People Tracking Under Social Constraints

Matthias Luber

Technische Fakultät
Albert-Ludwigs-Universität Freiburg im Breisgau

Dissertation zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften

Betreuer: Kai O. Arras

2014

People Tracking Under Social Constraints

Matthias Luber

Dissertation zur Erlangung des akademischen Grades Doktor der Naturwissenschaften
Technische Fakultät, Albert-Ludwigs-Universität Freiburg im Breisgau

Dekan:	Prof. Dr. Yiannos Manoli
Erstgutachter:	Prof. Dr. Kai O. Arras
Zweitgutachter:	Prof. Dr. Thomas Brox

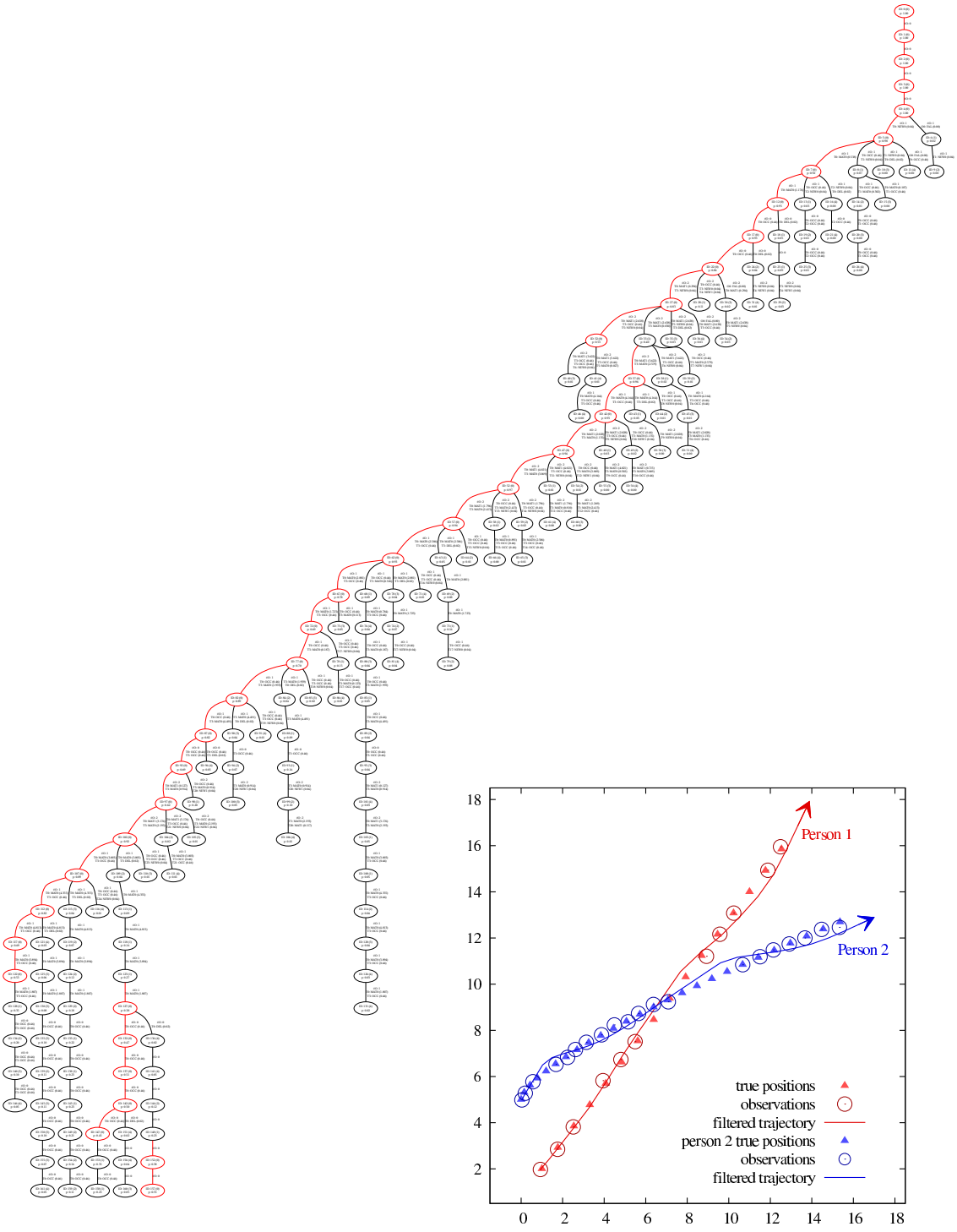


Figure 1: Visualization of the hypotheses tree with $k = 5$ hypotheses at each point in time of an example scene with two tracked persons. (*Tree*:) The nodes in the tree denote the hypotheses with different track to observation assignments, thus the various paths through the tree represent different explanations of how the state of the persons evolves over time. The best hypothesis at each time step is marked in red. (*Plot*:) The plot shows the real positions (rectangles), observed positions (circles), and filtered trajectories (lines) of the two persons.

Für Alexandra und Carlotta

Abstract

A major research topic in human oriented robotics is the development of state-of-the-art methods that enable robots to operate in crowded human environments. Moreover, human-robot interaction and hand-in-hand human-robot cooperation have also become a field of increasing activity. A fundamental requirement for safe, efficient, and socially acceptable robot behavior under social constraints is detailed knowledge about the presence, motion states, and goals of the surrounding individuals. This makes reliable people detection and tracking key technologies for robots.

The main components of tracking are detection, motion prediction, and data association. During the last decades extensive research has led to many approaches for detecting humans in different sensory modalities. Learning or modeling environment-specific goals yield refined motion predictions and advanced probabilistic techniques tackled the data association problem. However, the developed methods barely consider that human coexistence is based on (unwritten) social rules and normative behavior. People share social relations and respect the needs and desires of each other. So that accurate, robust, and reliable people detect and tracking has to account for such rules.

In this thesis, various advanced methods to model, learn, and integrate *spatial*, *temporal*, and especially *social* information into people detection, human motion prediction, and probabilistic data association are proposed. The investigated methods are derived from human behavior taking unwritten social constraints into account. To the best of my knowledge, this results into the most reliable and accurate people detection and tracking approaches in 2D range data. Learning from human observations, directly, allows to encode social rules using sound and common mathematical frameworks. This enables to track individuals and groups of people, simultaneously. Employing on-line learning to model target appearances and support detection the first reliable and accurate approach on 3D people detection and tracking in RGB-D data is proposed.

In more detail, spatio-temporal information supports people detection in capturing the huge variety of human appearances. By learning location-specific classifiers detection accuracy is increased significantly. However, perfect detection is impossible, thus false positives need to be filtered out. Spatio-temporal priors on detection events are either derived from a map of the environment or learned by tracking people and observing the classification failures. Furthermore, target-specific detectors are learned on-line to improve people detection in case of partial occlusions causing the generic detector to fail.

Human motion is very complex but follows non-random, non-linear patterns. Such place-dependent patterns are either learned by observing people or modeled with social forces to allow spatial and social informed human motion prediction. Latter integrates inner motivation, estimated goals, social rules, and physical constraints. When people are encountered in groups formed by social relations between individuals their geometrical arrangement is learned on-line to predicts maneuvering people jointly over intra-group constraints.

This thesis presents progress on spatial, temporal, and socially informed probabilistic data association. Learning and integrating spatio-temporal prior and target probabilities into the multi-hypothesis tracking framework people tracking is made substantially more accurate without compromising efficiency. Furthermore, a physically grounded occlusion model and a novel approach on socially information group detection guides the hypotheses generation and results in a refined probability distribution over hypotheses.

Tracking various dynamic objects makes it hard to design suitable models for their appearances and appearance dynamics, manually. This work proposes an unsupervised learning approach for

representing the time-varying appearance of objects using probabilistic exemplar-based models. Employing RGB-D cameras for 3D tracking the appearance of people is learned on-line using boosting. Target-specific appearance models support detection via a depth informed confidence search and tracking via a joint likelihood data association. Integrated into a decisional framework with the MHT on-line learning is controlled through track interpretation feedback that avoids drift.

All proposed methods are extensively analyzed using large real world experiments. It is shown, that the integration of spatial, temporal and social information enhances people detection, eases the interpretation of detection events, improves motion prediction, and guides data association. Measured with a state-of-the-art tracking metrics all methods increase the accuracy of people detection and tracking. Furthermore, their computational complexity is analyzed and it is shown, that despite the expensive of some models people tracking can still be applied in real-time.

Zusammenfassung

Aktuelle Forschungsprojekte beschäftigen sich immer häufiger mit der Fragestellung, wie Menschen und Maschinen koexistieren, sich gegenseitig unterstützen und miteinander gemeinsame Ziele erreichen können. Diese Arbeiten untersuchen nicht nur die direkte Mensch-Maschine Interaktion wie z.B. die Möglichkeiten der verbalen und nonverbalen Kommunikation, Potentiale und Probleme bei geteilten Arbeitsbereichen oder die Unterstützung des Menschen im Straßenverkehr, sondern vor allem die Integration von Maschinen in das soziale Umfeld unseres täglichen Lebens. Unser Alltag wird durch zum Teil ungeschriebene soziale Regeln und Normen bestimmt. Ein Abweichen von diesen Verhaltensregeln wird selbst vertrauten Menschen nicht leicht verziehen. Soll die Integration einer Maschine oder eines Roboters[1] in unseren Alltag gelingen, muß das menschliche Verhalten besser verstanden und in formale, mathematisch beschreibbare Regeln übersetzt werden.

Nun ist der Begriff des sozialen Verhaltens sehr unspezifisch und kann zum Beschreiben des menschlichen Zusammenlebens im Allgemeinen verwendet werden. Im Rahmen dieser Arbeit beschränkt sich der Ausdruck "soziales Verhalten" auf die Bewegung von Personen in ihrer Umwelt. Detailierter formuliert betrachtet diese Arbeit z.B., wie sich Personen im Fussgängerverkehr oder im Büro auf sozial kompatible Weise von einem Ort zum anderen bewegen. An einigen negativen Beispielen verdeutlicht, lässt sich leicht nachvollziehen, dass das Gehen auf der falschen Straßenseite, das Hindurchdrängen durch andere Personengruppen oder das Rennen in beruhigten Arbeitsräumen schnell als unangenehm und unangemessen empfunden wird. Im Gegensatz zu diesen Beispielen, für die sich das richtige Verhalten leicht durch wenige soziale Grundregeln beschreiben läßt, gibt es viele Situationen, die durch mehrere, meist subtile Verhaltensregeln beschrieben werden. Diese von Hand zu definieren ist weitaus schwieriger und in manchen Fällen sogar unmöglich. Eine Alternative zur akribischen Definition der Verhaltensregeln von Hand stellt das automatisierte Lernen dieser Regeln durch das Beobachten von Personen in ihrer täglichen Umgebung dar. Unter der Annahme, dass sich die beobachteten Personen sozial verhalten, läßt sich aus den gewonnenen Daten mit verschiedenen Methoden der Statistik und des maschinellen Lernens eine Menge an Verhaltensregeln ableiten. Befindet sich später ein Roboter in ähnlicher Situation, kann er auf Grund der gelernten Regeln sozial akzeptables Verhalten zeigen.

Durch die Notwendigkeit, Beobachtungsdaten von Personen in ihrem alltäglichen Leben zu sammeln, erhalten Algorithmen zur *Erkennung* von Personen sowie Algorithmen zur *Nachführung* der Bewegung mehrerer Personen[2] im Raum immer mehr Aufmerksamkeit. Darüber hinaus müssen Roboter, die sich im menschlichen Alltag aufhalten und (sozial) mit Menschen interagieren sollen, diese auch erkennen und ihre Bewegung nachvollziehen und bestenfalls sogar vorhersagen können. Aus diesen Gründen sind zuverlässige, robuste und schnelle Verfahren zur Personenerkennung und Bewegungsnachführung unabdingbar.

Algorithmen zur Bewegungsnachführung basieren auf der kontinuierlichen Beobachtung von Interaktionspartnern in der Umgebung des Roboters. Aus den Beobachtungsdaten werden die Positionen der Personen berechnet und mit Hilfe physikalischer Bewegungsmodelle zeitlich geglättet. Sobald die Bewegung einer Person bekannt ist, läßt sich daraus sogar eine Vorhersage ihres zukünftigen Verhaltens ableiten. Eine Schwierigkeit bei der Beobachtung von Personen ist allerdings, dass diese sich häufig gegenseitig verdecken oder selbst durch statische Objekte in der Umgebung verdeckt werden.

[1] Bedeutung nach Duden: (der menschlichen Gestalt nachgebildete) Apparatur, die bestimmte Funktionen eines Menschen ausführen kann; Maschinenmensch.

[2] Das Erkennen von Personen in Sensordaten wird im englischen als "people detection" und das Nachführen der Bewegung mehrerer Personen als "people tracking" bezeichnet. Von letzterem leitet sich der Titel dieser Arbeit ab.

