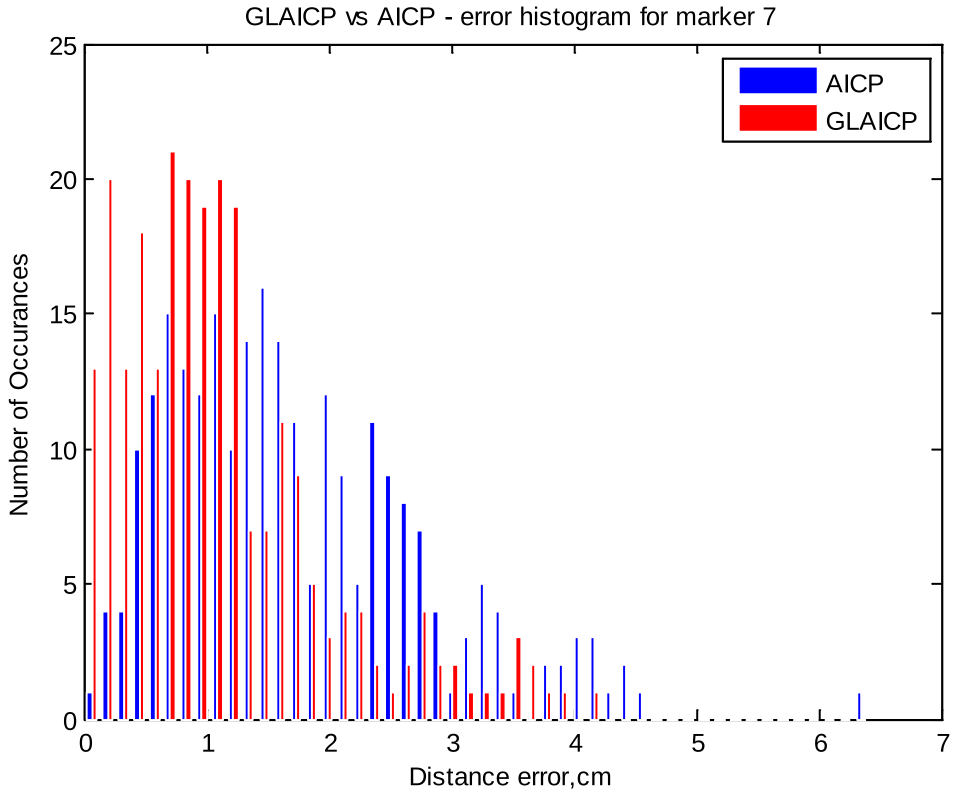

Error comparision, GLACP vs AICP for marker 5
error,cm
timeframe
AICP
GLAICP
0
1
2
3
4
5
6
0
50
100
150
200
250

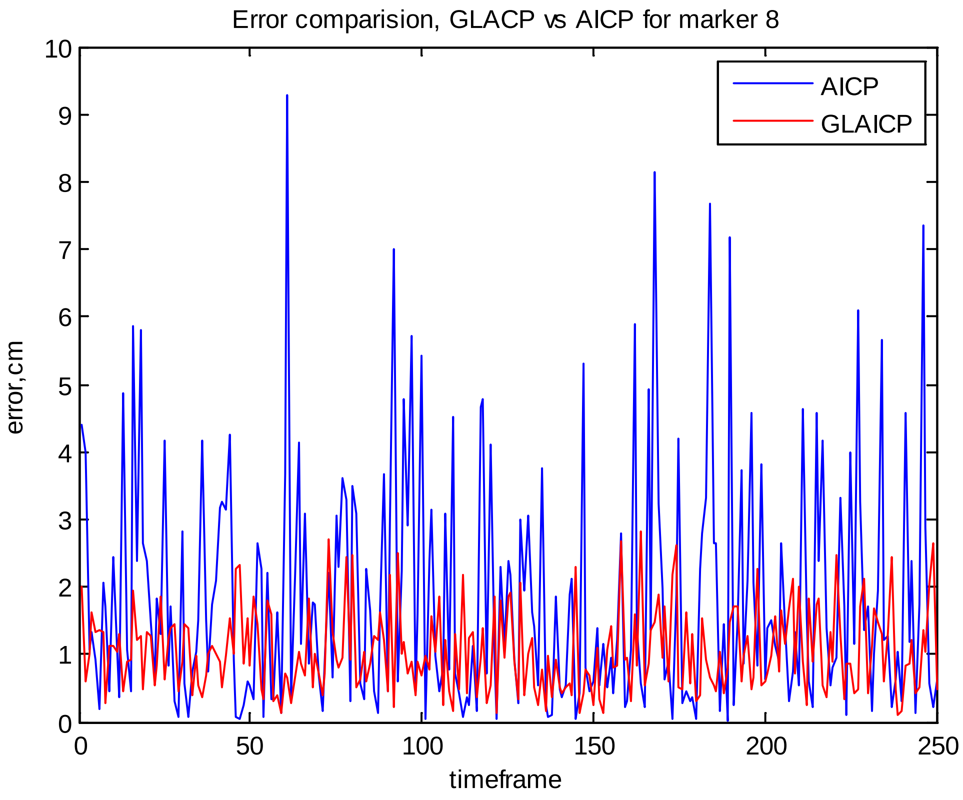

GLAICP vs AICP - error histogram for marker 5
Number of Occurances
Distance error,cm
AICP
GLAICP
0
5
10
15
20
25
30
0
1
2
3
4
5
6

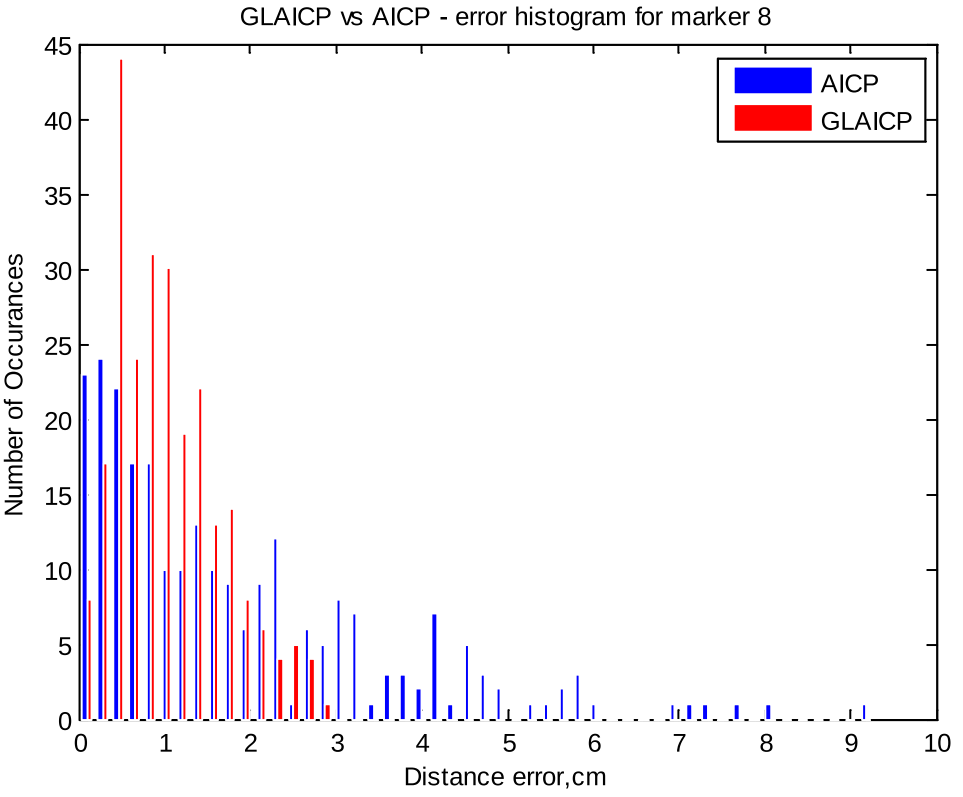

Error comparision, GLACP vs AICP for marker 6
error,cm
timeframe
AICP
GLAICP
0
1
2
3
4
5
6
7
8
9
10
0
50
100
150
200
250

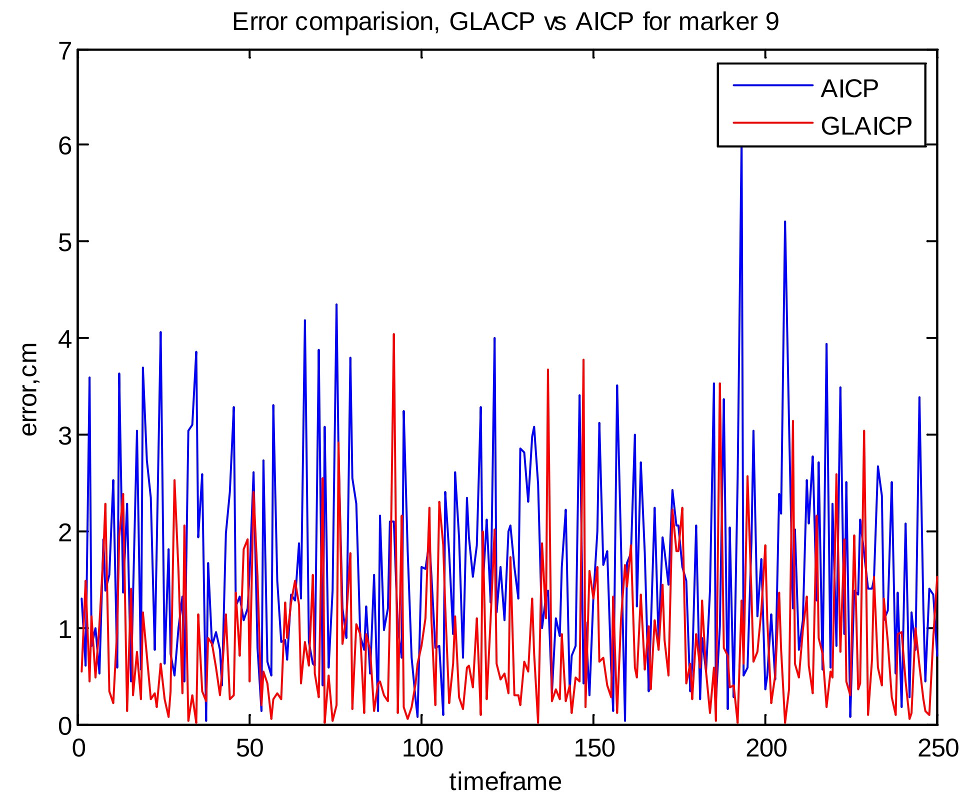

GLAICP vs AICP - error histogram for marker 6
Number of Occurances
Distance error,cm
AICP
GLAICP
0
5
10
15
20
25
30
35
0
1
2
3
4
5
6
7
8
9
10

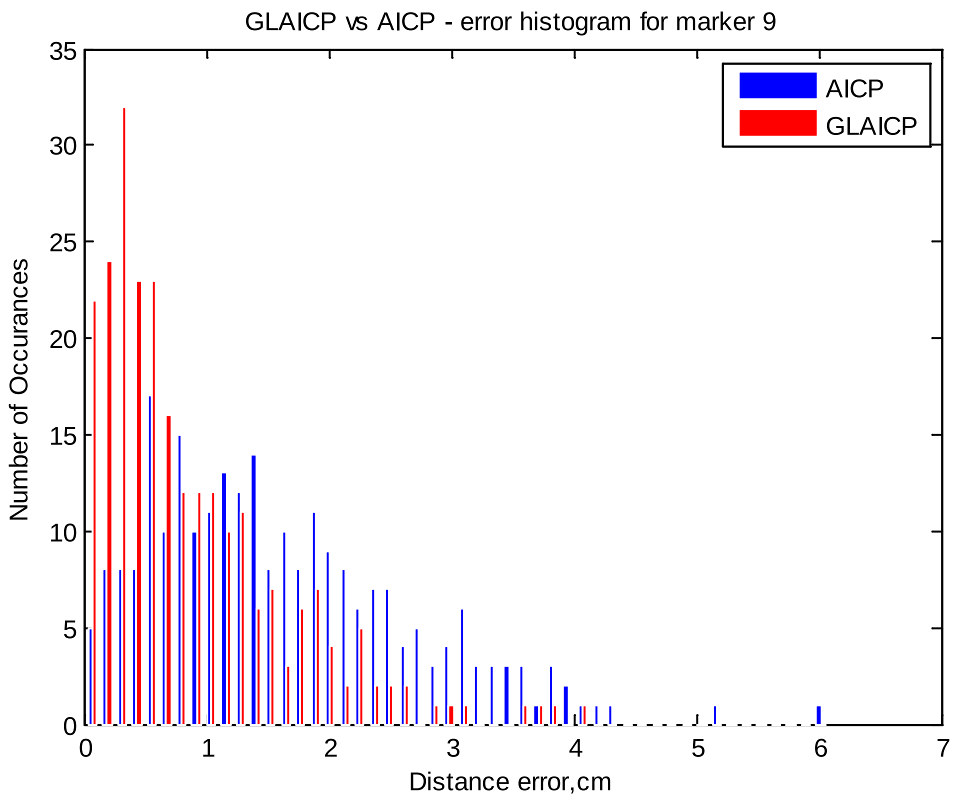

Error comparision, GLACP vs AICP for marker 7
AICP
GLAICP
error,cm
timeframe
0
1
2
3
4
5
6
7
0
50
100
150
200
250