Sind die Phasen der Verdeckung zu lang, gehen die Bewegungsinformationen verloren oder werden ungenau, ein korrektes Nachführen der Bewegung kann in diesen Fällen nicht mehr gewährleistet werden. Umgekehrt können falsche Messungen zu fehlerhaften Beobachtungen und Bewegungsinformationen führen. Es ist z.B. schwierig, zwischen echten Personen und Abbildungen von Personen auf Plakaten zu unterscheiden. Wird eine Abbildung fälschlicherweise als Person erkannt, wird auch eine falsche Bewegung – oder in diesem Fall ein Stillstehen – beobachtet. Ein weiteres Problem stellt die Identifikation einzelner Personen dar. Viele Sensoren stellen Daten zur Verfügung, in denen sich Personen so sehr gleichen, dass sie nicht voneinander unterschieden werden können. Kreuzen sich die Wege dieser Personen, kommt es zur Vertauschung der Bewegungsinformationen und damit zur Verwechslung der Identitäten. Dies ist in etwa mit der Verwechslung von Zwillingen vergleichbar, doch kann bereits eine ähnliche Körperstatur oder Kleidungsfarbe zu einer Verwechslung führen.

Die hier vorliegende Arbeit beschäftigt sich vor allem mit der Fragestellung, wie sich die Algorithmen zur Personenerkennung und Bewegungsnachführung unter Verwendung *räumlicher*, *zeitlicher* und *sozialer* Informationen verbessern lassen. Mann kann selbst beobachten, dass sich Fussgänger in weiten Parkanlagen anders verhalten als in engen Einkaufsstraßen. Zudem laufen sie an Arbeitstagen, wenn sie es eilig haben, schneller als am Wochenende. Gruppen von Personen – besonders Eltern mit ihren Kindern – trennen sich selten auf, auch dann nicht, wenn sie einzeln viel schneller vorankommen würden. Es gibt aber auch viele technische Herausforderungen. So liefern Kameras bei Tageslicht bessere Bilder als in der Dämmerung oder nachts. Verwechslungen von Personen oder gar falsche Messungen – wie das Erkennen von Personen auf Plakaten – sind im Durcheinander eines gut besuchten Marktes viel wahrscheinlicher als an Orten mit nur wenig Betriebsamkeit. Diese Arbeit untersucht den Einfluss, den Hintergrundwissen über menschliches Verhalten auf die Genauigkeit der Verfahren zur Personenerkennung und Bewegungsnachführung hat.

Ein Ziel dieser Arbeit ist es, die Bewegungsvorhersage von Personen zu verbessern. Dabei wird zum einen das sogenannte Soziale-Kräfte-Modell verwendet, um physikalische und soziale Einflüsse auf das Verhalten von Menschen zu simulieren. Des Weiteren werden Gruppenzugehörigkeiten untersucht, da sich aus dem Verhalten der gesamten Gruppe eine genauere Vorhersage der Bewegung einzelner Personen ableiten läßt. Ein weiterer Aspekt betrifft die Verwendung modellierter oder gelernter statistischer Informationen. Aus der Beobachtung menschlichen Verhaltens läßt sich ebenfalls eine sehr genaue Vorhersage der Bewegung einzelner Person berechnen. Darüber hinaus können Modelle der Umwelt herangezogen werden, um falsche Informationen z.B. an Plakatwänden herauszufiltern. Weiterhin betreten Menschen einen beobachteten Raum nicht an beliebigen Orten, sondern erscheinen in Türen, treten hinter Hindernissen hervor oder kommen um Ecken. Auch diese Informationen lassen sich lernen oder von Hand modellieren. Liefert der Sensor Daten, in denen sich einzelne Personen identifizieren lassen, kann man deren Aussehen lernen und zur späteren Wiedererkennung verwenden. Mit diesem Ansatz ist eine Verwechslung von Personen unwahrscheinlicher, die Verfahren zur Personenerkennung und Bewegungsnachführung werden robuster. Zuletzt lassen sich manche Teilnehmer im Strassenverkehr nicht nur durch ihr Aussehen, sondern auch durch ihre Bewegung unterscheiden. So sehen sich stehende Fussgänger und Rollschuhfahrer zwar sehr ähnlich, sobald sie sich in Bewegung setzen, ist eine Verwechslung aber so gut wie ausgeschlossen. Es wird gezeigt, dass sich mit Hilfe unüberwachter Lernverfahren dynamische Modelle verschiedener Verkehrsteilnehmer durch Beobachtung erstellen lassen.

Diese Arbeit beschäftigt sich aber nicht nur mit der Modellierung sensor- und personenspezifischen Verhaltens, sondern auch mit der Integration des gewonnenen Hintergrundwissens in modernste Algorithmen zur Personenerkennung und Bewegungsnachführung. Des Weiteren werden alle Ansätze unter Verwendung umfangreicher, realer Datensätze getestet und ausgewertet.

Acknowledgments Danksagung

First of all, I would like to thank my advisor Kai O. Arras for his great ideas and tremendous support. He had a major influence on the success of my research and this thesis. It was a pleasure to work in his laboratory and I enjoyed the opportunities to visit interesting conferences. I learned a lot during this time and I am convinced that this knowledge will help me in the future. I will remember the time in Freiburg as a formative and highly enjoyable period of my life.

I also like to thank Thomas Brox for reviewing this thesis and for acting as an external referee.

I want to thank Gian Diego Tipaldi, Luciano Spinello, and Dizan Vasquez for helping me to develop my scientific skills, working with me on interesting research questions, and creating an enjoyable and productive working atmosphere in the lab. Your contributions to this thesis can not be measured as well.

Furthermore, I want to thank Wolfram Burgard for giving me the opportunity to start my scientific career in his lab, introducing me into research, and teaching me scientific writing. A huge thank to Christian Plagemann who improved my presentation skills. It took a while but finally I got it. Also many thanks to all my friends and colleagues at the Autonomous Intelligent Systems lab. Axel Rottmann, Rainer Kümmerle, Bastian Steder, Michael Ruhnke, Daniel Meyer-Delius, Slawomir Grzonka, Barbara Frank, Jürgen Hess, Felix Endress, Christoph Sprunk, Boris Lau, Jürgen Sturm, and many more, it was a pleasure to share your experiences. Please, keep improving your soccer skills.

I also want to thank Maren Bennewitz for the interesting discussions and helpful advices. Thanks to Armin Hornung and Daniel Maier, members of the Humanoid Robotics Lab, for many interesting discussions during lunch and coffee breaks.

For the warm welcome in the DESIRE project I want to thank Yulia Sandamirskaya, Ulrich Reiser, Uwe Handmann, and Thilo Grundmann. It was a pleasure to be part of the team. I will never forget your commitments to get the system running.

Many thanks to my students Johannes A. Stork and Jens Silva. You have done great research and I really enjoyed the time you spend at our lab. I know we had some ambitious and challenging tasks but we did it. Many thanks to my students Markus Schwenk, Luc Lanners, Severin Gustorff, and Ivo Malenica for helping me developing our research platform DARYL and for collecting real data sets.

My deepest gratitude goes to my parents and my beloved wife Alexandra. For all their support and love they gave me in every period of my life I want to thank them with a few personal words in German below.

Ein paar besondere Worte des Dankes möchte ich an meine Eltern und meine geliebte Ehefrau Alexandra richten. Für die grenzenlose Unterstützung in den Jahren des Studierens und Forschens ist jeder Dank zu gering. Ihr habt die Rahmenbedingungen geschaffen die mir zur Verwirklichung meiner Ziele verholfen haben. Ich habe stets Euer vollstes Vertrauen gespürt. Dieses Vertrauen verbindet ein Leben lang.

This work has partly been supported by the German Research Foundation (DFG) under contract number SFB/TR-8 Spatial Cognition, the EC under contract number FP6-IST-045388, and the German Federal Ministry of Education and Research (BMBF) within the research projects DESIRE under grant number 01IME01F.

Contents

1 Introduction

Considerable research in the last decades, especially in the fields of artificial intelligence, machine learning, and robotics, enabled robots to enter and operate in human domains. Robotic applications include automatic production lines, ware houses, mining, aerospace, research, and many more. Usually robots are employed under human supervision and in isolation to the human workers. A fast and reliable hand-in-hand cooperation between humans and robots is still not possible, either due to safety reasons or due to task complexity. However, human-robot interaction offers many opportunities and advantages, thus researchers have shifted their attention from pure robotics related topics – like SLAM, object recognition, grasping, to name a few – to a new field called *human oriented robotics*, recently. Projects in this field focus on both questions: How to improve robotics itself ? and How to better integrate robots into domestic and professional every day life of humans ?

With robots operating in human populated environments, people detection and tracking becomes a key technology. For instance, precise knowledge about the presence, position, motion, and movement intention of people is basis for navigation and collaboration tasks. Not being imperative for collision free navigation in principle, considering humans and their needs is fundamental for socially compliant robot behavior. While, so far, humans have typically been treated as (static or dynamic) obstacles in the environment, recent approaches for robot navigation account for their dynamics and incorporate social rules to mimic human behavior.

Successfully employing robots in various environments and ensuring their safe and reliable operation depends on people detection and tracking algorithms, that are accurate and robust under wide ranges of environmental conditions. This challenging task cannot be solved, reliably, by detection only. To give some examples, vision-based detectors break down in the dark, laser-based methods suffer from little information and fail in far distances, and camera-based range sensors – like the MS Kinect – provide no data in close ranges. Even a detector combining multiple of these sensors fails if people are hidden behind static obstacles or occlude each other. Especially latter occurs quite often when observing people in groups from a first-person perspective. In such a case, a reliable detection of single individuals is almost impossible. In addition this, photographs of people, mirror images or reflections, and random clutter produce false detections that need to be filtered out. Otherwise the robot might be confused in executing its task. Last, in some sensory data people have similar (or even identical) appearance, thus a pure detection-based approach fails to maintain the identities of people.

These examples show, that detection only will not be sufficient and that tracking itself must evolve to cope with the uncertainties of false positives, occlusions, and misdetections. In the past, data association has been improved from simple strategies like (global) nearest neighbor (NN, GNN) to more advanced methods like the (joint) probabilistic data association filters (PDAF, JPDAF) or the multi-hypothesis tracking (MHT) approach. Recently, sampling based approaches like Markov chain Monte Carlo (MCMC) based tracking have been introduced. Besides data association, refined motion prediction is addressed in related literature. Strategies include learning preferred trajectories from human observation, modeling environmental-specific local and global goals, or employing interactive multiple models (IMM) for more general predictions of human motion. However, all of these approaches integrate human-specific information into their predictions to improve accuracy. The same idea can be applied to data association.

In this work, we focus on the integration of a wide range of human-specific information into detection, motion prediction, and data association. Furthermore, strategies on exchanging information between these main components of tracking are investigated to break the mold that data is only forwarded from the detection to the tracking stage (known as Tracking-By-Detection algorithms). Thereby, the goal is to improve the accuracy and robustness of the tracking system in its entirety.

In more detail, this thesis addresses the integration of *spatial*, *temporal*, and *social* information to support the interpretation of detection events, to further improve motion prediction, and to guide data association. Thereby, statistical and individual information is either modeled in advance, learned off-line from training data, or learned on-line by observing humans and their behaviors. As discussed more extensively in the next section on contributions, background knowledge is used to identify systematic detection failures and aids to avoid tracking of "ghost" targets. Social and physical constraints aid to improve motion prediction of individuals and groups. Analysing grouping behaviors of people supports motion prediction and data association and increases the accuracy of tracking single people in groups, dramatically, even if they cannot be detected reliably due to occlusions. Moreover, with sensor providing rich information, on-line learned target-specific appearance and dynamics models are used to detect and classify targets more reliably.

To the best of my knowledge, the proposed approaches result in the best people detection and tracking system in 2D range data and the first accurate and reliable system to track people in 3D using data from a Velodyne 3D laser range finder or MS Kinect RGB-D cameras.

1.1 Contributions

As outlined in the previous section, the main contribution of the thesis is the integration of *spatial*, *temporal*, and especially *social* information into the people detection and tracking framework to improve its accuracy and robustness. A more detailed overview of the kind of information that is employed to improve people detection, human motion prediction, and data association is given below.

Each target tracking system has to classify the sensory data to detect the designated targets. If robots are equipped with 2D laser range finders – often employed for safety reasons, mapping, localization, and collision free navigation – the ability to detect people in laser range data has the advantage that no further modifications of the robot are required. Currently, target-specific classifiers are learned off-line from manually annotated training data. The integration of incomming information is either not possible or requires expensive re-training. The contributions on improved people detection in this work are twofold: First, we train a cascade of multiple place-dependent detectors to increase the accuracy of the a-priori detector in general. Each detector employs a set of geometrical and statistical features trained with AdaBoost. Second, with rich sensor modalities, that allow the identification of individual targets[3] we use on-line learning of target-specific appearance models to detect people in case the a-priori detector fails. On-line learning is implemented using on-line-boosting. Both detectors are integrated into a decisional framework with a multi-hypothesis tracker that controls on-line learning through a track interpretation feedback. As we will show, combining the a-priori and on-line detectors leads to reliable 3D tracking with increased tracking performance and avoids drift of the on-line detectors.

Once the desired targets can be detected, a motion model is required to predict their future states. Most related work make only weak assumptions on the motion of humans and employ either the Brownian model, the constant velocity model, or constant acceleration model. Since these models

[3] RGB-D sensors that provide rich color and depth information have just been launched, recently.

are not very adequate to predict the highly dynamic movements of people, in this work we improve motion prediction by integrating spatial and social information. By observing the motion of people, we learn the walkable area of an environment and describe it by a spatial Poisson process. Employing rejection sampling, we derive a place-dependent motion model whose predictions follow the space usage patterns that people usually take. Furthermore, we use the social force model to describe physical and social constraints of other people and static objects in the environment. In addition with on-line estimated short-term goals, motion prediction of individual people is improved. For people that share social relations with others and walk in groups we learn geometrical (spatial) relations informed by priors from the social science community in an on-line fashion. Using a particle-based approach our method is able to jointly predict human motion over intra-group constraints. Tracking accuracy and reliability is improved even in case of lengthy occlusion events when single people can not be detected.

A major problem, while tracking multiple dynamic targets is to associate the previously known targets with the current observations (known as data association problem). In case targets have identical appearance – which is the case for 2D range data – crossing trajectories can easily cause track confusions (or identifier switches). To avoid the confusion of people walking in groups, we use the formerly mentioned social constraints and geometric relations to guide data association in a multi-hypothesis tracking framework. We will show, that our approach is able to decrease the number of identifier switches, dramatically.

Further, tracking algorithms have to declare unassigned observations to emanate from new targets or clutter. For that purpose, we learn statistics of how people use the environment and where detectors fail systematically from human observations. The statistics are represented by spatio-temporal Poisson distributions in a so called spatial affordance map. From this representation we infer locations where people enter and leave the environment. Static obstacles that impair detection are filtered out. Unassigned targets are caused by people that are either hidden or have left the monitored area. Consequently, the tracker has to declare their tracks as occluded or obsolete. We learn spatial distributions on both of these events from manually annotated data and on-line from human observations. In addition to learning, we propose a physical model to calculate the occlusion probability of people using the current scene information. Therefore, we check the visibility of their predicted positions using a particle-based approach and ray tracing. The learned statistics and the physical occlusion model support tracking, thus the numbers of missed and wrongly tracked targets decreass, significantly.

All proposed models are valid for people tracking in general and can be integrated into any probabilistic target tracking framework, regardless the sensor modality, the filtering approach, or the space in which targets are represented. Researcher have developed many different probabilistic tracking algorithms. Known to be one of the most general approaches, able to handle the entire life-cycle of tracks – from creation and confirmation to occlusion and deletion – we selected the multi-hypothesis tracking (MHT) approach in this thesis. To integrate the learned and modeled information and to achieve refined hypotheses distributions, we extended the multi-hypothesis tracking framework by multiple aspects. Detailed theoretical information and descriptions on feasible implementations of those extensions are also provided in this thesis.

Tracking and classifying various dynamic objects – like humans, animals, vehicles – makes it hard to manually design suitable models for their appearances and appearance dynamics. Thus we present an unsupervised learning approach for representing the time-varying appearance of objects in 2D range data using probabilistic exemplar-based models. Employing a clustering procedure that builds a set of object classes from given observation sequences our system is able to autonomously learn useful models for, e.g., pedestrians, skaters, or cyclists without any external class information.

1.2 Publications

Parts of the thesis have been published in the following journal articles, conferences, symposia, and workshop proceedings:

- M. Luber and K. O. Arras, Multi-Hypothesis Social Grouping and Tracking for Mobile Robots. In *Proceedings of Robotics: Science and Systems (RSS)*, Berlin, Germany, 2013. Best Student Paper Award Nominee.
- K. O. Arras, B. Lau, S. Grzonka, M. Luber, O. Mozos, D. Meyer-Delius, and W. Burgard. Range-based people detection and tracking for socially enabled service robots. In *Towards Service Robots for Everyday Environments: Recent Advances in Designing Service Robots for Complex Tasks in Everyday Environments*, 235–280, 2012.
- M. Luber, L. Spinello, and K. O. Arras. People Tracking in RGB-D data With On-line Boosted Target Models. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Francisco, USA, 2011.
- M. Luber, L. Spinello, and K. O. Arras. Learning to Detect and Track People in RGB-D Data. In *The Workshop on RGB-D Cameras. Robotics: Science and Systems (RSS)*, Los Angeles, USA, 2011.
- M. Luber, G. D. Tipaldi, and K. O. Arras. Better Models For People Tracking. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, Shanghai, China, 2011.
- M. Luber, G. D. Tipaldi, and K. O. Arras. Place-dependent people tracking. In *International Journal of Robotics Research (IJRR)*, 30(3):280–293, March 2011.
- L. Spinello, M. Luber, and K. O. Arras. Tracking People in 3D Using a Bottom-Up Top-Down Detector. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, Shanghai, China, 2011.
- M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras. People Tracking with Human Motion Predictions from Social Forces. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, Anchorage, USA, 2010.
- M. Luber, G. D. Tipaldi, J. A. Stork, and K. O. Arras. People Tracking with Social Force-Based Motion Prediction. In *International Conference on Cognitive Systems (CogSys)*, Zurich, Switzerland, 2010.
- M. Luber, G. D. Tipaldi, and K. O. Arras. Place-Dependent People Tracking. In *Proceedings of the International Symposium of Robotics Research (ISRR)*, Lucerne, Switzerland, 2009.
- M. Luber, G. D. Tipaldi, and K. O. Arras. Spatially Grounded Multi-Hypothesis Tracking of People. In *The Workshop on People Detection and Tracking. International Conference on Robotics & Automation (ICRA)*, Kobe, Japan, 2009.
- M. Luber, K. O. Arras, C. Plagemann, and W. Burgard. Classifying dynamic objects: An unsupervised learning approach. In *Autonomous Robots*, 26(2-3):141–151, 2009.
- M. Luber, K. O. Arras, C. Plagemann, and W. Burgard. Classifying Dynamic Objects: An Unsupervised Learning Approach. In *Proceedings of Robotics: Science and Systems (RSS)*, Zurich, Switzerland, 2008.

Outside the scope of this thesis, the following articles have been published:

- M. Luber, J. Silva, L. Spinello, and K. O. Arras. Socially-Aware Robot Navigation: A Learning Approach. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vilamoura, Portugal, 2012.

- S. Embgen, M. Luber, C. Becker-Asano, M. Ragni, V. Evers, and K. O. Arras, Robot-Specific Social Cues in Emotional Body Language. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*, Paris, France, 2012

- K. O. Arras, S. Grzonka, M. Luber, and W. Burgard. Efficient People Tracking in Laser Range Data using a Multi-Hypothesis Leg-Tracker with Adaptive Occlusion Probabilities. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, Pasadena, USA, 2008.

1.3 Collaborations

Parts of this thesis have resulted from collaboration with colleagues and I would like to thank all of them for putting hard work in the joint projects. Specifically, the *boosted features* people detector, presented by Arras, Martínez Mozos, and Burgard [2007], has been extended by several geometrical and statistical features (see Chapter 2) which have been developed jointly with Luciano Spinello. Employing *social forces* to guide motion prediction by social and physical constraints (see Chapter 4) was originally addressed in the co-supervised bachelor thesis of Johannes A. Stork. Learning *spatial affordances* (see Chapter 6) and integrating them into the MHT formalism is based on joint work with Gian Diego Tipaldi. Tracking people in 3D (see Chapter 9) and feeding back the tracking information into detection has been investigated jointly with Luciano Spinello.

1.4 Outline

This thesis is structured in four main parts.

In *Part I: Basic Techniques for People Detection and Tracking*, the spatial informed boosted features people detector using a cascade of range specialized basis detectors is introduced and analyzed in various complex environments (Chapter 2). Subsequently, the multi-hypothesis tracking (MHT) approach employing spatio-temporal information is presented and a detailed summary of the developments made in MHT tracking in the past decades and in this thesis is provided (Chapter 3).

The *Part II: Social and Spatio-Temporal Constraints: Model-Based Approaches* shows, how computational models developed in the cognitive and social science communities can be employed to describe individual and collective pedestrian behavior. Accounting for social and physical constraints combined in the social force model refined motion predictions translate into more informed probability distributions over hypotheses and finally into a more robust tracking behavior (Chapter 4). Furthermore, human-specific models on the occurrence of new track, false alarm, occlusion, and deletion events are developed to support probabilistic data association. Integrated into the MHT framework tracking is made substantially more accurate without compromising efficiency (Chapter 5).

In *Part III: Social and Spatio-Temporal Constraints: Learning-Based Approaches*, learned probability distributions on the formerly mentioned tracking events encode probabilities and frequencies of human behavior. These spatial priors are learned from human observations using non-homogeneous

spatial Poisson processes. Integrated into the MHT more informed probability distributions over hypotheses lead to a more accurate tracking behavior. Further, a place-dependent motion model whose predictions follow the space usage patterns of people is presented (Chapter 6). Tracking the social relations between individuals over time group affiliations are inferred to improve data association. Moreover, learning geometric relations between people in groups on-line, human motion is predicted jointly using intra-group constraints and a particle-based approach. Both support data association in a multiple hypothesis tracking framework with adaptive occlusion probabilities (Chapter 7).

In *Part IV: Learning Appearances and Appearance Dynamics*, an unsupervised learning approach to find suitable models for the appearance and dynamics of various dynamic entities is presented. Employing exemplar-based models for representing the time-varying appearance of objects and a clustering procedure that builds a set of object classes useful models for, e.g., pedestrians, skaters, or cyclists are learned without being provided with any external class information (Chapter 8). Further, the first reliable and accurate approach on 3D people detection and tracking in RGB-D data is proposed. Individual target appearance models are learned on-line using boosting to select among different types of RGB-D features. Combined with a novel multi-cue person detector and integrated into the MHT tracking is continued even in case the a priori detector fails – e.g. during lengthy events of partial occlusion. Employing a refined depth informed confidence search in 3D increases the tracking accuracy. A decisional framework within the MHT using track interpretation feedback controls on-line learning and avoids drift of the on-line detectors (Chapter 9).

In all chapters, extensive experiments in various indoor and outdoor environments demonstrate the benefits of the presented methods. Furthermore, their computational complexity is analyzed to assure that tracking can still be applied in real-time, even when multiple people are present. Chapter 10, finally, recapitulates the contributions of the thesis, discusses its results, and proposes future research directions for people tracking in general.

Part I

Basic Techniques for People Detection and Tracking

2 People Detection in 2D Range Data

Robust and reliable people detection is a key component for robots operating in human environments. Current approaches on people detection learn generic classifiers from manually annotated training data that are employed to detect people at all locations of the monitored environment. But for most sensor modalities the appearance of people heavily depends on their relative position to the sensor and time-dependent external conditions. Thus, the detection accuracy scales with the variety of the provided sensor information. To resolve this problem, spatio-temporal information can be utilized to learn place and time dependent classifiers for improved people detection.

This chapter focuses on the integration of spatial information into the boosted features people detector proposed by Arras et al. [2007] to derive a set of place-dependent detectors. Assuming that the robot is equipped with 2D laser range finders the distance of objects is available allowing to train a set of range-specific classifiers. Each classifier represents a specialized person detector with improved detection accuracy. In addition to the place dependency, the original set of geometric and statistical features employed by Arras et al. is extended to better capture the variability in the appearance of people.

Extensive experiments in large outdoor environments analyze the contributions of the introduced features and demonstrate that the proposed approach exceeds state-of-the-art methods in range based people detection. When detecting people in up to 20 meters distance the overall accuracy is 80% with 91% true positives and 94% true negatives, respectively. Furthermore, the proposed detector can be transfered to new environments containing people of different appearance.

This chapter is structured as follows. Introduction and related work are presented in section 2.1 and section 2.2, respectively. Section 2.3 explains the boosted features people detection in 2D range data, reviews the original feature set, and presents additional features. Subsequently section 2.4 introduces the proposed cascade of place-dependent strong classifiers. In section 2.5 the experimental results are presented followed by the conclusions in section 2.6.

2.1 Introduction

People detection is a key technology for social robots operating in populated environments. Moreover, it is fundamental for tracking people in the surroundings, especially when using the tracking-by-detection (TBD) scheme. Both – detection and tracking – in combination enables the generation of socially acceptable behavior. While most related work in this area focuses on vision based approaches to solve the detection task, range sensing is a particularly interesting sensor modality due to its accuracy, large field of view, and robustness with respect to illumination changes and vibrations. Furthermore, robots sharing space with humans employ range sensing for collision free navigation and other safety reasons. Thus people detection in range data does not imply any further modifications of the mobile robot.