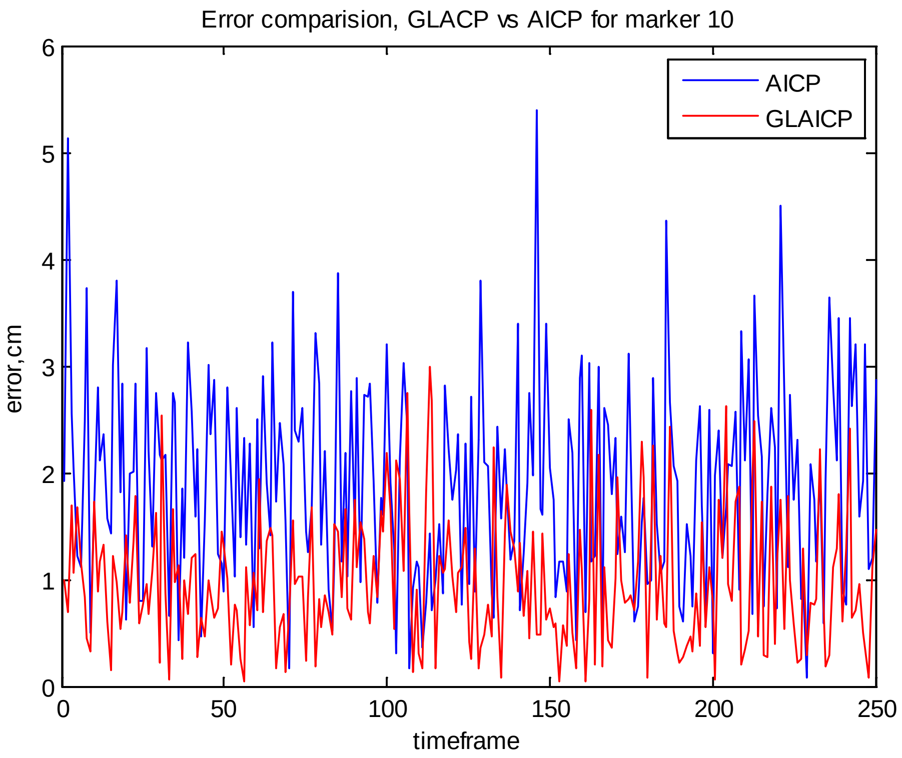

GLAICP vs AICP - error histogram for marker 7
AICP
GLAICP
Number of Occurances
Distance error,cm
0
5
10
15
20
25
0
1
2
3
4
5
6
7

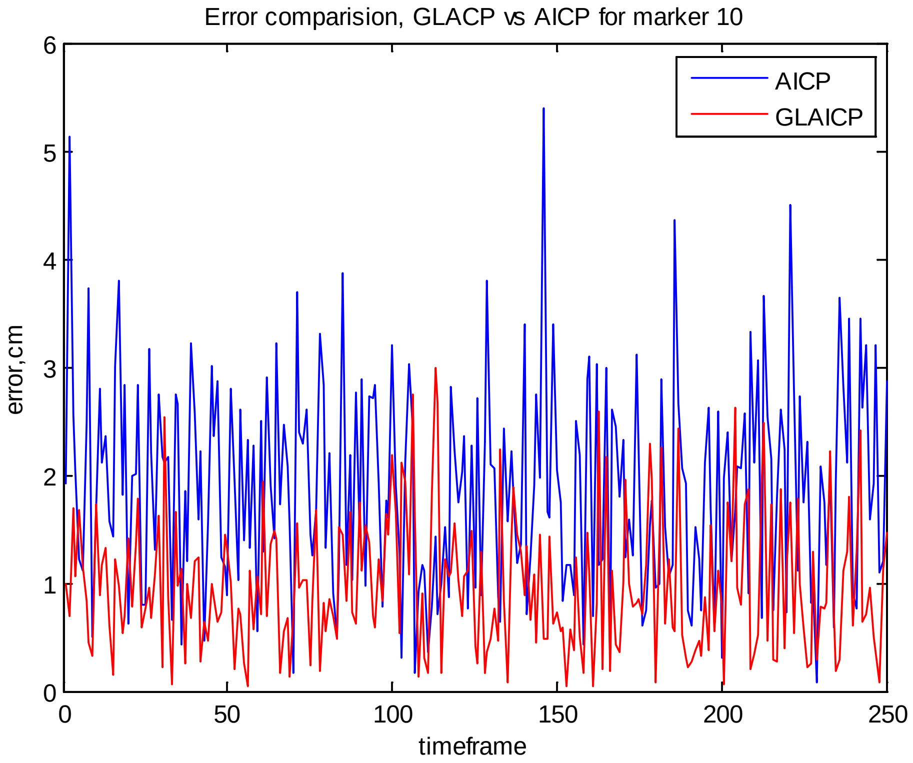

Error comparision, GLACP vs AICP for marker 8
AICP
GLAICP
error,cm
timeframe
0
1
2
3
4
5
6
7
8
9
10
0
50
100
150
200
250

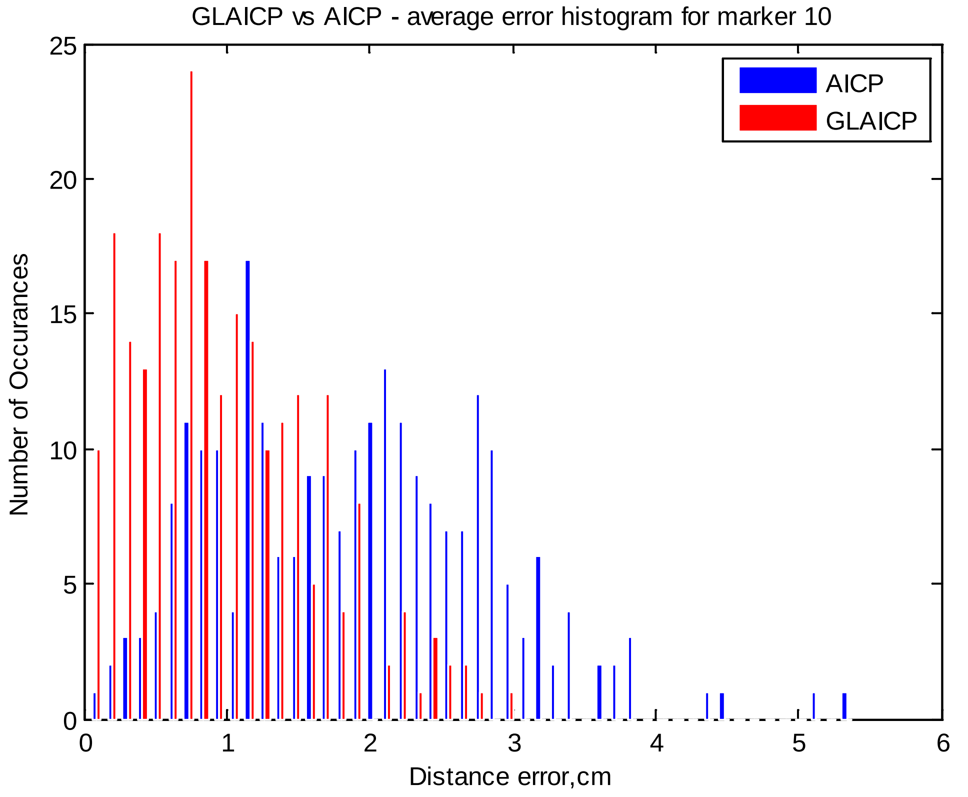

GLAICP vs AICP - error histogram for marker 8
AICP
GLAICP
Number of Occurances
Distance error,cm
0
5
10
15
20
25
30
35
40
45
0
1
2
3
4
5
6
7
8
9
10

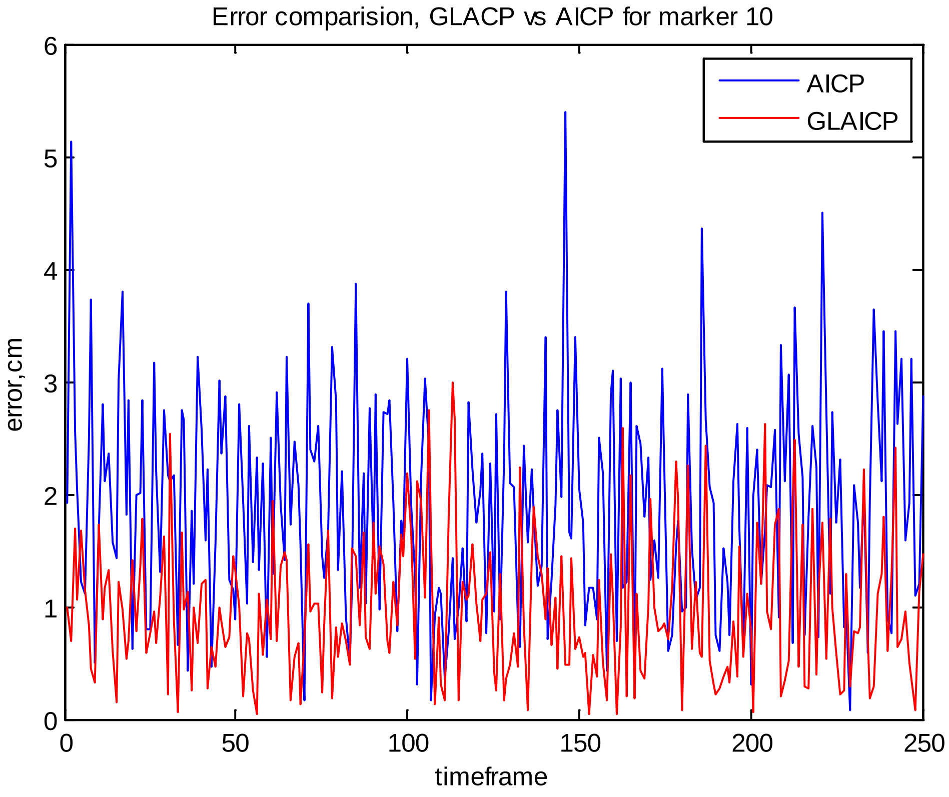

Error comparision, GLACP vs AICP for marker 9
AICP
GLAICP
error,cm
timeframe
0
1
2
3
4
5
6
7
0
50
100
150
200
250

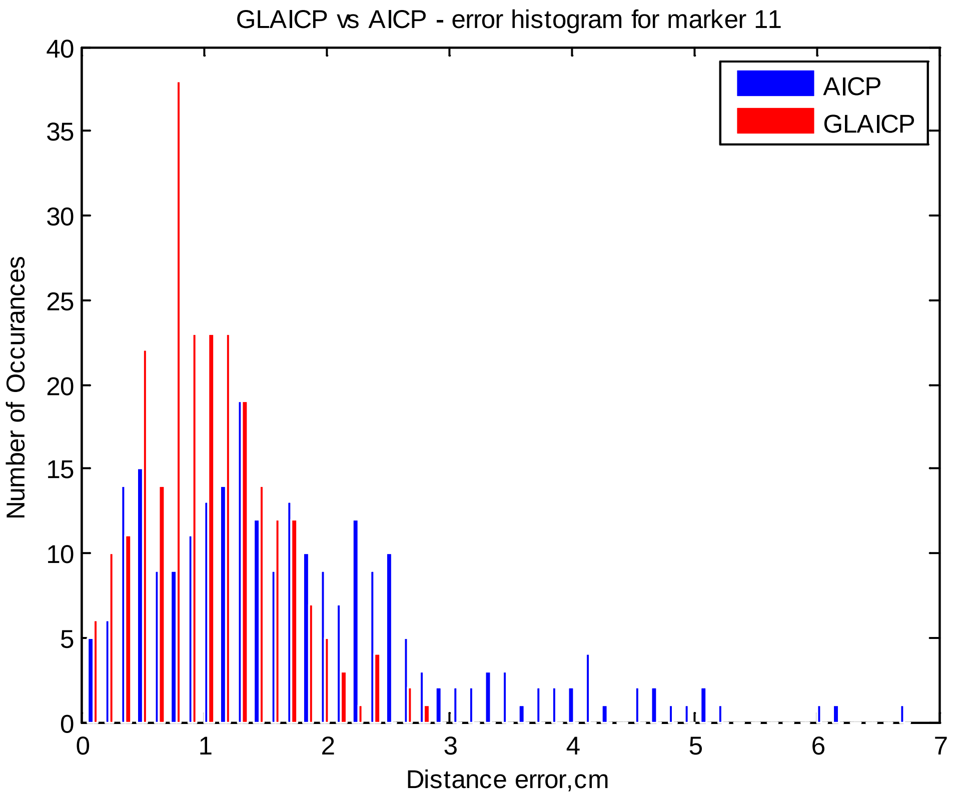

GLAICP vs AICP - error histogram for marker 9
AICP
GLAICP
Number of Occurances
Distance error,cm
0
5
10
15
20
25
30
35
0
1
2
3
4
5
6
7