In the context of people tracking, the focus has mostly been shifted on tracking algorithms and data association rather than on people detection. The motivation of this work is the belief that the definition of appropriate features and the integration of spatio-temporal information has a major impact on the accuracy, robustness, and reliability of people detection algorithms. For this reason

the boosted features detector proposed by Arras et al. [2007] is extended in two manners. First, the original set of geometric features is extended to better capture the huge variety in human appearance. And second, the generic detector is extended to forward the classification process to an expert detector based on spatial information. (The integration of temporal information is mainly ignored in this section as laser based detection is robust against time-dependent changes of the environment.) The resulting cascade of range-specific people detectors is more robust against the diversity of people's appearances due to distance changes and allows more accurate classification.

The detection approach presented in this chapter is employed as people detector in the remainder of this thesis with the exception of Chapter 8 and Chapter 9 that are concerned with detecting and tracking arbitrary objects or people in RGB-D data.

2.2 Related Work

Many researchers have addressed the task of people detection in 2D laser range data. In early works of Kluge et al. [2001], Fod et al. [2002], and Schulz et al. [2003], people are detected using ad-hoc classifiers, looking for moving local minima in the scan.

The first principled learning approach has been taken by Arras et al. [2007] where a classifier for 2D point clouds has been learned by boosting a set of geometric and statistical features. AdaBoost has been successfully used in different applications for object recognition. Viola and Jones [2002] boost simple features based on gray level differences to create a fast face detector using images. In Treptow and Zell [2004] AdaBoost is used to track a ball without color information in the context of RoboCup. The work of Verschae et al. [2008] focuses on training nested cascades of boosted classifiers for the purpose of face detection with high accuracy, robustness, and training speed.

As there is a natural accuracy limitation when using only a single layer of 2D range data, several authors have been using multiple co-planar 2D laser scanners like Gidel et al. [2010] and Carballo et al. [2008]. In the work of Martínez Mozos et al. [2010] the authors apply boosting on each of three layers and use a probabilistic scheme to combine the three classifiers in a flattened 2D space.

In the field of people detection in 3D data, Navarro-Serment et al. [2010] collapse the 3D scan into a virtual 2D slice to find salient vertical objects above ground. They align a window to the principal data direction, compute a set of features, and classify pedestrians using a set of SVMs. In the work of Bajracharya et al. [2009] people are detected in point clouds from stereo vision by processing vertical objects and considering a set of geometrical and statistical features of the cloud based on a fixed pedestrian model. Both of them can be seen as top-down detection procedures. Unlike these works that require a ground plane assumption, Spinello et al. [2011] overcome this limitation via a voting approach of classified parts and a top-down verification procedure that learns an optimal feature set and volume tessellation.

In the computer vision literature, the problem of detecting, tracking and modeling humans has been extensively studied by Dalal and Triggs [2005], Felzenszwalb et al. [2008], Leibe et al. [2005], and Enzweiler and Gavrila [2009]. A major difference to range-based systems is that the richness of image data makes it straightforward to learn target appearance models.

Dense depth data from stereo are used by Beymer and Konolige [1999] to support foreground segmentation in an otherwise vision-based people detection and tracking system. They use a set of binary person templates to detect people in images and demonstrate multi-person tracking with learned appearance-based target models. The work of Leibe et al. [2008] and Ess et al. [2009a] detect people in intensity images and track them in 3D. In Enzweiler et al. [2010] a stereo system for combining intensity images, stereo disparity maps, and optical flow is used to detect people. Multi-modal detection and tracking of people is performed in Spinello et al. [2010b] where a trainable 2D

range data and camera system is presented.

This work focus on the integration of spatial information that has not been used by any of the related works. By learning place-dependent classifiers the problem of changing appearances depending on relative position of people and sensor can be eliminated. Thereby, a novel, place-dependent classification schemes based on a cascaded of range-specific classifiers is introduced. Moreover, defining a suitable set of features for 2D laser based people detection the variety in the appearance of people – for example when wearing luggage or heavy coats– is well captured.

2.3 Boosted Features People Detector in 2D Range Data

The approach of Arras et al. [2007] applies boosting to train a strong classifier composed of a set of weak classifiers using simple features for the purpose of leg detection. The features defined in subsection 2.3.1 are calculated on segmented groups of neighboring beams corresponding to human legs in range data. These segments are found using a simple jump distance condition on the range of consecutive beams. As the boosted features detector is employed to train range-specific classifiers it is explained in more detail in this section.

Boosting is a general method to find a highly accurate *strong* classifier by combining many *weak* classifiers, each of them being only moderately accurate. Typically, each weak classifier contains a simple rule which can be used to generate a predicted classification for any instance. One requirement to each weak classifier is that its accuracy is better than a random guess. Further, the employed AdaBoost algorithm introduced by Schapire and Singer [1999] extending the original approach of Freund and Schapire [1997] assumes each weak classifier generates not only predicted classifications, but also self-rated confidence scores which estimate the reliability of each of its predictions.

The input to the algorithm is a set of annotated training examples $\mathcal{E} = \{(\varepsilon_1, l_1), \ldots, (\varepsilon_N, l_N)\}$, where each ε_i[4] is an example and $l_i \in \{-1, +1\}$ indicates whether ε_i is negative or positive, respectively. In a series of $k \leq K$ iterations, the algorithm trains all available weak classifiers using a weight distribution δ over the training examples and selects the best one called h_j, repeatedly, with $1 \leq j \leq k$. The selected weak classifier h_j is expected to have the smallest classification error w.r.t. the weighted training examples. The idea of the algorithm is to modify the distribution δ at each iteration to increasing the weight (also called importance) of those examples that have been classified incorrectly by the previously selected weak classifier h_j. The final strong classifier H is a weighted majority vote of the k best weak classifiers. Given an unknown example ε its label l is calculated using the strong classifier H by

$$l(\varepsilon) \;=\; H(\varepsilon) \;=\; \operatorname{sgn}\left(\sum_{j=1}^{k} \alpha_j \; h_j(\varepsilon)\right), \tag{2.1}$$

where $l(\varepsilon) \in \{-1, +1\}$ is the classification label assigned to ε by the classifier H and $\vec{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$ is the weight vector of the weak classifiers. Large weights are assigned to good weak classifiers whereas poor classifiers receive small weights.

[4] Each training example ε_i consists of all feature values calculated from the corresponding laser end points and is therefore also called feature descriptor.

Algorithm 1: The generalized AdaBoost learning algorithm.

Input : Set of N examples $\mathcal{E} = \{(\varepsilon_1, l_1), \ldots, (\varepsilon_N, l_N)\}$ with annotations $l \in \{-1, +1\}$ denoting positive ($l = +1$) and negative ($l = -1$) examples, respectively.
Set of M features $\mathcal{F} = \{f_1, \ldots, f_M\}$.
K denoting the maximum number of weak classifiers to train.

Output : Vector of weak classifiers $\vec{h} = \{h_1, \ldots, h_k\}$ and
vector of weights $\vec{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$ both defining the strong classifier H.
$k < K$ denoting the true number of trained weak classifiers.

Variables: Vector of weights $\vec{\delta} = \{\delta_1, \ldots, \delta_N\}$ denoting the importance of the examples.
Vector of temporary weak classifiers $\vec{\omega} = \{\omega_1, \ldots, \omega_M\}$ trained in each iteration.

```
/* Initialize uniform example weights */
for i ← 1 to N do
    if l_i = +1 then
        δ_i = 1/(2a); // where a denotes the total number of positive examples
    else
        δ_i = 1/(2b); // with b the total number of negative examples

/* main loop, learn the weak classifiers */
for k ← 1 to K do
    // 1.) normalize weights distribution δ⃗ so that Σ_i δ_i = 1
    for i ← 1 to N do
        δ_i = δ_i / Σ_i δ_i;
    // 2.) train weak classifiers ω_j using f_j ∈ F, E, and δ⃗
    for j ← 1 to M do
        ω_j = train(f_j, E, δ⃗);
    // 3.) calculate errors of weak classifiers ω_j
    for j ← 1 to M do
        r_j = Σ_{i=1}^{N} (δ_i l_i ω_j(ε_i));   // ω_j(ε_i) ∈ {−1, +1}
    // 4.) choose best weak classifiers
    r_k = min_j(|r_j|);
    if r_k ≥ 1/2 then
        break;  // stop training of weak classifiers and return
    α_k = 1/2 log((1 + r_k)/(1 − r_k));
    h_k = min_{|r_j|}(ω_j);
    // 5.) update weights δ_i
    for i ← 1 to N do
        δ_i = δ_i exp(−α_k l_i h_k(ε_i));

/* training of H done, the strong classifier is given by: */
/* H(ε) = sgn( Σ_{j=1}^{k} α_j h_j(ε) ), with h_j(ε_i) ∈ {−1, +1} */
return h⃗, α⃗, k;
```

Figure 2.1: The generalized AdaBoost learning algorithm.

The weak classifiers are trained using decision stumps following the approach of Viola and Jones [2002]. For each feature f_j, the corresponding weak classifier h_j determines the optimal threshold classification function, such that the minimum number of examples are misclassified. In detail, a weak classifier h_j consists of single-valued feature f_j, a threshold θ_j, and a parity $p_j \in \{-1, +1\}$ indicating the direction of the inequality sign and has the form:

$$h_j(\varepsilon) \;=\; \begin{cases} 1 & \text{if } \; p_j \, f_j(\varepsilon) < p_j \, \theta_j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

In each iteration of the algorithm, the values for θ_j and p_j are learned, so that the misclassification rate w.r.t. the weighted training examples is minimized. The complete AdaBoost approach is outlined in algorithm 1.

2.3.1 Original Set of Geometrical Feature

The list compiled below presents the geometrical features introduced by Arras et al. [2007]. The features are explained in detail as some of them serve to define further features in the extended set (see subsection 2.3.2) and for the sake of completeness. In the following the set of these features is denoted as *original feature set* and employed as baseline in the experimental section.

All features are calculated based on a given segment S_i that consists of a sequence of laser range readings $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots\}$. Each beam $\mathbf{b}_j$ corresponds to a tuple $\mathbf{b}_j = (\phi_j, \rho_j)$ that defines a point in a 2D plane with polar coordinates where ϕ_j denotes the angle of the beam relative to the laser sensor and ρ_i the length of the beam or the distance to the measured object, respectively. Further, the Cartesian coordinates $\mathbf{x}_j = (x_j, y_j)$ can be calculated given $x_j = \rho_j cos(\phi_j)$ and $y_j = \rho_j sin(\phi_j)$. Interesting to know, the characteristics of a laser range finder guarantees the points to be sorted in ascending ordered w.r.t their angle ϕ_j, hence $\phi_j < \phi_{j+1}, \; \forall \; j$. With that assumption many features can be calculated very efficiently. A feature f is defined as a function $f : S \to \mathbb{R}^1$ that takes a segment S as argument and returns a real value. For each segment $S_i = \{\mathbf{b}_1, \mathbf{x}_1, \mathbf{b}_2, \mathbf{x}_2, \dots\}$ the following features are determined:

1. Number of laser end points in segment S_i. It is assumed, that people have a certain width causing a specific number of range readings, defined as

$$n_i \;=\; |S_i|.$$

2. Standard deviation (or variance from center of gravity) of all points in the segment,

$$\mu_i \;=\; \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_j, \qquad \sigma_i \;=\; \sqrt{\frac{1}{n-1} \sum_{j=1}^{n_i} ||\mathbf{x}_j - \mu_i||^2}.$$

   As people are allowed to appear at all positions in the sensor field of view, especially in case of a mobile platform, the mean μ_i must not be considered as feature.

3. Mean average deviation from median. The median of a distribution $f(x)$ is the value where the cumulative distribution function $F(x) = 1/2$. Given an ordered set of K scalar random samples x_j the median $\tilde{x}_i$ of segment S_i is defined as

$$\tilde{x}_i \;=\; \begin{cases} x_{\frac{(K+1)}{2}} & \text{if } K \text{ is odd,} \\ \frac{1}{2} x_{\frac{K}{2}} + x_{\frac{K}{2}+1} & \text{if } K \text{ is even.} \end{cases}$$

This feature is designed to measure the segment compactness more robustly than the standard deviation above. Opposed to the mean, the median is less sensitive to outliers. In the multi-dimensional case, $\tilde{x}_i$ is calculate in each dimension independently.

4. Jump distance from preceding segment that is the Euclidean distance between the first point of the inspected segment S_i and the last point in the preceding segment S_{i-1}, calculate as

$$\delta_i^{prec} = ||\mathbf{x}_1^i - \mathbf{x}_{n_{i-1}}^{i-1}||^2.$$

Theoretically, it is not guaranteed that these two points are the closest ones from both segments. However, experiments prove good discrimination using the proposed features.

5. Jump distance to succeeding segment calculated in the same fashion as feature δ_i^{prec},

$$\delta_i^{succ} = ||\mathbf{x}_{n_i}^i - \mathbf{x}_1^{i+1}||^2.$$

6. Width of the segment estimated by the Euclidean distance between the first and the last laser reading in the segment S_i,

$$w_i = ||x_1 - x_{n_i}||^2.$$

Similar to the previous two features the true width of segment S_i can differ from this estimate, theoretically. However, for the sake of computational simplicity and shown by the experiments this estimate is sufficient.