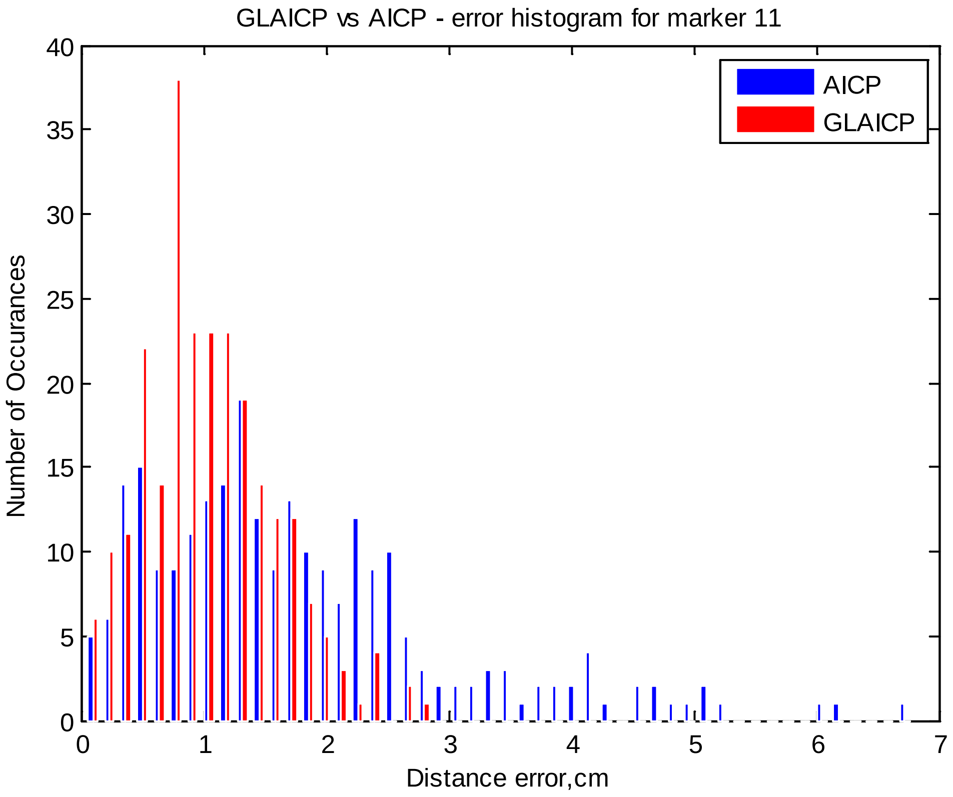

Error comparision, GLACP vs AICP for marker 10
AICP
GLAICP
error,cm
timeframe
0
1
2
3
4
5
6
0
50
100
150
200
250

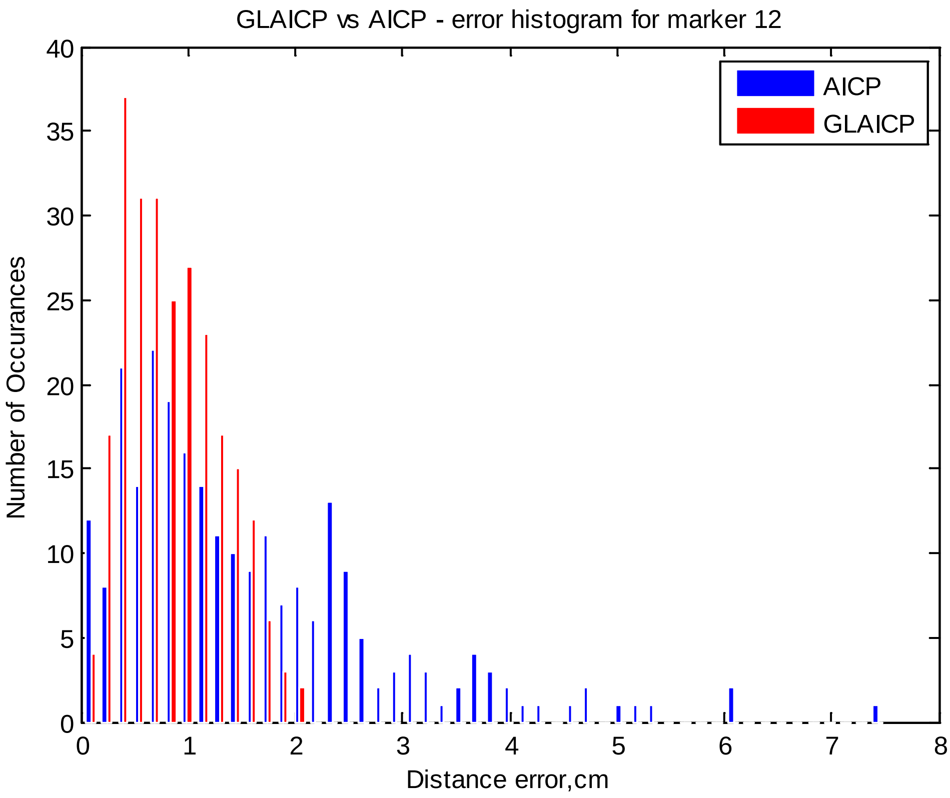

Error comparision, GLACP vs AICP for marker 10
AICP
GLAICP
error,cm
timeframe
0
1
2
3
4
5
6
0
50
100
150
200
250


GLAICP vs AICP - average error histogram for marker 10
AICP
GLAICP
Number of Occurances
Distance error,cm
0
5
10
15
20
25
0
1
2
3
4
5
6


Error comparision, GLACP vs AICP for marker 10
AICP
GLAICP
error,cm
timeframe
0
1
2
3
4
5
6
0
50
100
150
200
250

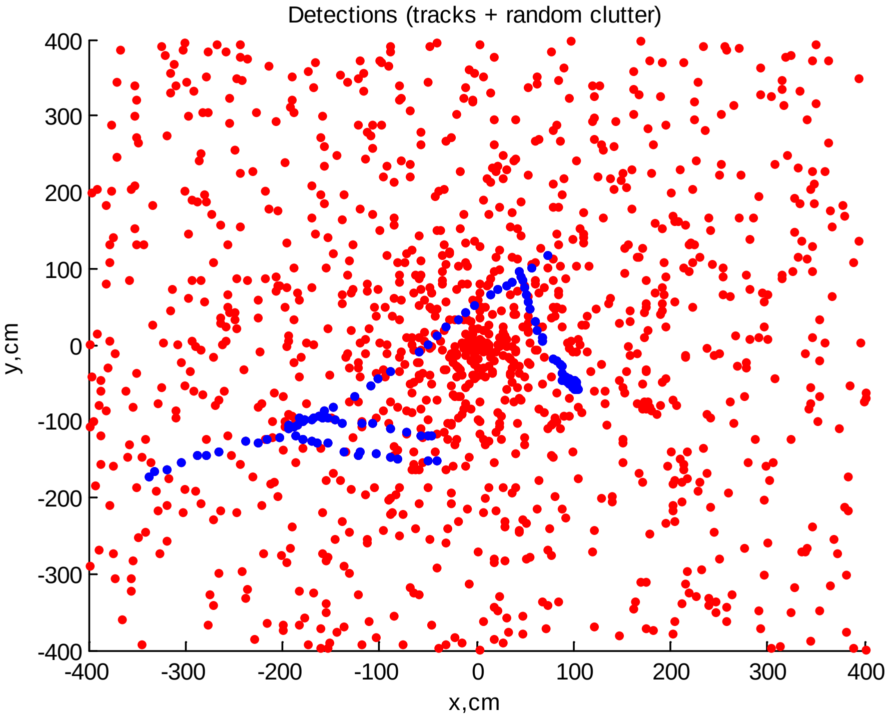

GLAICP vs AICP - error histogram for marker 11
Number of Occurances
Distance error,cm
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5
6
7
🟦 AICP
🟥 GLAICP

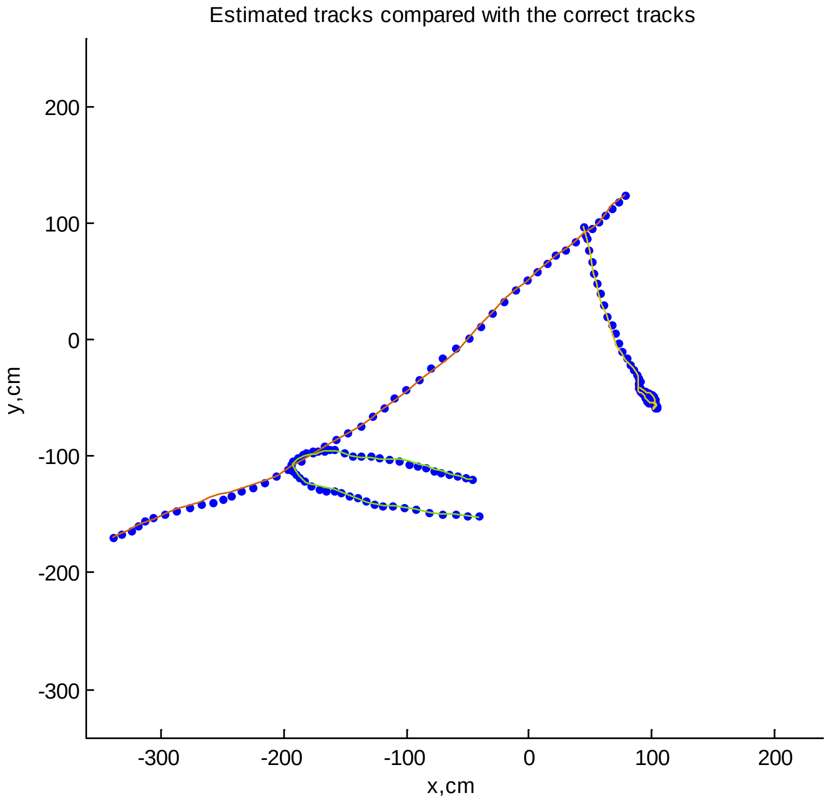

GLAICP vs AICP - error histogram for marker 11
Number of Occurances
Distance error,cm
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5
6
7
🟦 AICP
🟥 GLAICP

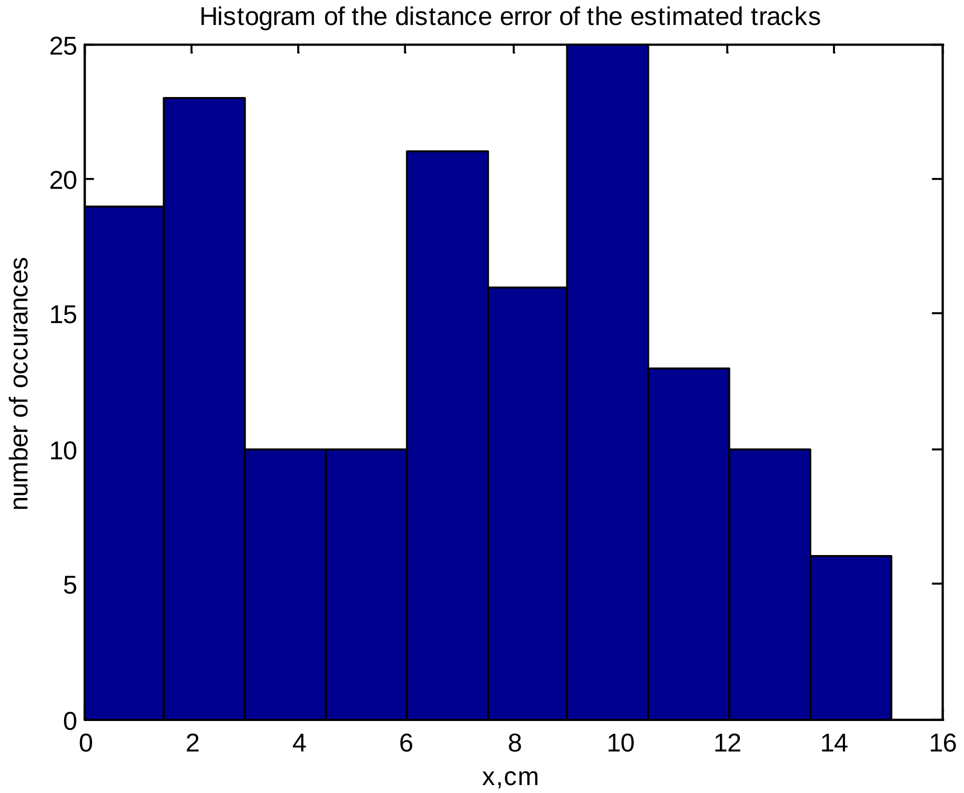

GLAICP vs AICP - error histogram for marker 12
AICP
GLAICP
Number of Occurances
Distance error,cm
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5
6
7
8

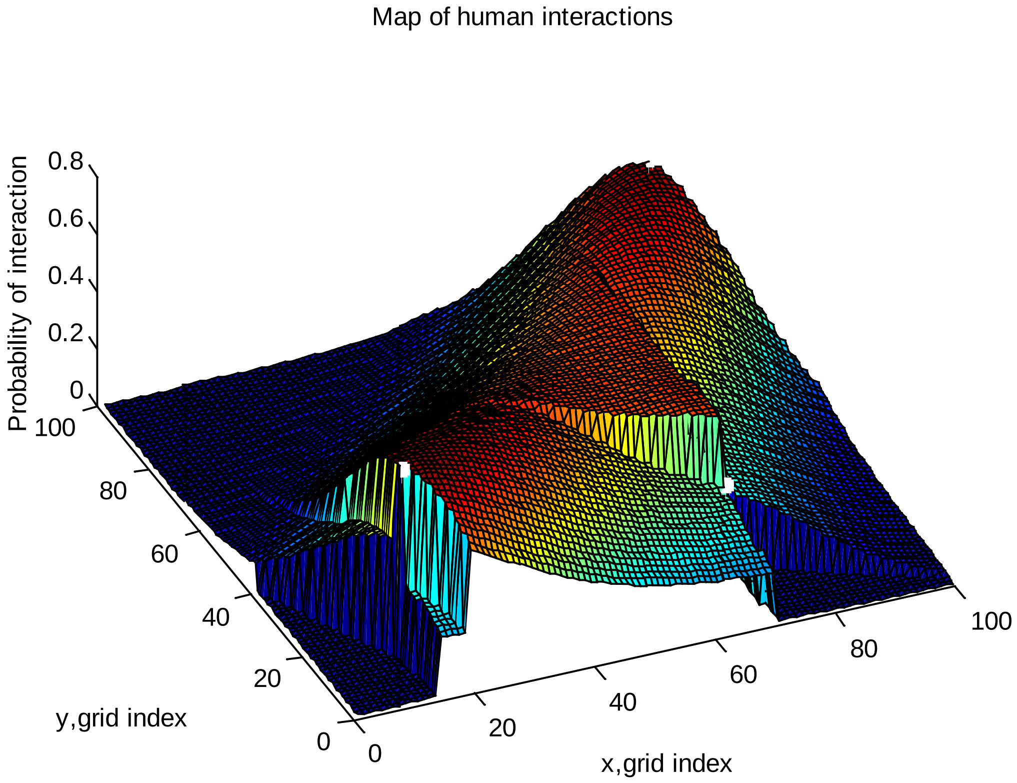
Appendix H: Multiple human tracking results

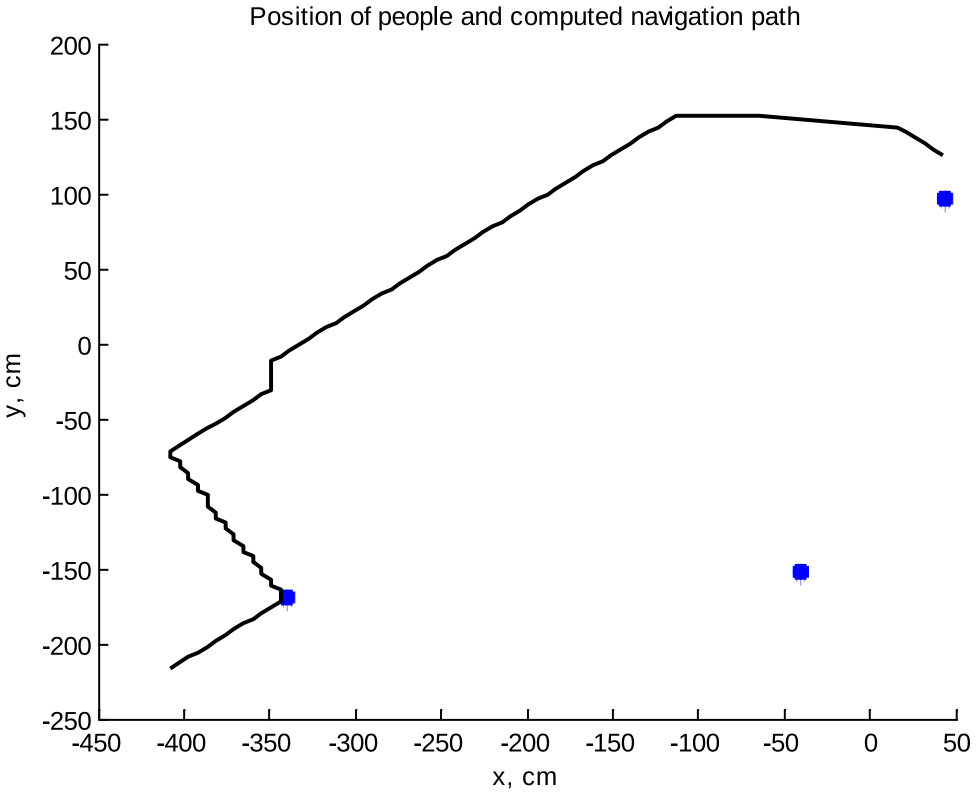
Experiment One - Three human tracks



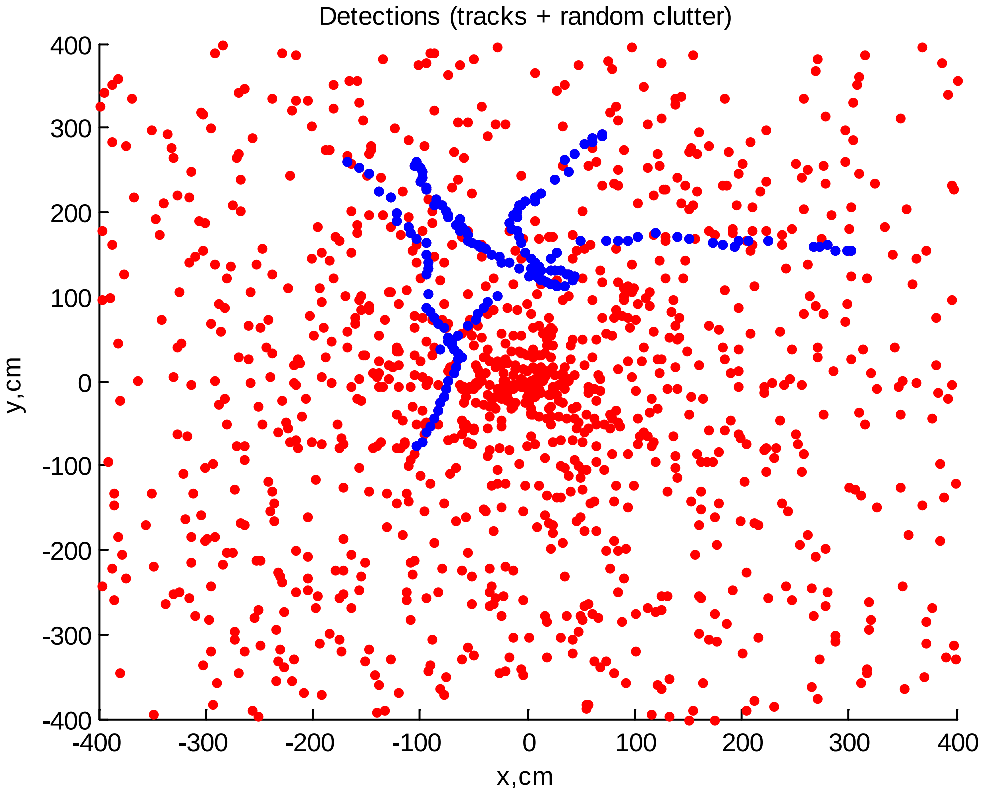


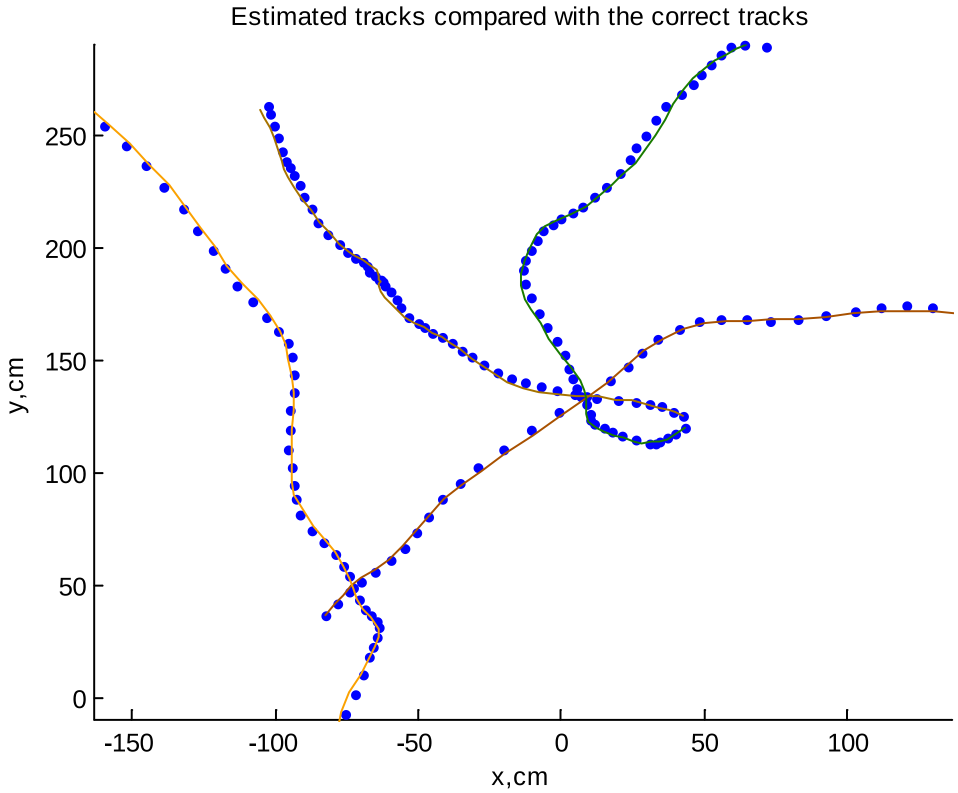

Histogram of the distance error of the estimated tracks
number of occurances
0
5
10
15
20
25
0
2
4
6
8
10
12
14
16
x,cm

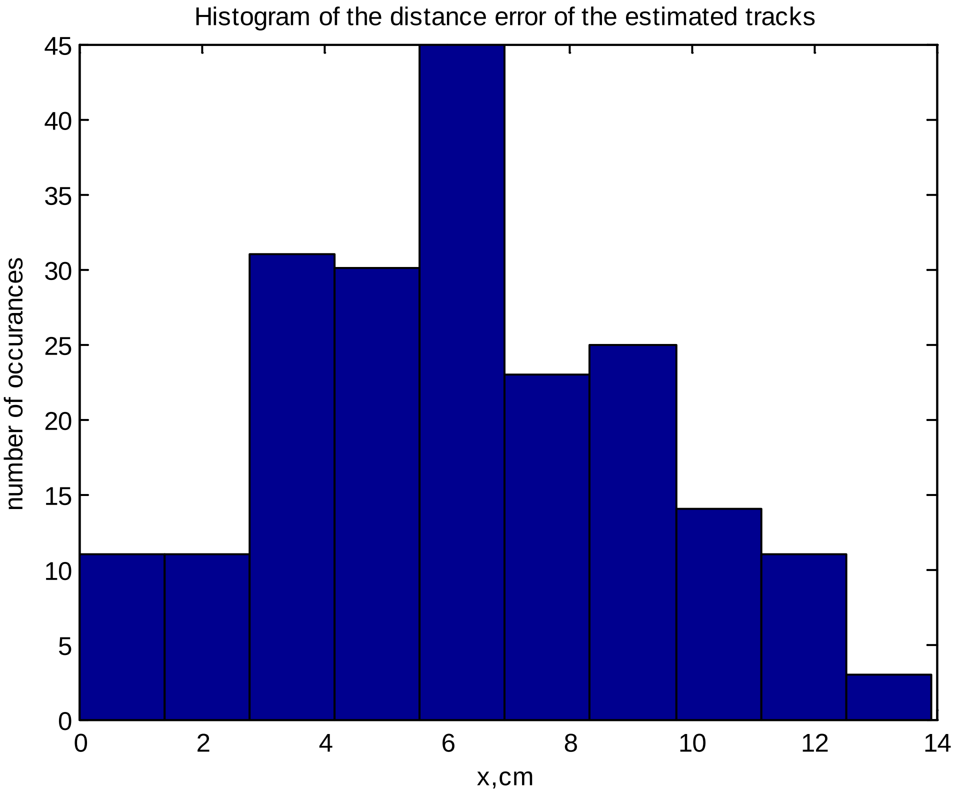

Map of human interactions
Probability of interaction
0
0.2
0.4
0.6
0.8
y,grid index
0
20
40
60
80
100
x,grid index
0
20
40
60
80
100