7. Linearity of segment S_i that measures the straightness of the segment and corresponds to the residual sum of squares to a line fitted into the segment in the least squares sense. Using the segment points in polar coordinates, fitting a line in the Hessian (α, r)-representation that minimizes perpendicular errors from the points onto the line has a closed form solution. Using the (unweighted) expressions from Arras [2003] the residual sum of squares is calculated as

$$l_i = \sum_{j=1}^{n_i} (x_j \, cos(\alpha) + y_j \, sin(\alpha) - r)^2.$$

8. Circularity of the circle fitted into the points of segment S_i. Given a set of points in Cartesian coordinates fitting a circle in the least squares sense is to parameterize the problem by the vector of unknowns $x = (x_c \quad y_c \quad x_c^2 + y_c^2 - r_c^2)^T$ where x_c, y_c, and r_c denote the circle center and radius, respectively. Fitting the circle is equivalent to solve the overdetermined equation system

$$A = \begin{pmatrix} -2x_1 & -2y_1 & 1 \\ -2x_2 & -2y_2 & 1 \\ \vdots & \vdots & \vdots \\ -2x_{n_i} & -2y_{n_i} & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} -2x_1 - 2y_1 \\ -2x_2 - 2y_2 \\ \vdots \\ -2x_{n_i} - 2y_{n_i} \end{pmatrix},$$

using the pseudo-inverse $x = (A^T A)^{-1} A^T \cdot b$. The circularity defined by the residual sum of squares is

$$c_i = \sum_{j=1}^{n_i} \left(r_c - \sqrt{(x_c - x_j)^2 + (y_c - y_j)^2} \right)^2.$$

9. Radius $r_i = r_c$ of the circle fitted into the points of segment S_i. This feature provides an alternative measure of the size of a segment S_i and is calculated following the least squares approach described above.

10. Boundary length of segment S_i. Based on the Euclidean distances $d_{j,j+1}$ between two adjacent points j and $j+1$ this feature measures the length of the poly-line corresponding to the segment S_i as

$$b_i = \sum_{j=1}^{n_i-1} d_{j,j+1},$$

with $d_{j,j+1} = ||\mathbf{x}_j - \mathbf{x}_{j+1}||^2$.

11. The boundary regularity of the contour poly-line is measured by the standard deviation of the distances $d_{j,j+1}$ of the adjacent points in segment S_i,

$$\mu_i^d = \frac{1}{n_i - 2} \sum_{j=2}^{n_i-1} d_{j,j+1}, \qquad \sigma_i^d = \sqrt{\frac{1}{n_i - 3} \sum_{j=1}^{n_i-2} ||d_{j,j+1} - \mu_i^d||^2},$$

where the mean Euclidean distance between adjacent points μ_i^d was not included in the original feature set.

12. Mean curvature μ_i^κ defined as the average of all curvature estimates κ_j at the points of segment S_i. Given a sequence of three succeeding points $\mathbf{x}_{j-1}$, $\mathbf{x}_j$, and $\mathbf{x}_{j+1}$, the curvature at point $\mathbf{x}_j$ is calculated using the following approximation. Let A denote the area of the triangle defined by $\mathbf{x}_{j-1}, \mathbf{x}_j, \mathbf{x}_{j+1}$ and $d_{(j-1,j)}, d_{(j,j+1)}, d_{(j-1,j+1)}$ the Euclidean distances between these three points, an approximation of the discrete curvature of the boundary at $\mathbf{x}_j$ is given by

$$\kappa_j = \frac{4A}{d_{(j-1,j)}\, d_{(j,j+1)}\, d_{(j-1,j+1)}}.$$

The mean curvature of segment S_i is then calculated as

$$\mu_i^\kappa = \frac{1}{n_i - 2} \sum_{j=2}^{n_i-1} \kappa_j,$$

which provides an alternative measurement of r_c as curvature and radius are inverse proportional.

13. Mean angular difference measuring the convexity or concavity of segment S_i. This feature calculates the average of the angles β_j between the vectors $\overline{\mathbf{x}_{j-1}\,\mathbf{x}_j}$ and $\overline{\mathbf{x}_j\,\mathbf{x}_{j+1}}$, as

$$\mu_j^c = \frac{1}{n_i - 2} \sum_{j=1}^{n_i-1} \beta_j,$$

with $\beta_j = \sphericalangle(\overline{\mathbf{x}_{j-1}\,\mathbf{x}_j}, \overline{\mathbf{x}_j\,\mathbf{x}_{j+1}})$.

The original feature set proposed by Arras et al. [2007] includes a feature measuring the average speed of segments by associating their points over time. In this work the motion feature is not taken into accout as the detector is supposed to be able to detect both, standing and moving people.

2.3.2 Extended Set of Geometrical Features

In addition to the original features introduced above additional geometric properties are inspected. The impact of those features on the classification results is not known in advance. However, the idea is to provide as many features as possible and let the boosting algorithm learn their significance. The new features, being less intuitive as the original ones, are listed below.

14. Mean μ_i^ρ and variance σ_i^ρ of the difference in range of consecutive beams. Based on the range readings ρ_j these features describe the smoothness of the segments in another form,

$$\mu_i^\rho = \frac{1}{n_i - 1} \sum_{j=1}^{n_i-1} |\rho_j - \rho_{j+1}|, \qquad \sigma_i^\rho = \sqrt{\frac{1}{n_i - 2} \sum_{j=1}^{n_i-1} || \, |\rho_j - \rho_{j+1}| - \mu_i^\rho \, ||^2}.$$

15. Aspect ratio of the minimum and maximum standard deviation of the x- and y-coordinates, defined as

$$\Phi_i^\sigma = \begin{cases} {}^{(1+\sigma_x)}/_{(1+\sigma_y)} & \text{if } \sigma_x < \sigma_y, \\ {}^{(1+\sigma_y)}/_{(1+\sigma_x)} & \text{otherwise,} \end{cases}$$

where σ_x and σ_y are the standard deviations of the x- and y-coordinates, respectively. This feature provides a robust measure of the segments with compared to its depth. This ratio is assumed to be in certain limits for people.

16. Error of the quadratic and cubic b-spline calculated using the n_i points in segment S_i as control points. A B-spline curve of degree d for a collection of n_i control points $\mathbf{x}_k$ is defined by

$$P_d(u) = \sum_{k=1}^{n_i} \mathbf{x}_k B_{k,d}(u),$$

where $B_{k,d}(u)$ are the B-spline basis functions (blending functions) of degree d given by the Cox-de Boor recursion formula, that is

$$B_{k,1}(u) = \begin{cases} 1 & \text{if } u_k \leq u \leq i_{k+1}, \\ 0 & \text{otherwise,} \end{cases}$$

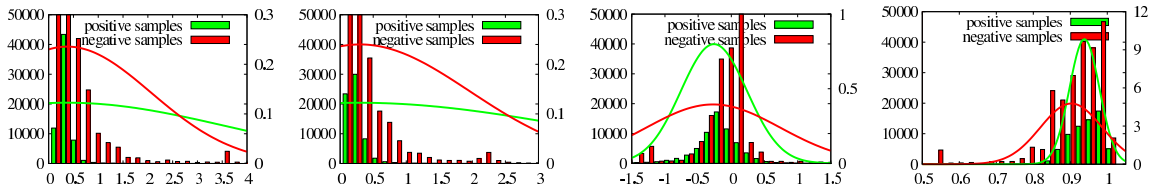
$$B_{k,d}(u) = \frac{u - u_k}{u_{k+d-1} - u_k} B_{k,d-1}(u) + \frac{u_{k+d} - u}{u_{k+d} - u_{k+1}} B_{k+1,d-1}(u).$$

The errors ε_i^d with $d \in \{2, 3\}$ are computed as

$$\varepsilon_i^d = \frac{1}{n_i} \sum_j \min_k (P_d(\mathbf{x}_k) - \mathbf{x}_j)^2.$$

17. Area of segment S_i measurement by the 2-dimensional region enclosed by the polygon defined by the laser end points. For a non-self-intersecting polygon with n_i vertices, the area is given by

$$A_i = \frac{1}{2} \sum_{j=1}^{n_i-1} (x_i y_{i+1} - x_{i+1} y_i).$$



(a) Mean angular diff. (b) Mean curvature. (c) 3^{rd} Fourier coefficient. (d) Aspect ratio Φ^{σ}.

Figure 2.2: Histograms and estimated Gaussian distributions of the four best features. The feature values of positive and negative samples are drawn in green or red, respectively. Fitted Gaussians visualize the variety in the data. Feature values and Gaussians are calculated on the manually annotated detections of the Freiburg city center data set.

18. Distance between the shortest and longest beam in segment S_i and their aspect ratio

$$\delta_i \;=\; \max_j(\rho_j) - \min_j(\rho_j), \qquad \Phi_i^{\rho} \;=\; \frac{\min_j(\rho_j)}{\max_j(\rho_j)}.$$

Former estimates the depth of an object while latter is another smoothness measure.

19. First three Fourier coefficients describing the shape of segment S_i. A sequence of n_i complex numbers $X_1, \ldots, X_{n_i}$ can be transformed into its frequency domain representation $Y_1, \ldots, Y_{n_i}$ using the discrete Fourier transform (DFT, Cochran et al. [1967]) formula

$$Y_k \;=\; \sum_{j=1}^{n_i} X_j e^{-2\pi i \frac{jk}{n_i}},$$

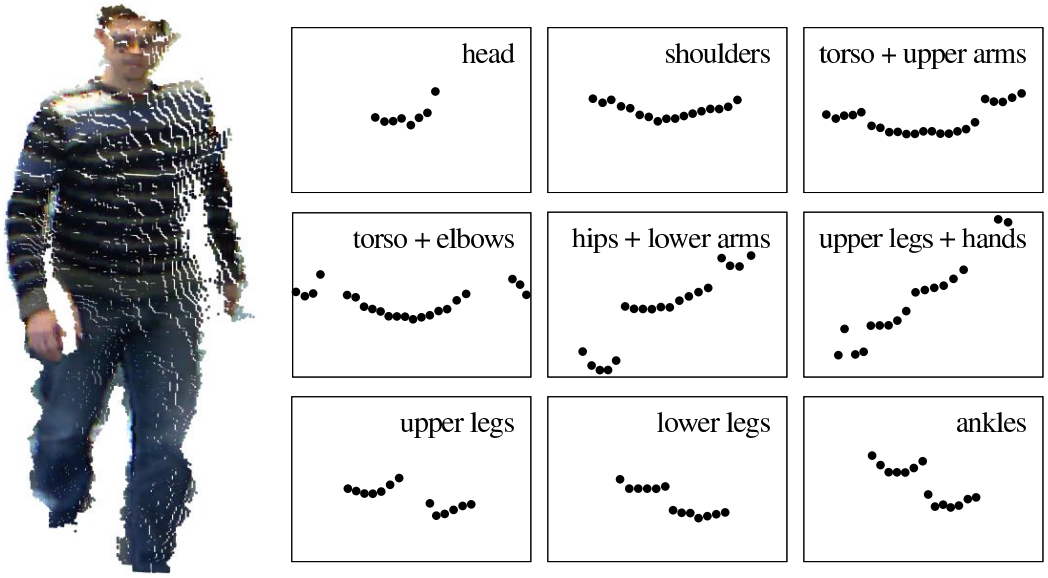
where Y_k are called Fourier coefficients. The complex input X_j is composed of the Cartesian coordinates of the points in segment S_i, hence $X_j = x_j + y_j i$. The first three Fourier coefficients Y_1, Y_2, and Y_3 are added as features.

Further Extensions

Additional features could, for example, investigate the reflection intensity of the laser beams. Carballo et al. [2010] proposed to take advantage of the average intensity, intensity variation, average difference of intensity, and the intensity uniformity to extend the set of geometric features. Furthermore, a multi-layered approach (as presented in Martínez Mozos et al. [2009]) is proposed. The work of Xavier et al. [2005] contains a feature to describe "arc-like" shapes for the purpose of leg detection.

2.4 Place Dependent People Detection

Learning a generic classifier implies the strong assumption that the appearance of people is unconditioned on their position in space and the current time of observation. Especially former does not hold for people detection in 2D range data where the appearance of people heavily depends on the relative height of the sensor (shown in Figure 2.3b). An uneven ground or a multi-level environment easily leads to observations of different body parts, thus training a generic classifier becomes challenging. Furthermore, due to a fixed sampling rate geometrical and statistical properties of objects change with their distance to the sensor. In other sensor modalities – e.g. cameras – people appear differently if the illumination conditions change. Both characteristics demand for place and



(a) 3D range data of a MS Kinect RGB-D camera.

(b) 2D range data recorded with a SICK LMS 291 laser range finder mounted at (from top left to bottom right) 1.7, 1.5, 1.3, 1.1, 0.9, 0.7, 0.5, 0.3, and 0.1 meter height above ground, respectively.

Figure 2.3: Visualization of 3D and 2D range data of a person at a distance of 2.5 meter. While the 3D range sensor observes the whole person the 2D sensor provides only few information. Depending on the height of the sensor different body parts are observed.

time-dependent classifiers that are experts for the respective circumstances. To implemented spatio-temporal detection, Eq. 2.1 is conditioned on the position $\mathbf{z}_i$ of the analyzed segment S_i and the current time t, hence $H(S_i \mid \mathbf{z}_i, t)$.

The ground conditions of the monitored area are usually unknown in advance, thus a detector trained on exact 2D positions might include wrong assumptions. However, the fact that the geometrical and statistical features depend on the range remains true. The generic detector is trained for people inside a maximum detection range ρ_{max} that depends on the situation – e.g. indoor or outdoor environment[5]. As the range of a segment S_i can be defined as the average range $\mu_i^\rho = \frac{1}{n_i}\sum_{j=1}^{n_i}\rho_j$ of the contained laser beams a set of range-specific classifiers $\widehat{H} = \{H_1, H_2, \dots\}$ specialized to detect people in intervals of certain minimum and maximum range values $I_p = [\rho_{min}^p, \rho_{max}^p)$ can be trained.

In the proposed approach the sensor field of view is divided uniformly[6] into a number of P intervals $\{I_1, \dots I_P\}$. Given a maximum detection range ρ_{max}, the number of intervals P, and the average range of a segment μ_i^ρ the index p of the specialized classifier H_p is determined by

$$p = \begin{cases} 1 + \frac{P\,\mu_i^\rho}{\rho_{max}} & \text{if } \mu_i^\rho < \rho_{max}\,, \\ P & \text{otherwise.} \end{cases} \tag{2.3}$$

Classifying S_i within the interval I_p using $\widehat{H}$ is performed by the sub-classifier specialized for the

[5] Using a SICK LMS 291 laser range finder with 0.5° resolution ρ_{max} is ∼20 meter. Beyond ρ_{max} the number of laser range readings on the surface of people drops dramatically making reliable people detection unfeasible.

[6] Other forms of subdividing the sensor field of view are possible. Additional experiments dividing the space into intervals of equal areas have been performed and yield comparable results.

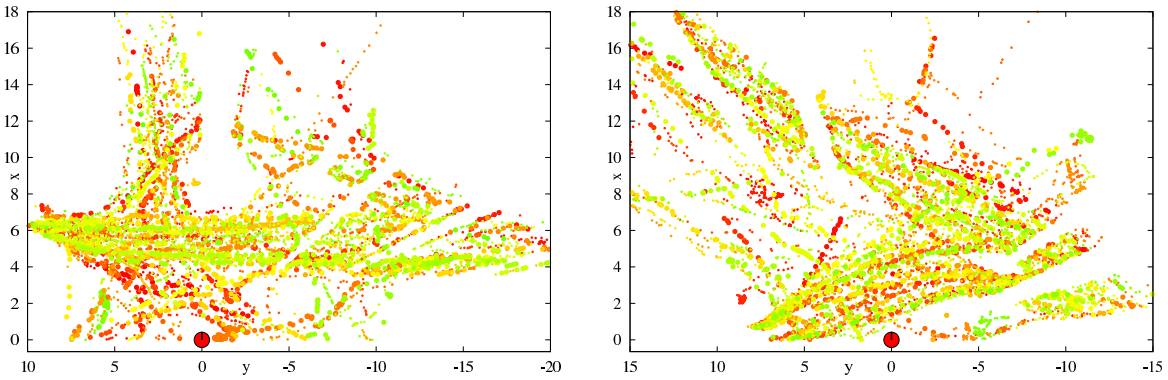


Figure 2.4: Visualization of the manually annotated detections of the Freiburg city center (*left*) and Freiburg main station data set (*right*), respectively. The laser range finder (marked with a red dot) is located at position $(0, 0)$ heading into the $+x$ direction. Detections of different people are drawn in slightly different colors to give an expression of the trajectories.

specific range, hence

$$\widehat{H}(S_i \mid \mu_i^{\rho}, t) \;=\; \begin{cases} H_p(S_i \mid t) & \text{if } \mu_i^{\rho} \leq \rho_{max}\,, \\ -1^{(7)} & \text{otherwise,} \end{cases} \tag{2.4}$$

where $H_p(S_i \mid t) = H_p(S_i) = \operatorname{sgn}\left(\sum_{j=1}^{k} \alpha_j^p \; h_j^p(S_i)\right)$ is the p^{th} sub-classifier trained on segments observed in the interval $I_p = [\rho_{min}^p, \rho_{max}^p)$. The time dependency is ignored in this work as laser range finders are robust against illumination changes. During training and classification each segment S_i is first categorized according to its average range μ_i^{ρ} and then passed to the classifiers H_p, specialized for the range interval I_p that satisfies $\rho_{min}^p \leq \mu_i^{\rho} < \rho_{max}^p$.

2.4.1 Segmentation of 2D Range Data

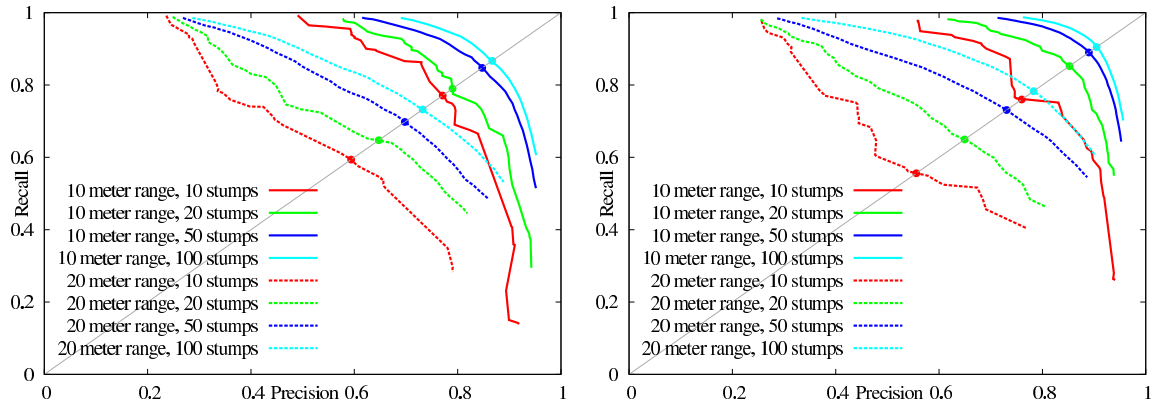
The training algorithm shown in algorithm 1 and the classification shown in Eq. 2.1 and Eq. 2.4 rely on segments of laser end points found by partitioning the set of laser range readings provided by the sensor into meaningful subsets. In contrast to the jump distance segmentation employed in Arras et al. [2007] it is proposes to use hierarchical clustering. While a simple criterion – like a large change of the range of consecutive beams – is sufficient to find the legs of people it was found to be inaccurate to find and group the points on both legs of a person or the torso and the arms together. Which body parts are observed exactly depends on the people's size and the height of the sensor[8] (see Figure 2.3). However, in this work, agglomerative hierarchical clustering with single linkage as presented by Day and Edelsbrunner [1984] is employed. The approach is able to bridge the gaps between different body parts of the observed people.

2.5 Experiments

In this section the introduced place-dependent people detector is evaluated and compared to state of the art on large outdoor data sets (see Figure 2.4) using different feature sets.

[7] Alternatively, an additional strong classifier H_{P+1} trained on examples observed in the interval $[\rho_{max}, \infty)$ can be added to allow detections in the complete sensor field of view. However, experiments have shown a very low detection accuracy beyond a maximum range of 20 meter, anyway.

[8] In the experiments the sensor was mounted at $\sim$0.85 meter observing hips and arms of adults.



(a) Original feature set.

(b) Extended feature set.

Figure 2.5: Precision recall (PR) curves and equal error rates (EER) of the detector proposed by Arras et al. using the original feature set (a) and extended feature set (b), respectively. The performances are compared under varying maximum detection ranges and number of weak classifiers. With 20 meter range and 50 decision stumps the EER rises from 69.8% (original set) to 73.1% (extended set). The rather small improvement shows that the definition of more geometrical features does not solve the detection problem in general.

2.5.1 Extended Feature Set

This part of the experimental evaluation analyzes the influence of the different feature sets, numbers of weak classifiers, and maximum detection range on the accuracy of people detection. Therefore, these parameters are varied. In the next subsection, the proposed cascade of specialized classifiers is compared to the boosted features detector by Arras et al. [2007] that consists of a unique strong classifier and serves as baseline.

First, the influence of the number of available geometrical descriptors is examined. The assumption is that a bigger variety in the (geometric) features enables the AdaBoost algorithm to train a "stronger" classifier with a higher classification accuracy. Further, it is assumed that a higher number of weak classifiers trained increases the detection accuracy as well as a lower maximum detection range. A classifier trained for a smaller range interval is expected to have a higher accuracy[9]. To compare the detection accuracy with varied numbers of maximum detection range, weak classifiers, and available features, respectively, precision recall (PR) curves are generated by counting the number of true positives (tp), false positives (fp), true negatives (tn), and false negatives (fn) using various detection thresholds θ. Therefore, the signum function in the AdaBoost classification scheme shown in Eq. 2.1 is replaced by the threshold θ, thus

$$H(\varepsilon) \;=\; \begin{cases} +1 & \text{if } \left(\sum_{j=1}^{k} \alpha_j \; h_j(\varepsilon)\right) \geq \theta \\ -1 & \text{otherwise.} \end{cases} \tag{2.5}$$

Finally, precision and recall values are determined using

$$\mathit{Precision} \;=\; \frac{tp}{tp+fp}\,, \qquad \mathit{Recall} \;=\; \frac{tp}{tp+fn}\,. \tag{2.6}$$

[9] Segments that are further away from the sensor than the specified maximum detection range are not taken into account to calculate the detection accuracy.

	original feature set	extended feature set
1	Mean angular difference μ^c (f_{13})	Mean angular difference μ^c (f_{13})
2	Mean curvature μ^κ (f_{12})	**3^{rd} Fourier coefficient (f_{19})**
3	Radius r^c of fitted circles (f_9)	**Aspect ratio Φ^ρ (f_{18})**
4	Distance to succeeding segment δ^{succ} (f_5)	**Area A (f_{17})**
5	Segment width ω (f_6)	**Aspect ratio Φ^σ (f_{15})**
6	Distance to previous segment δ^{prev} (f_4)	Distance to succeeding segment δ^{succ} (f_5)
7	Circularity c (f_8)	Linearity l (f_7)

Table 2.1: The best seven features selected by the unique strong classifier during the learning process employing the original and extended feature set, respectively. The new features added to the extended feature set are highlighted.

In addition, the equal error rate (EER) that is the rate at which both precision and recall errors are equal is calculated. The value of the EER can be easily obtained from the PR curve. The method with the highest EER performs best. The resulting PR curves and EER with $\theta \in [-1, +1]$ are shown in Figure 2.5.

The presented graphs confirm the assumptions made above. First, it is shown that the detection accuracy increases with the number of weak classifier[10] by taking advantage of the rich variety in the different features. Keeping a constant maximum detection range of 20 meter and using the original feature set the classifier trained on 10 features reaches an EER of 59.4% while the classifier trained on 100 features has an EER of 73.3% (+13.9%). The presented figures are carried out using 10-fold cross-validation making an overfitting effect very unlikely. An opposite impact on the EER is caused by the maximum detection range. Using 100 weak classifiers and the original feature set the EER of the detector trained for 10 meter is 86.7% while the EER of the detector trained for 20 meter drops to 73.3% (−13.4%). The impact of the extended feature set is slightly smaller but still valuable. Keeping a constant number of 100 weak classifiers and a maximum detection range of 20 meter, respectively, the EER increase from 73.3% to 78.3% (+5.0%). The seven best features selected by the AdaBoost algorithm are shown in Table 2.1. The distribution of the feature values of the two best features in each set are shown in Figure 2.2.

2.5.2 Place Dependend Detector

The analysis done so far revealed that the maximum detection range has a major impact on the accuracy of the unique classifier. Setting the maximum detection range above 10 meter the accuracy drops below 80.0%. But many applications rely on robust detections even in far ranges. Hence, a detector trained on multiple small range intervals with a high accuracy in each interval is required. In the following experiment, the unique classifier covering the complete detection range is compared to the proposed cascaded approach. The maximum detection range is set to 20 meter split into four equal intervals of 5 meter each. To train the specialized detectors an additional overlapping margin of 0.5 meter has been added to the intervals to select the positive and negative examples. The precision recall curves and equal error rates presented in Figure 2.6 (*left*) demonstrate the improvement of the cascade of specialized classifiers.

The experimental results show that the cascaded detector achieves always a higher accuracy than the baseline using either the original or the extended feature set. In detail, using 50 weak classifiers and the original feature set the EER increases from 69.8% to 74.1% (+4.3%). Employing the extended

[10] The proposed method employs decision stumps as weak classifiers.