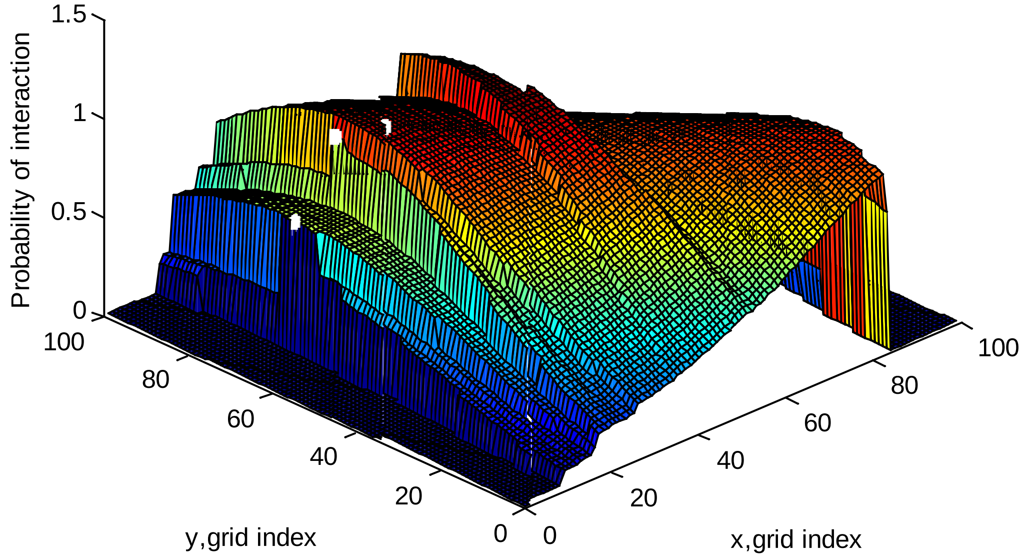
Experiment Two - Four human tracks



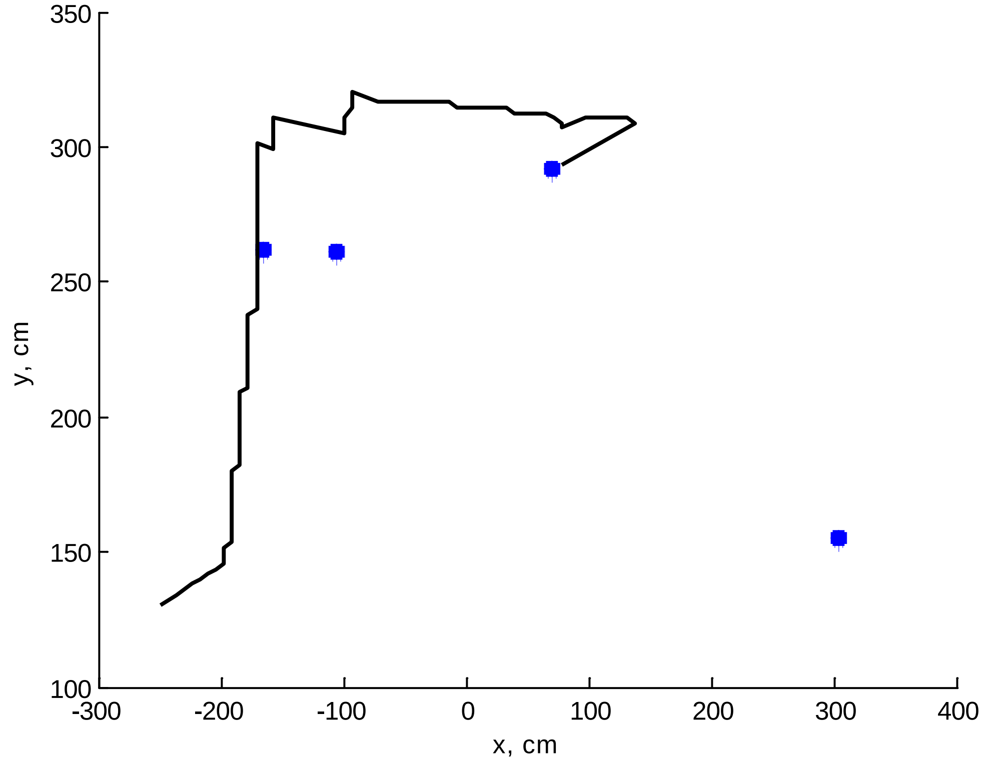

Estimated tracks compared with the correct tracks
y,cm
250
200
150
100
50
0
-150
-100
-50
0
50
100
x,cm


Histogram of the distance error of the estimated tracks
number of occurances
45
40
35
30
25
20
15
10
5
0
0
2
4
6
8
10
12
14
x,cm

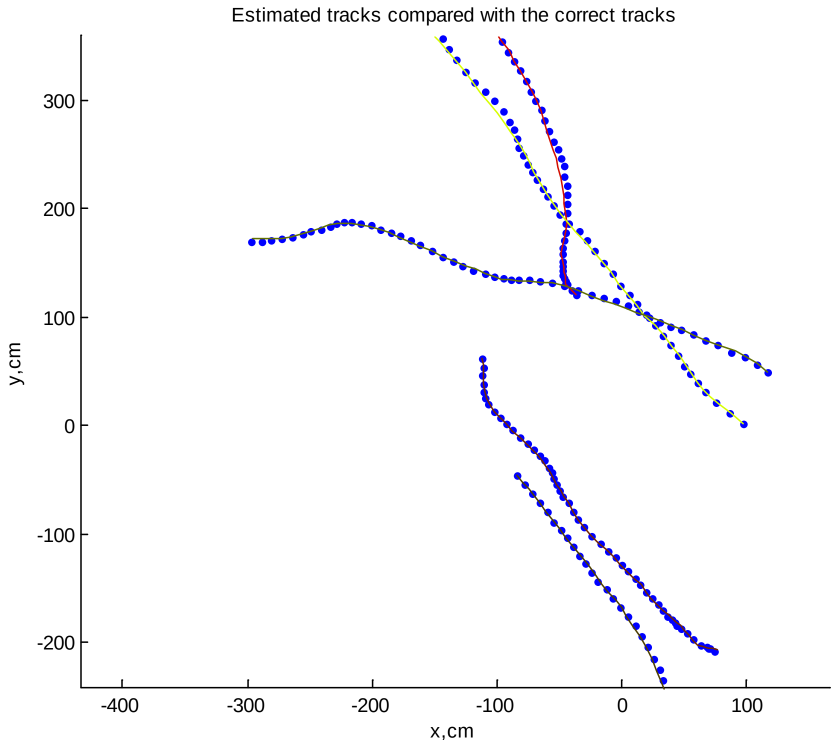
Map of human interactions

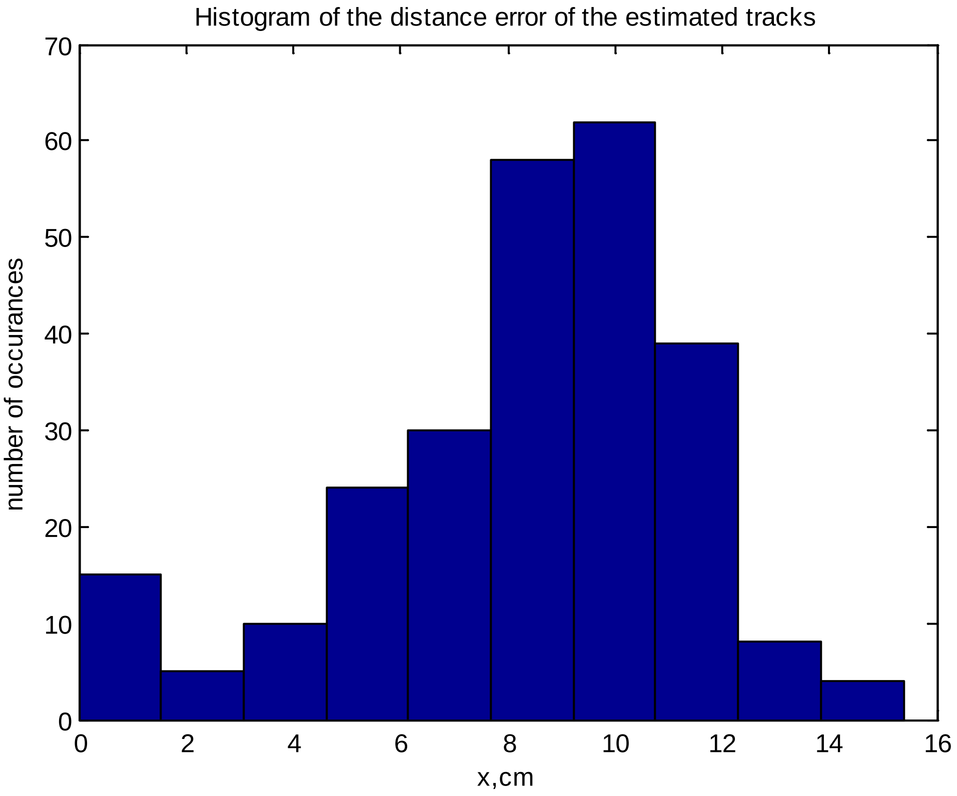


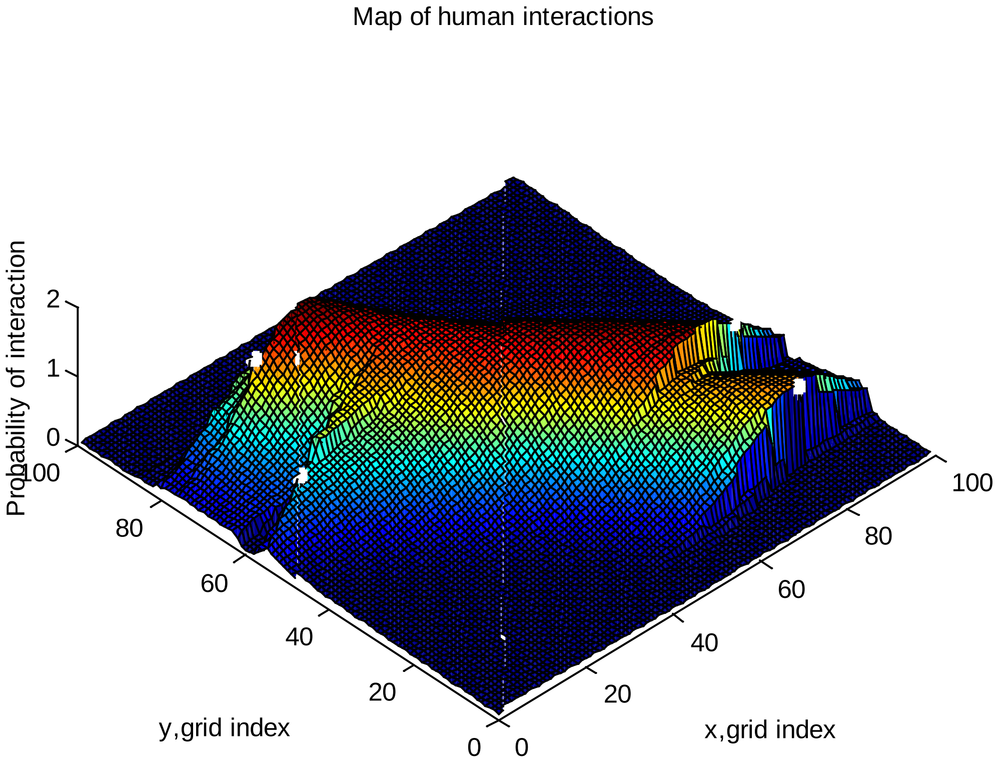
Position of people and computed navigation path

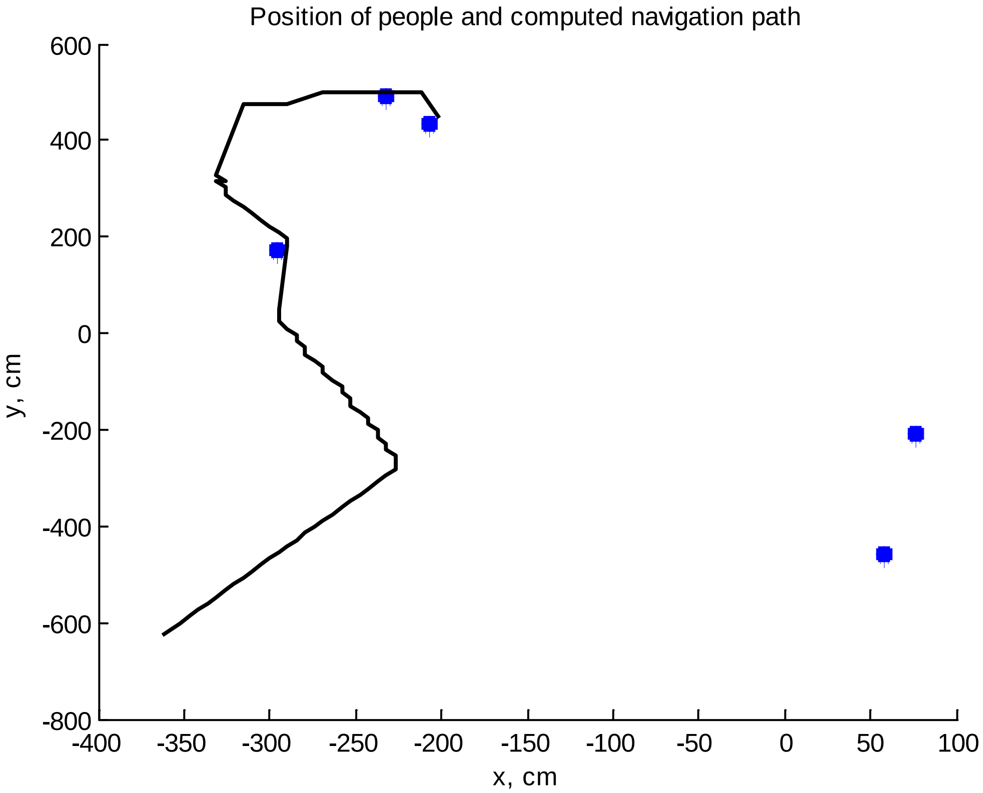


Experiment Three - Five human tracks

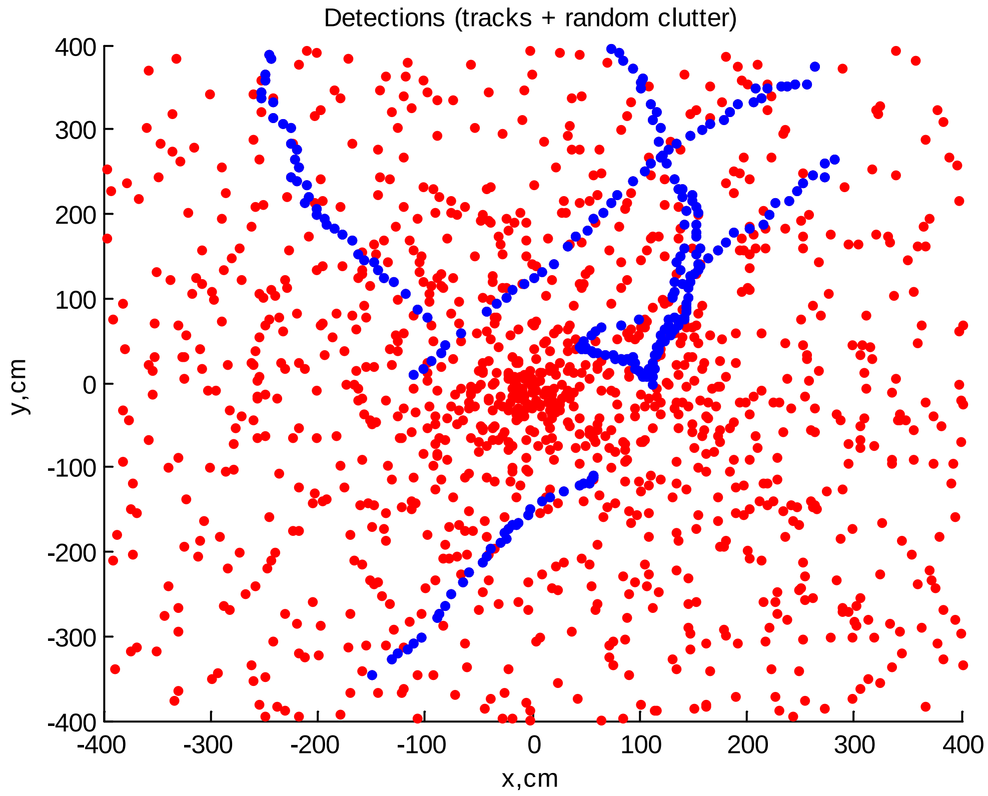


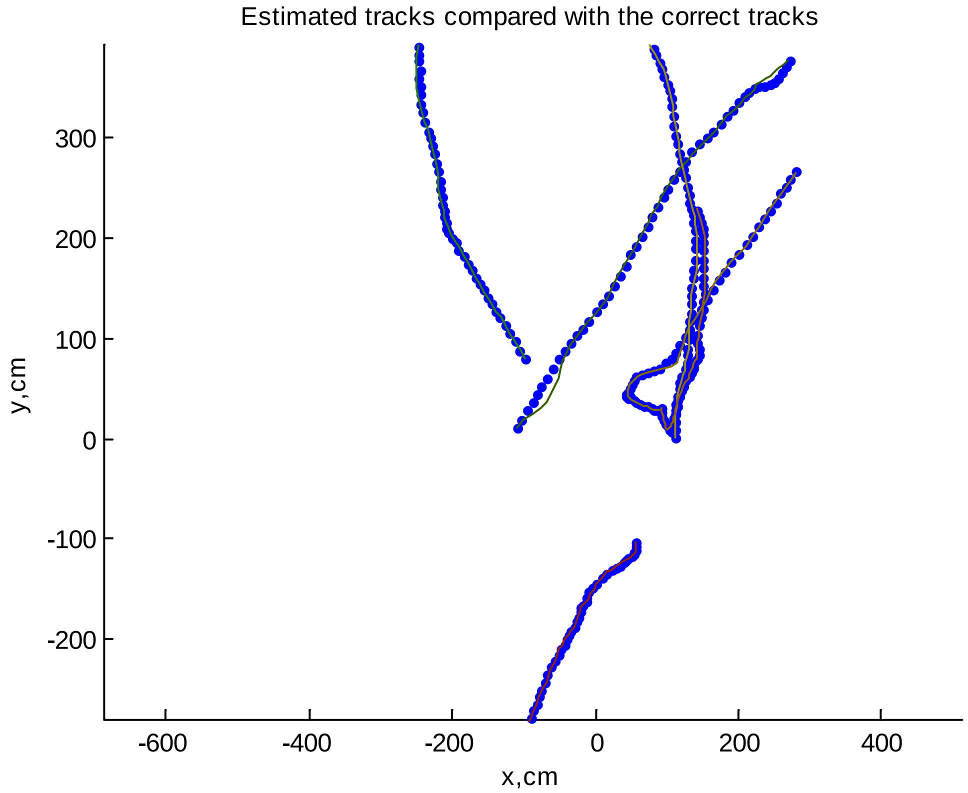


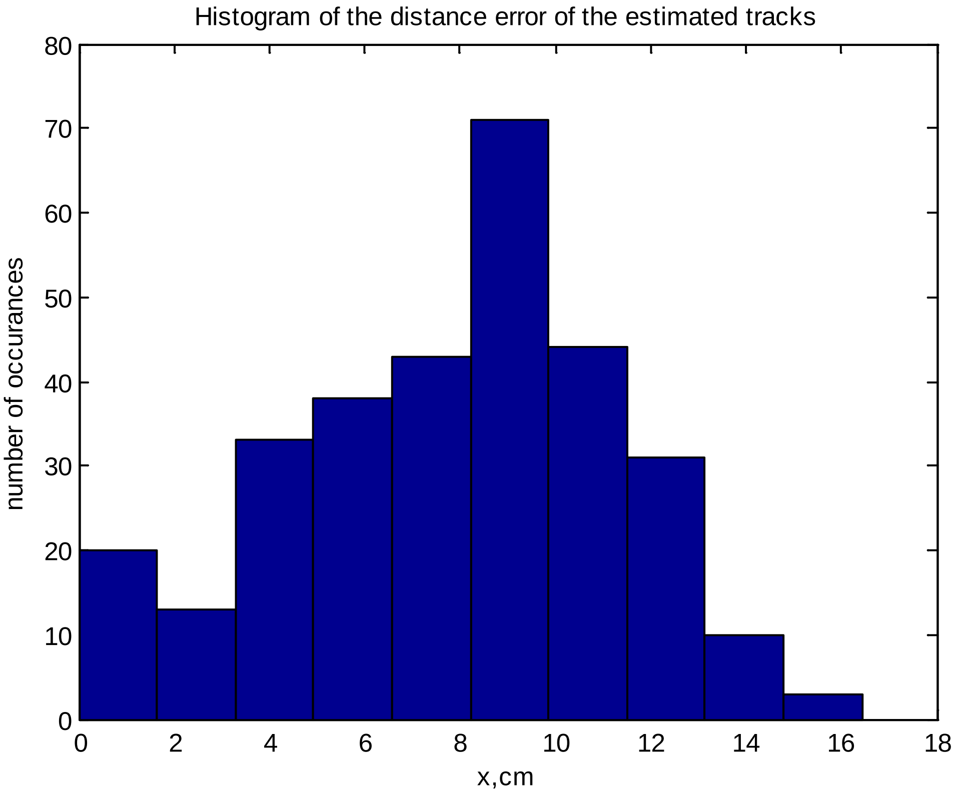

Map of human interactions
Probability of interaction
2
1
0
100
80
60
40
20
0
y,grid index
0
20
40
60
80
100
x,grid index

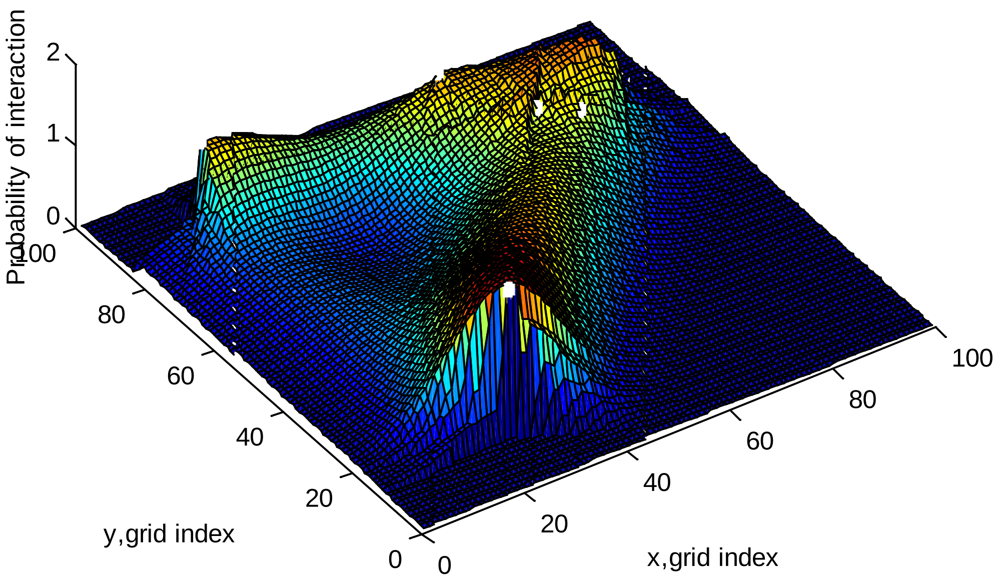

Position of people and computed navigation path
600
400
200
0
-200
-400
-600
-800
y, cm
-400
-350
-300
-250
-200
-150
-100
-50
0
50
100
x, cm

Experiment Four - Six human tracks

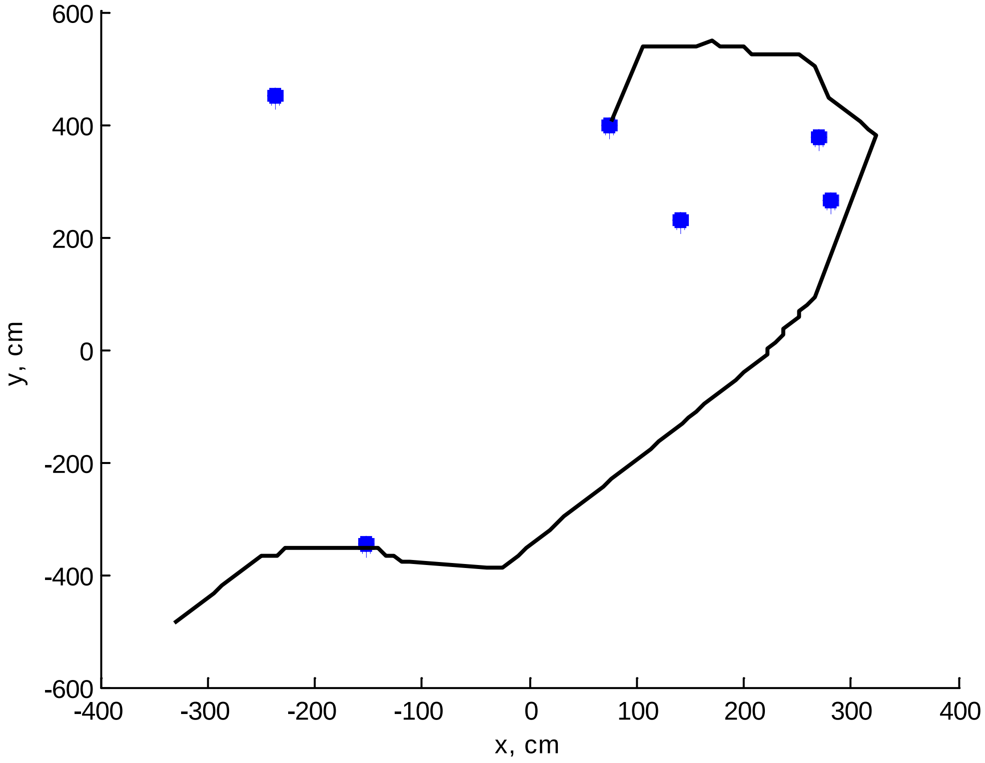



Estimated tracks compared with the correct tracks
300
200
100
0
-100
-200
y,cm
-600
-400
-200
0
200
400
x,cm