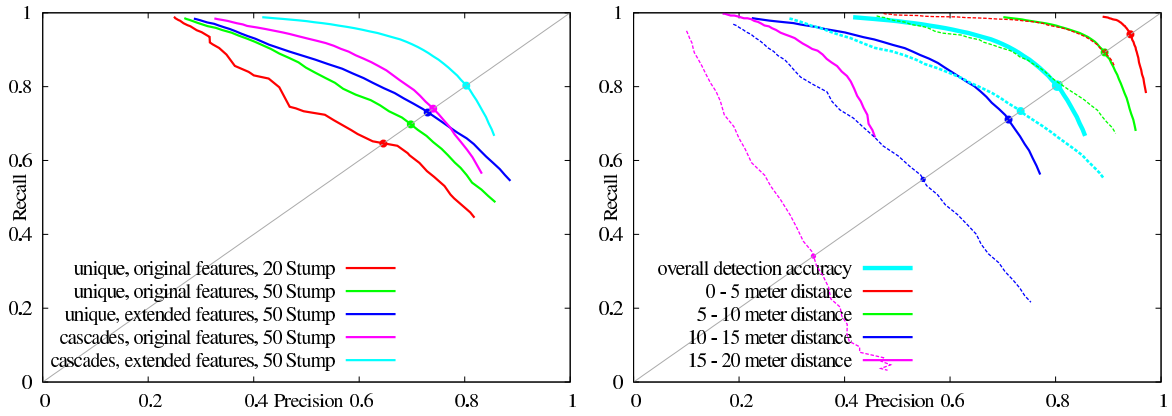


Figure 2.6: *Left:* PR curves and ERR of the unique classifier of Arras et al. (red, green, and blue) and the proposed cascade of classifiers (magenta and cyan) using different feature sets. In general the accuracy increases with the number of features and weak classifiers. The proposed method outperforms the unique classifiers. With extended feature set and 50 decision stumps the EER increases from 73.0% (unique classifier, shown in blue) to 80.3% (cascade of classifiers, shown in cyan).
Right: Comparative PR curves and EER of the best unique classifier (dashed lines) and the proposed cascade of classifiers (solid lines) at different detection ranges. The EERs of the classifiers covering the intervals of $[0-5)$, $[5-10)$, and $[10-15)$ meter increase by 4.8%, 8.5%, and 16.2%, respectively. Both detectors are trained using 50 weak classifiers.

feature set the EER raises from 73.0% to 80.3% (+7.3%). These figures show the benefit of the cascaded detector and the ability to train a high accuracy using the presented approach. However, an open question is the accuracy of the specialized classifiers. A high overall accuracy could be caused by an excellent classifier for close distances while the classifier for far ranges performs poorly. Therefore, a detailed comparison of the accuracies at specific distances has been carried out. The results are shown in Figure 2.6 (*right*).

The analysis of the precision recall curves and EERs shows that the accuracy of the specialized classifiers is higher in all cases. This is not surprising as the specialized classifiers have been trained in their specific intervals only. In other words, the training data was not polluted by examples that are very unlikely to occur in the classification process. An interesting aspect is that the increase in the accuracy raises with the distance. The accuracy of the classifier trained on the $[0-5)$ meter interval increases from 89.4% to 94.2% (+4.8%) while the accuracy of the $[10-15)$ meter classifier rises form 54.9% to 71.1% (+16.2%). These figures indicate that examples from people in close ranges have a major[11] impact on the training of the unique classifier. Further, the lower accuracy values in higher distances are explained by the decreased data density and increased noise of the laser range finder.

Some more details on the experimental results using 50 weak classifiers and a maximum detection range of 20 meter are given. Table 2.2 shows the numbers of true and false positives and negatives, respectively. These numbers have been determined using the standard classification method in Eq. 2.1 that is utilized during tracking. It can be seen that the numbers of true positives and true negatives increase with the extension of the feature set and the use of the cascaded classifier while the numbers of false positives and false negatives decrease. This shows that both proposed extensions lead to more accurate detections. Using both approaches the true positive and true negative rates rise from 82.0%

[11] Examples from people in close distance are assigned with higher weights.

	original approach				place dependent approach			
	original feature set		extended feature set		original feature set		original feature set	
true label	positive	negative	positive	negative	positive	negative	positive	negative
positive	5368	1177	5537	1008	5780	765	**5979**	**566**
(6545)	(82.0%)	(18.0%)	(84.6%)	(15.4%)	(88.3%)	(11.7%)	**(91.4%)**	**(8.6%)**
negative	4548	38683	4178	39053	3910	39321	**2535**	**40696**
(43231)	(10.5%)	(89.5%)	(9.7%)	(90.3%)	(9.0%)	(91.0%)	**(5.9%)**	**(94.1%)**

Table 2.2: Confusion matrix of classification results. The two rows correspond to the true number of positive and negative examples. The columns denote the number of examples classified as positive (person) and negative (no person), respectively, using different classification approaches. The presented approach using a cascade of classifiers and the extended set of features performs best. 91.4% of the positive detections belong to people.

to 91.4% and from 89.5% to 94.1%, respectively. This corresponds to an improvement of 11.5% and 5.1% in total. On the other hand, the false positive and false negative rates drop from 10.5% to 5.9% and from 18.6% to 8.6%. These are improvements of 43.8% and 53.8%, respectively. All values have been determined using 10-fold cross-validation.

2.5.3 Transferability to New Environments

The last experiment on people detection addresses the transferability of the proposed detector to new environments with data never seen in advance. More specifically, not only the new environment differs in general also the appearance of the observed people varies. For example, the people in the Freiburg inner city data set push strollers or bikes, while the people at the main station carry luggage. To analyze the change in the accuracy of the proposed cascaded people detector it is trained on the data recorded in the inner city and used to classify the data recorded at the main station. The accuracy of the detector trained with varying number of weak classifiers is shown in Figure 2.7(*left*). It is also interesting to knwon how well the range specialized detectors perform in comparison to the baseline approach of Arras et al. [2007]. Therefore, the range-dependent accuracies have been investigate. The results shown in Figure 2.7(*right*). All accuracy values are averaged over ten runs on different parts of the data sets using 10-fold cross validation.

The results show, as expected, that the accuracy decreases in comparison to the situations above where the training and testing data were collected in the same domain. However, the cascaded detector still reaches very high accuracy values. With a maximum detection range of 20 meter the EERs vary from 67.0% to 71.5% using 10 to 100 weak classifiers, respectively. The improvements with increased numbers of weak classifiers are rather small compared to the results obtained in Figure 2.5 and Figure 2.6. This is due to the fact that some appearances in the testing set – like people with hugh luggage – never appear during training.

The detailed analysis of the range-specific accuracy values identifies the major problem of both detectors using 50 weak classifiers. Especially at far distance only little information is available, thus the classification task is very hard. Even though the the specialized detector increases the accuracy at far distances (15 - 20 meter) from 10.5% to 15.5% it is very poor in general. At close distances (0 - 5 meter) the EERs are still very competitive reaching 89.3% and 91.7%. The overall detection accuracy of the proposed detectors is 68.5% and 69.6%. The results indicate that both approaches yields good generalization when employed for people detection in near ranges. With increasing size of the monitored area the proposed cascaded people detector exceeds the baseline method.

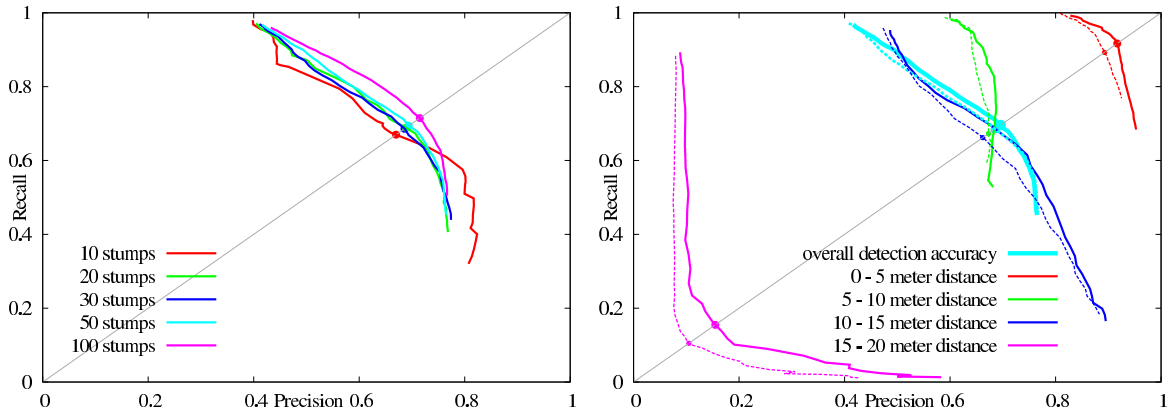


Figure 2.7: PR curves and EERs of the proposed people detector trained on the Freiburg city center data set and tested on the Freiburg main station data set. *Left*: Colors indicate different numbers of employed weak classifiers. *Right:* Solid lines show the accuracies at different range intervals of the proposed cascaded people detector. Dashed lines show the results of the baseline approach of Arras et al.. All values are average over ten runs on different parts of the data sets.

2.6 Conclusions

In this chapter, a novel, place-dependent approach for people detection in 2D laser range data was presented. The method is especially suitable for scenarios in which little information on the appearance of the desired objects is available. However, to improve people detection the boosted features people detector by Arras et al. [2007] is informed with spatial information. As the distance of objects is provided by the sensor the main idea is to train multiple detectors that are experts for specific range intervals instead of training a generic classifier that has to account for the huge variety in the data. This results in a cascade of range-specific boosted features detectors classifying people with respect to their distance to the sensor. As the spatial displacement is not the only reason for the variety in human appearance the set of employed geometrical and statistical features was extended to provide a larger and more expressive description of people.

Extensive experiments on large outdoor data sets show that the proposed approach exceeds the state-of-the-art baseline method. Especially in far distances where little information is available the specialized detectors are still very accurate. Using the proposed approach the detection accuracy in distances of more than 15 meters increases by 16%. The overall detection accuracy is 80% with 91% true positives and 94% true negatives, respectively. The influence of the additional geometrical features presented in this work is similar. Using the extended set of features the detection accuracy increases by 6%. The detector transferred to a new environment has still an accuracy above 70% in general and reaches 91% in close distances. This demonstrates that the results are not due to overfitting and that the proposed detector generalizes well to new environments with never seen data.

Future work on people detection includes the analysis of strategies that learn the spatial partitioning (like range intervals or even more complex subdivisions), segmentation thresholds, and the boosted features, jointly. Additionally, the use of more accurate domain partitioning weak classifiers that outperform the employed decision stumps will be investigated. Furthermore, temporal information not considered in this work provides additional a priori information on the appearance of people, thus a spatio-temporal informed people detector should further improve detection accuracy.

3 Multi-Hypothesis Tracking of People

Detecting people is a key component for robots sharing an environment with humans. Social robots, indeed, need to be capable of more. Their fundamental skill is to maintain the positions and behaviors of individuals over time to allow a robust and reliable human-robot interaction. To this end people detections are associate over time using a technique called people tracking.

The most challenging task in people tracking is to find the correct correspondences between known targets and incoming observations. The so called data association problem becomes especially hard when targets have identical appearance making an identification impossible. During the last decades many different data association technique have been developed. This work does not focus on the development of a completely new solution to this problem but investigates how spatial, temporal, and social information can be learned, modeled, and integrated into tracking to guide the data association process in different ways.

The impact of spatio-temporal and social information on tracking is analyzed using the multi-hypotheses tracking (MHT) approach. This chapter presents the original formulation of the MHT, reviews the progress made in the past, and provides the theoretical background of the extensions required to integrate the models proposed in this thesis. Furthermore, efficient implementation strategies on hypotheses generation and management are illustrated.

This chapter is structured as follows. Sections 3.1 and 3.2 introduce MHT tracking and the notations used in the remainder of this thesis. In section 3.3 the original formulation of the MHT is explained in detail. The extension with explicit deletion labels and explicit occlusion labels is presented in section 3.4 and section 3.5, respectively. The integration of spatio-temporal target priors is illustrated in section 3.6 while section 3.7 depicts the spatio-temporal target probabilities. A short overview of efficient implementation strategies is given in section 3.8 follow by section 3.9 that concludes the paper.