Histogram of the distance error of the estimated tracks
80
70
60
50
40
30
20
10
0
number of occurances
0
2
4
6
8
10
12
14
16
18
x,cm

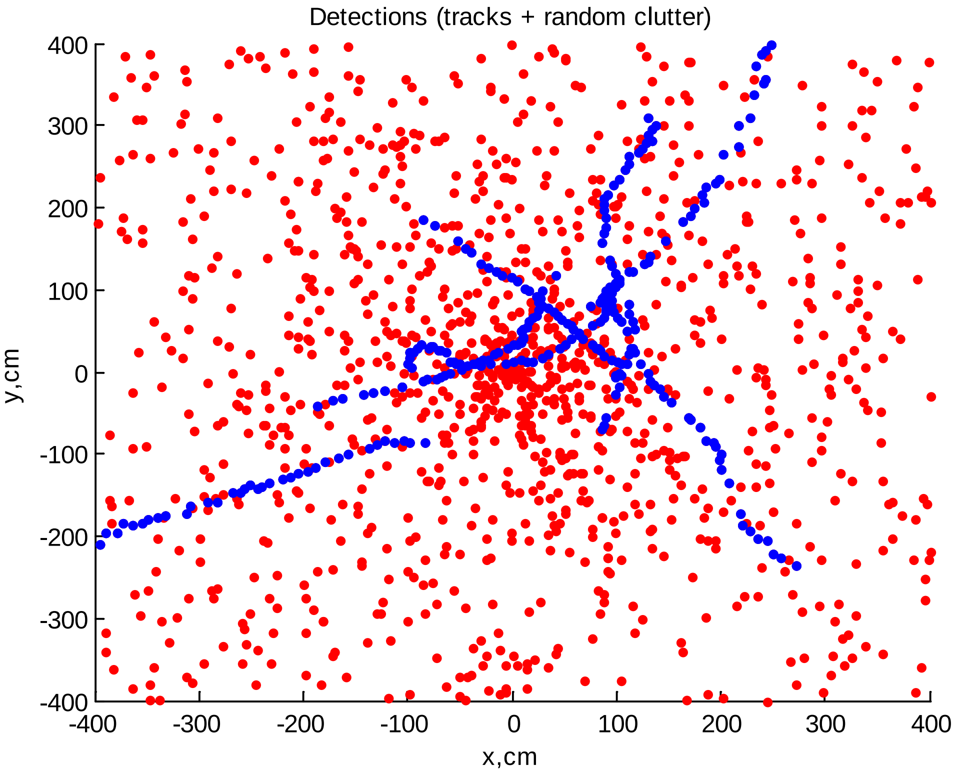


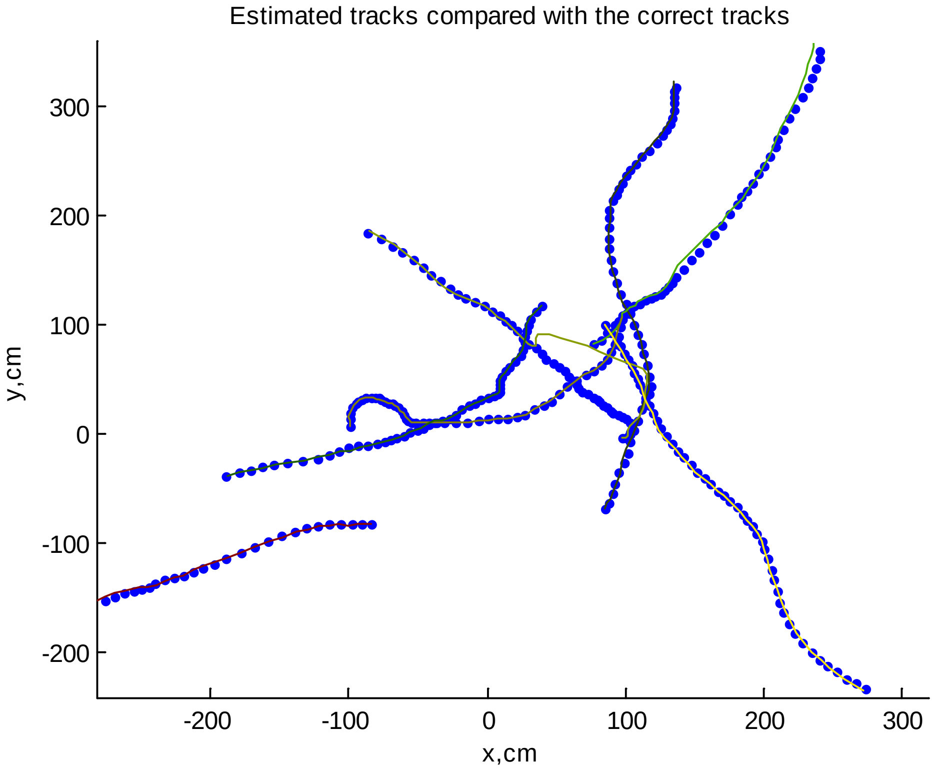
Map of human interactions

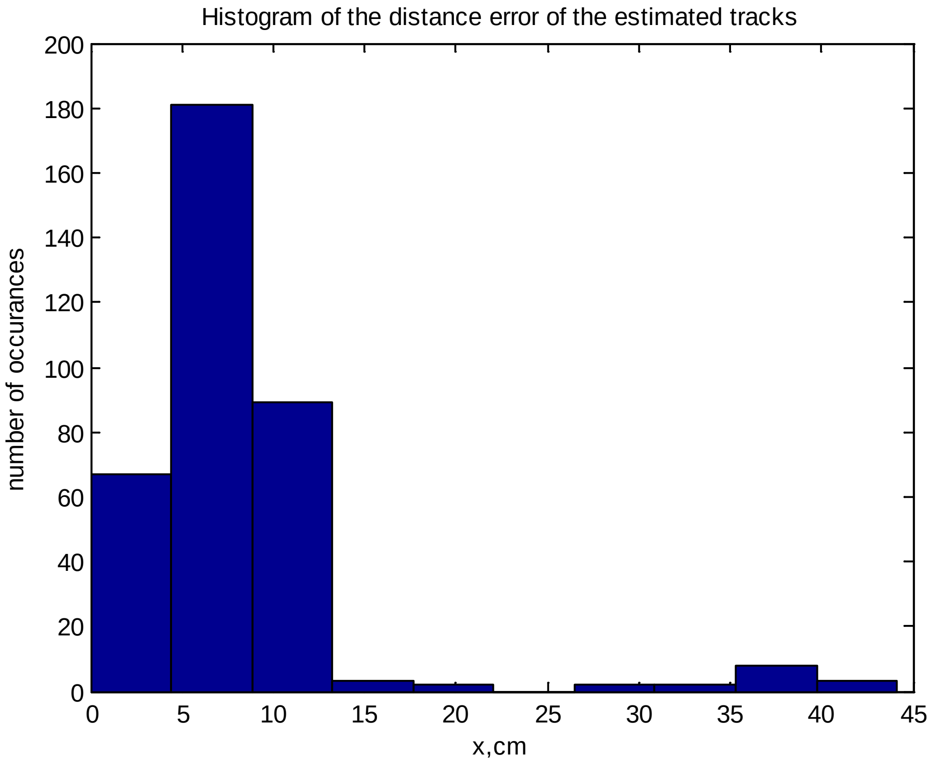


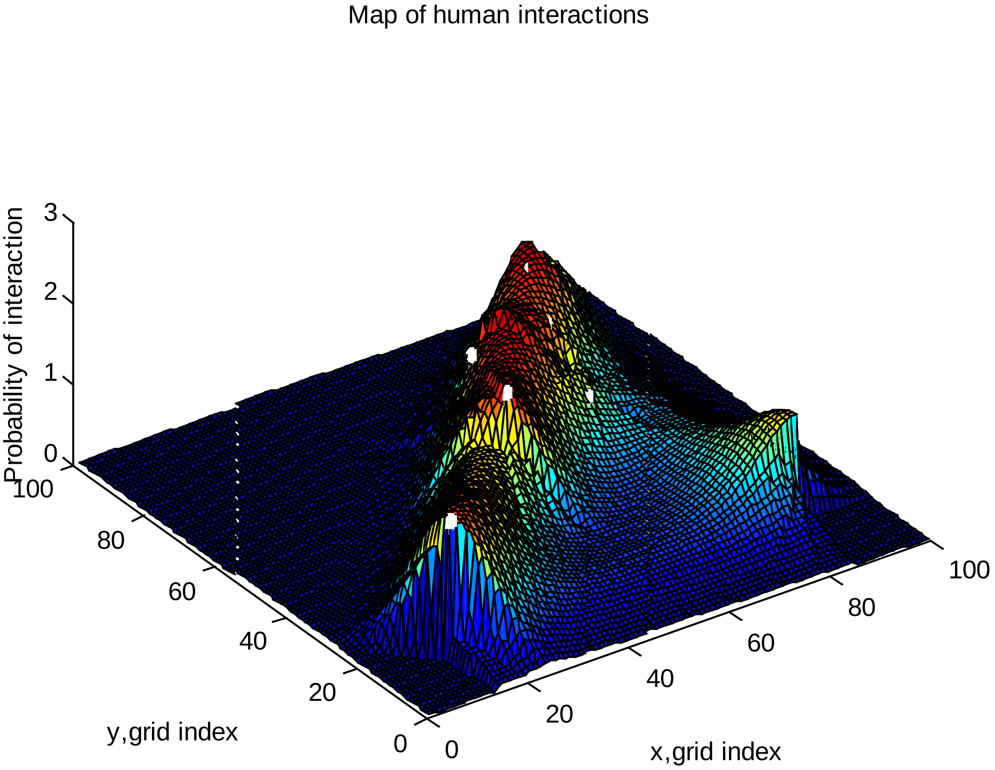
Position of people and computed navigation path

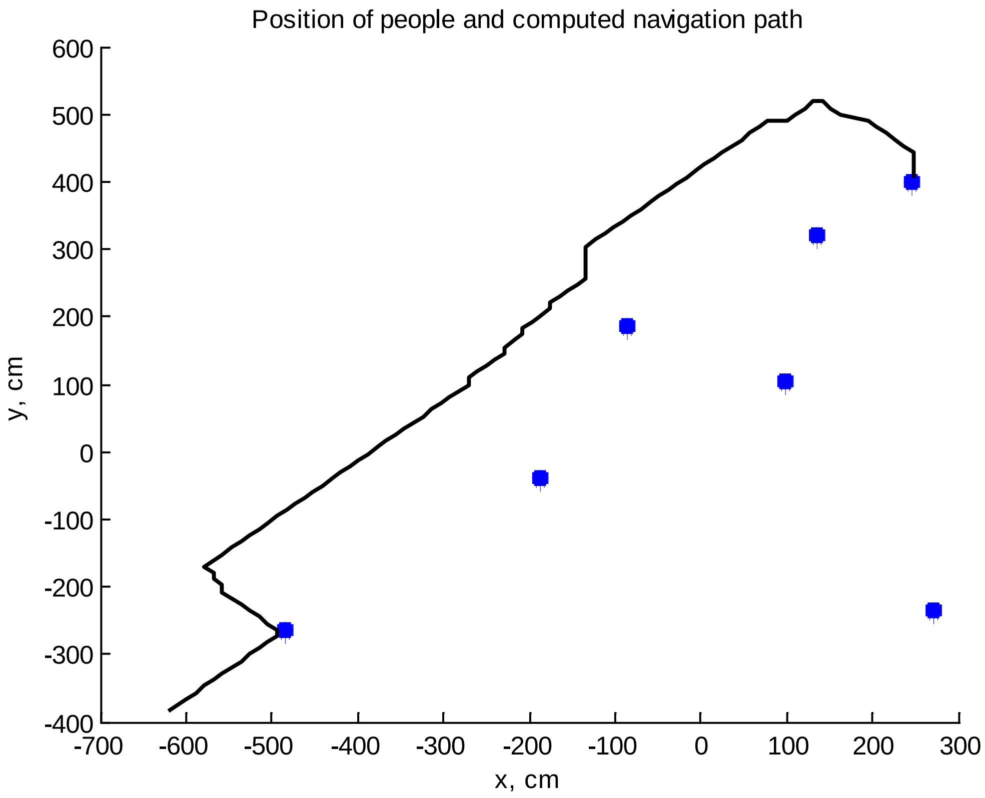
Experiment Five - Seven human tracks



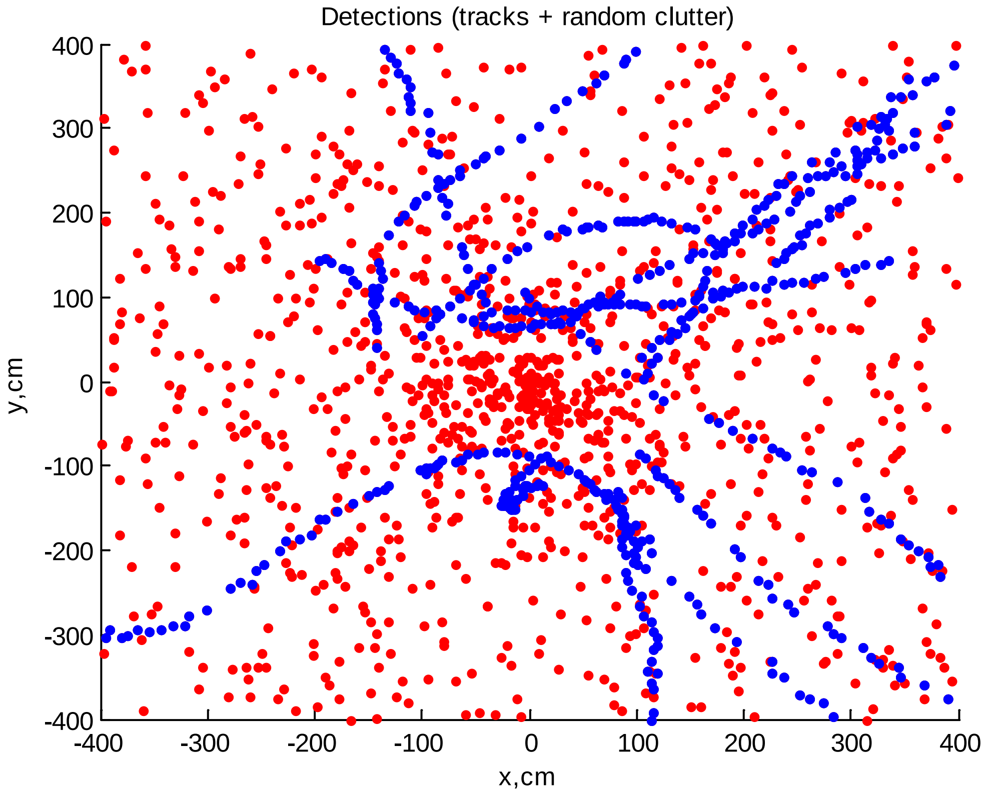


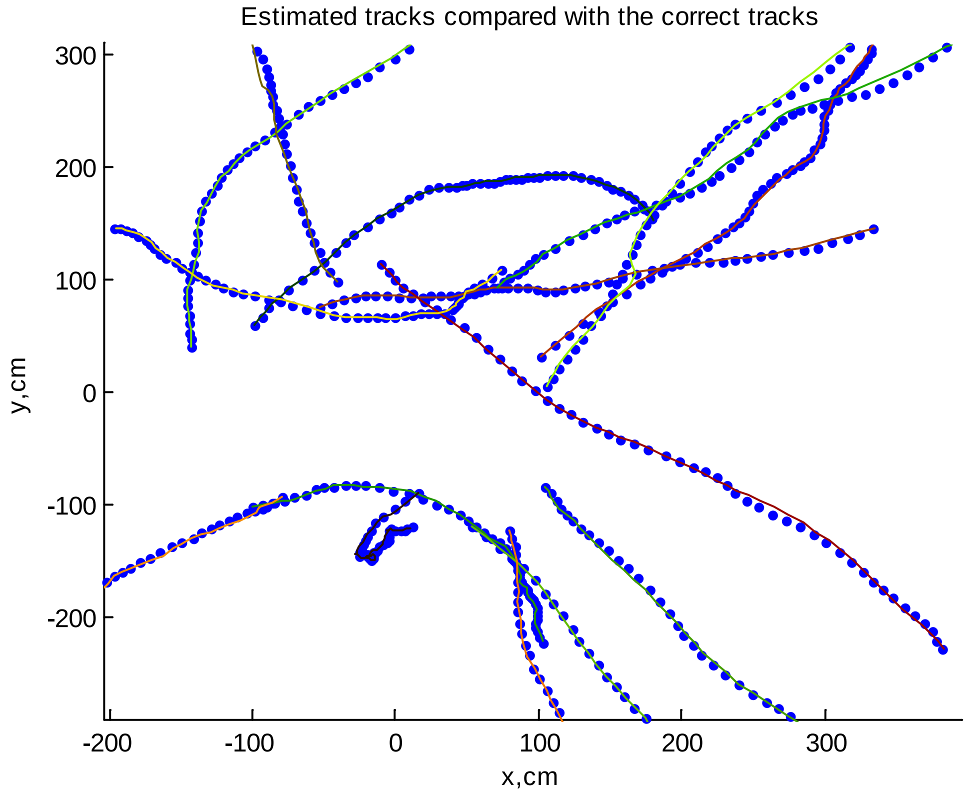

Histogram of the distance error of the estimated tracks
number of occurances
200
180
160
140
120
100
80
60
40
20
0
0
5
10
15
20
25
30
35
40
45
x,cm

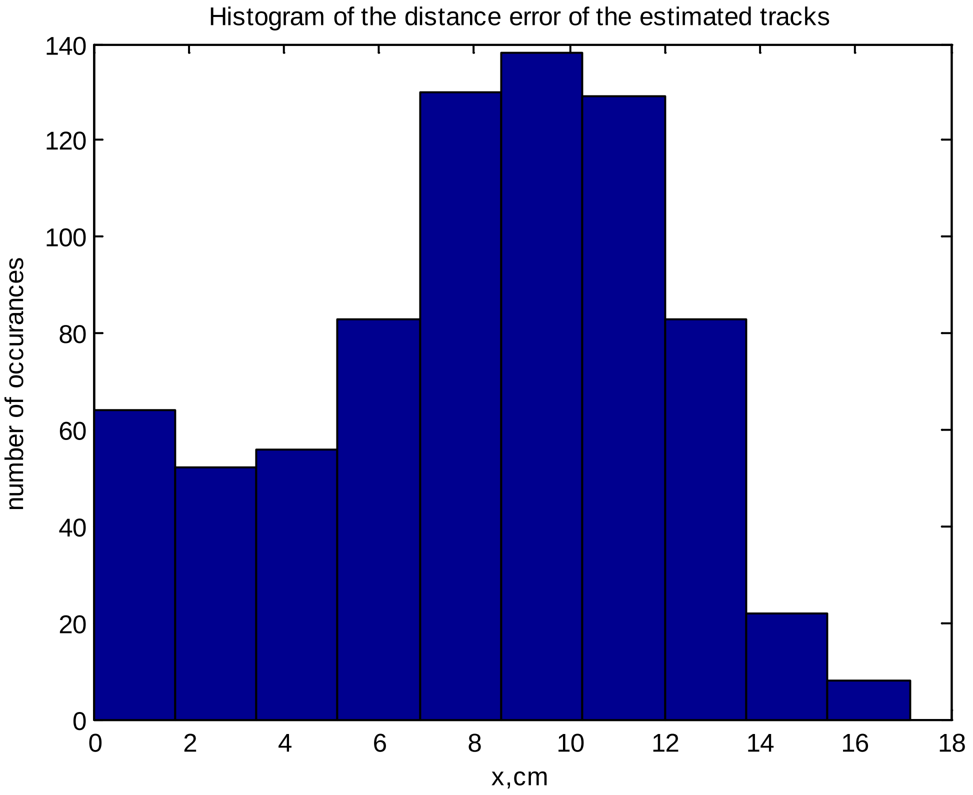

Map of human interactions
Probability of interaction
3
2
1
0
100
80
60
40
20
0
y,grid index
0
20
40
60
80
100
x,grid index

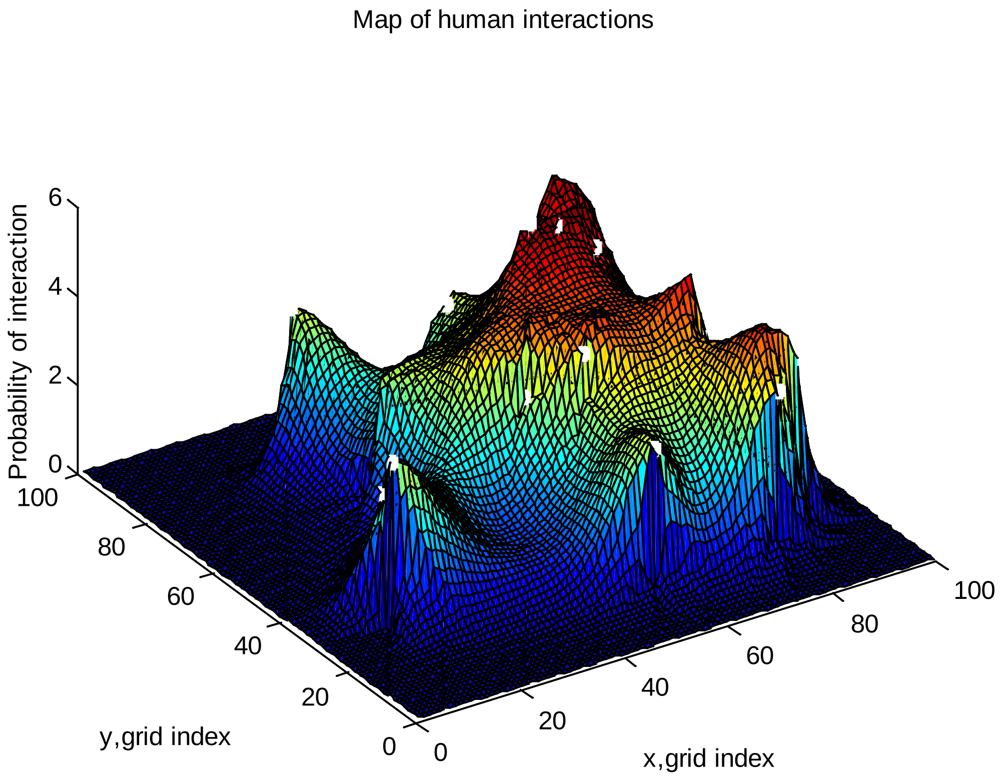


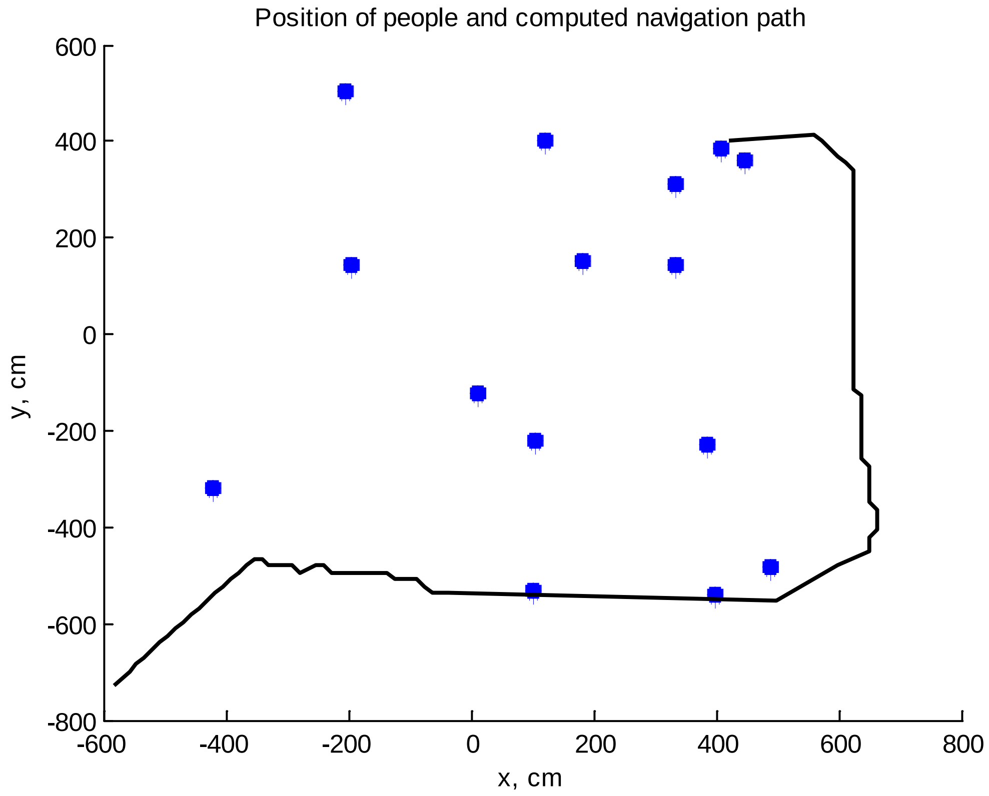
Experiment Six - Fifteen human tracks




Estimated tracks compared with the correct tracks
300
200
100
0
-100
-200
y,cm
-200
-100
0
100
200
300
x,cm


Histogram of the distance error of the estimated tracks
140
120
100
80
60
40
20
0
number of occurances
0
2
4
6
8
10
12
14
16
18
x,cm


Map of human interactions
Probability of interaction
6
4
2
0
100
80
60
40
20
0
y,grid index
0
20
40
60
80
100
x,grid index


Position of people and computed navigation path
y, cm
600
400
200
0
-200
-400
-600
-800
-600
-400
-200
0
200
400
600
800
x, cm