3.1 Introduction

In the original paper by Reid [1979], the **M**ulti-**H**ypothesis **T**racking (MHT) approach, an algorithm for tracking multiple targets was presented first. The developed algorithm is capable to interpret observations as *detected* with existing tracks, *new tracks*, or *false alarms* and to interpret tracks as *detected* when they match with an observation or *not detected*. The states of all targets are estimated using individual Kalman filters. The key development of the approach is a method to calculate the probabilities of various data association hypotheses, given the approach its name. In Cox and Hingorani [1996] the MHT is extended with the additional interpretation of tracks as *deleted*. This allows to model obsolete tracks, explicitly, and to deal with disappearing targets making the approach suitable for scenarios in which target may leave the sensor field of view. Furthermore, an efficient implementation of the MHT using an algorithm introduced by Murty [1968] to calculate the k-best data association hypotheses in polynomial time is presented. In 2008 Arras et al. extended the MHT framework to multiple track interpretation events including *occlusions* in addition to *detections* and *deletions*. Theoretically, an arbitrary number of track and observation labels is possible. The benefit of this approach is that the occlusion probabilities of tracks can be adjusted, individually, making the

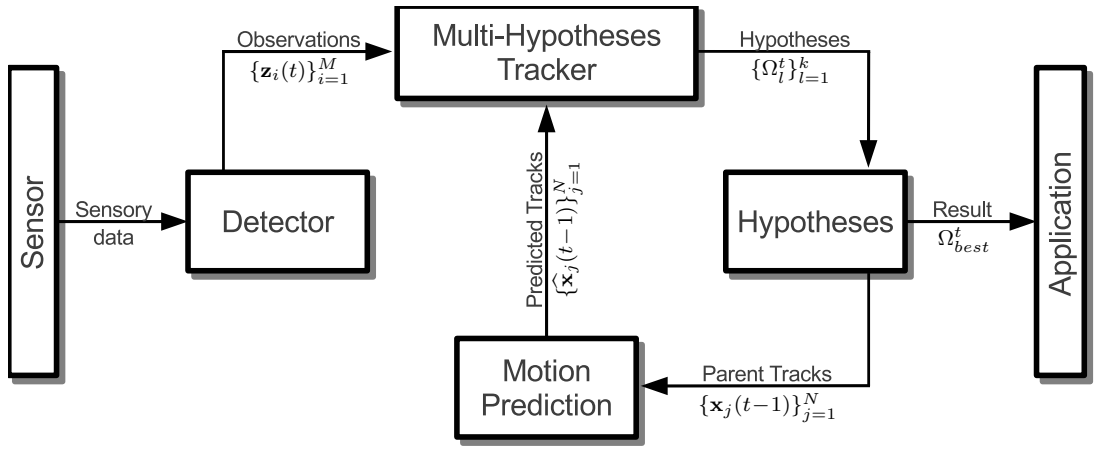


Figure 3.1: Illustration of the Multi-Hypothesis Tracking framework. The main components involved in tracking-by-detection are detection, motion prediction, and data association (here realized by the MHT).

approach more reliable in case of lengthy occlusion events. This is especially important as occlusions occur frequently when observing the environment with sensor mounted at the robot.

The contributions of this work include the extension of the regular MHT with human-specific spatial priors. This allows to model place-dependent *new track* and *false alarm* probabilities. As shown in Chapter 5 and Chapter 6 those priors enable the integration of – modeled or learned – human-specific behaviors. Another extension addresses track-specific *occlusion* and *deletion* probabilities. As analyzed in Chapter 5 and Chapter 7 these place-dependent adaptations enable tracking to account for spatio-temporal and social influences on the tracking event probabilities.

3.2 Notations

This section introduces the basic notations used in this work. Let $\mathcal{Z}^t$ be the sequence $\{\mathcal{Z}(0), \ldots, \mathcal{Z}(t)\}$ of all sensor readings from the beginning of the tracking process (assumed to be at $t = 0$) to the current time t. Further, each sensor reading $\mathcal{Z}(t)$ provides a set of measurements $\mathcal{B} = \{\mathbf{b}_1(t), \ldots, \mathbf{b}_{B_t}(t)\}$ and a set of observations $\mathcal{Z}(t) = \{\mathbf{z}_1(t), \ldots, \mathbf{z}_{M_t}(t)\}$ containing detected targets or misdetections caused by clutter. The terms *sensor reading*, *measurement*, *observation*, and *detection* can easily be confused but are not interchangeable in general. Actually, the denotation *sensor reading* describes the complete sensor information provided at the same time (e.g. a camera image or a complete range scan). In contrast the term *measurement* denotes a unit of the raw sensor information (e.g. a pixel of a camera image or a beam of a laser range finder). As these measurements contain only little information they are usually combined and filtered by a chain of clustering, segmentation, and detection algorithms, respectively. The result of these processing steps is the set of *observations*. The term *detections* is used for observations that belong to true targets. Unfortunately, no detection algorithm can provide this information reliably. The number of measurements B_t and observations M_t is assumed to be known in advance. Usually, B_t is constant over time (e.g. constant number of camera pixels or laser range finder beams) while M_t may vary. Further, each sensor reading $\mathcal{Z}(t)$ is assumed to be independent from the set of previous readings $\mathcal{Z}^{t-1} = \{\mathcal{Z}(0), \ldots, \mathcal{Z}(t-1)\}$.

Besides the observations $\mathbf{z}_i$ the second important entities used in tracking are the tracks $\mathbf{x}_j$ themselves. The set of all tracks present at time t is denoted by $\{\mathbf{x}_1(t), \ldots, \mathbf{x}_{N_t}(t)\}$ with N_t being the

Notation	Description
t	Index of the current time frame
	Observation related
$\mathcal{Z}(t)$	Sensor reading at time t containing a set of observations
$\mathcal{Z}^{t-1}$, $\mathcal{Z}^{t}$	Set of sensor readings from $t=0$ to $t-1$ and t, respectively
$\mathbf{z}_i(t)$, $\mathbf{z}_i$	$\mathbf{i}^{th}$ observation at time t
M_t	Number of observations at time t
V	Volume of the sensor field of view
	Hypothesis and Track related
Ω_l^t	$\mathbf{l}^{th}$-best hypothesis at time t
$\Omega_{p(l)}^{t-1}$	Parent hypothesis of the l^{th}-best hypothesis at time $t-1$
$\mathbf{x}_j^l(t)$, $\mathbf{x}_j(t)$, $\mathbf{x}_j$	$\mathbf{j}^{th}$ track in the l^{th}-best hypothesis at time t
N_t^l, N_t	Number of tracks in hypothesis l at time t
$N_{det}^l(t)$, N_{det}	Number of detected targets in hypothesis l at time t
$N_{new}^l(t)$, N_{new}	Number of new targets in hypothesis l at time t
$N_{fal}^l(t)$, N_{fal}	Number of false alarms in hypothesis l at time t
p_{det}, p_{occ}, p_{del}	General probabilities of detected, occluded, and deleted tracks
$p_{det}(\mathbf{x}_j(t-1), t)$	Detection probability at the position of track $\mathbf{x}_j(t-1)$ at time t
$p_{occ}(\mathbf{x}_j(t-1), t)$	Occlusion probability at the position of track $\mathbf{x}_j(t-1)$ at time t
$p_{del}(\mathbf{x}_j(t-1), t)$	Deletion probability at the position of track $\mathbf{x}_j(t-1)$ at time t
λ_{new}, λ_{fal}	General expected number of new track and false alarm events
$\lambda_{new}(\mathbf{z}_i(t), t)$	Expected number of new track events at the position of $\mathbf{z}_i$ at time t
$\lambda_{fal}(\mathbf{z}_i(t), t)$	Expected number of false alarm events at the position of $\mathbf{z}_i$ at time t
	Assignment related
$\psi_l(t)$	Set of assignments defining the l^{th}-best hypothesis at time t
$\tau(\psi_l(t))$	Set of indicator variables of observations originating from known tracks
$\tau^i(\psi_l(t))$, τ^i	Indicator that $\mathbf{z}_i(t)$ originates from a previously known track
$\phi(\psi_l(t))$	Set of indicator variables of observations from clutter
$\phi^i(\psi_l(t))$, ϕ^i	Indicator that $\mathbf{z}_i(t)$ is a false alarm
$\nu(\psi_l(t))$	Set of indicator variables of observations from new tracks
$\nu^i(\psi_l(t))$, ν^i	Indicator that $\mathbf{z}_i(t)$ creates a new track
$\delta(\psi_l(t))$	Set of indicator variables of detected tracks
$\delta^j(\psi_l(t))$, δ^j	Indicator that $\mathbf{x}_j(t-1)$ is detected in the current frame t
$\chi(\psi_l(t))$	Set of indicator variables of deleted tracks
$\chi^j(\psi_l(t))$, χ^j	Indicator that $\mathbf{x}_j(t-1)$ is deleted in the current frame t
$\omega(\psi_l(t))$	Set of indicator variables of occluded tracks
$\omega^j(\psi_l(t))$, ω^j	Indicator that $\mathbf{x}_j(t-1)$ is occluded in the current frame t

Table 3.1: Notations used in the **M**ulti-**H**ypothesis **T**racking framework.

number of tracks existing at time t. Each track $\mathbf{x}_j(t)$ represent both the current state of the corresponding target at time t as well as the targets complete history. The trajectory of each target can be derived from its consecutive states over time.

As mentioned in the introduction, the MHT reasons about various evolutions of world. Each of these possible explanations is called a hypothesis Ω_l^t, where t again denotes the current time and l the index or ranking of the considered hypothesis. Note, each hypothesis describes the interpretation of all observations and evolution of all tracks in the environment given the sequence of sensor readings $\mathcal{Z}^t$. However, each hypothesis Ω_l^t arise from a parent hypothesis $\Omega_{p(l)}^{t-1}$ and a set of assignments $\psi_l(t)$ that reasons about the N_{t-1} tracks from the previous and the M_t observations from the current time step. More precisely, an assignment declares $N_{det}^l(t)$ observations emerging from previously existing tracks, $N_{new}^l(t)$ observations from new tracks, and $N_{fal}^l(t)$ observations from false alarms in clutter, respectively. On the other side, each assignment declares $N_{det}^l(t)$ tracks as detected (or matched), $N_{occ}^l(t)$ as occluded, and $N_{del}^l(t)$ tracks are obsolete, respectively. For the sake of compactness of notation, the index l of the hypothesis and index t are omitted for the previously mentioned numbers. In the following they are notated as N_{det}, N_{occ}, N_{del}, N_{new}, and N_{fal}, respectively[12]. As shown in the next section, each hypothesis receives a probability denoted as $p(\Omega_l^t \mid \mathcal{Z}^t)$ to evaluate its quality. To achieve the probability of an assignment set $\psi_l(t)$ each of the assignment events – e.g. matching, new track, or occlusion – obtains its own probability, likelihood, or expected occurrence rate called, p_{det}, p_{occ}, p_{del}, λ_{new}, and λ_{fal}, respectively. An overview of the notations used in the MHT framework are presented in Table 3.1. A visualization of the MHT framework is provided in Figure 3.1.

3.3 Original Formulation of the MHT

The original formulation of the MHT by Reid [1979] expressed the generation of new hypotheses as follows. Given a set of incoming observations $\mathcal{Z}(t)$ a new hypothesis Ω_l^t is created from a set of assignments[13] $\psi_l(t)$ and a previous hypothesis $\Omega_{p(l)}^{t-1}$ based on all measurements up to time $t-1$,

$$\Omega_l^t \;=\; \left\{\psi_l(t),\; \Omega_{p(l)}^{t-1}\right\}. \tag{3.1}$$

The current set of assignments $\psi_l(t)$ assigns each observation to one of the observation labels *detected*, *new track*, or *false alarm*, respectively. Furthermore, $\psi_l(t)$ contains the specific information of N_{det} observations that arise from previously established tracks, N_{fal} observations that are identified as false alarms from clutter, and finally N_{new} observations that are marked as new tracks. As all observations have to be assigned to one of these track labels, the condition

$$N_{det} + N_{fal} + N_{new} \;= M_t \tag{3.2}$$

[12] Although the index l of the hypothesis is omitted for N_{det}, N_{occ}, N_{del}, N_{new}, and N_{fal} it is important to remember that these numbers vary in different hypotheses.

[13] In related literature the current set of assignment is also denotes as *current association event*.

must be hold. To provide the mentioned information in a formal way the following assignment set dependent indicator variables are introduced

$$\begin{aligned}
\tau_i &= \tau_i[\psi_l(t)] = \begin{cases} 1 & \text{if observation } \mathbf{z}_i(t) \text{ originates from a previously known track,} \\ 0 & \text{otherwise,} \end{cases} \\
\phi_i &= \phi_i[\psi_l(t)] = \begin{cases} 1 & \text{if observation } \mathbf{z}_i(t) \text{ is deemed as false alarm,} \\ 0 & \text{otherwise,} \end{cases} \\
\nu_i &= \nu_i[\psi_l(t)] = \begin{cases} 1 & \text{if observation } \mathbf{z}_i(t) \text{ is marked as new track,}^{14} \\ 0 & \text{otherwise,} \end{cases} \\
\delta_j &= \delta_j[\psi_l(t)] = \begin{cases} 1 & \text{if track } \mathbf{x}_j(t-1) \text{ from } \Omega_{p(l)}^{t-1} \text{ is detected in the current frame,} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.3}$$

All tracks assigned to an observation are called *detected* or *matched* tracks. For any observation i that arises from a previously existing track j the additional information of the track index is encoded in the assignment set as well

$$\Delta_i = \Delta_i[\psi_l(t)] = \begin{cases} j & \text{if observation } i \text{ originates from track } j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

From these defined indicator variables the number of detected tracks in $\psi_l(t)$ can be obtained as $N_{det} = \sum_{i=1}^{M_t} \tau_i$, the number of false alarms as $N_{fal} = \sum_{i=1}^{M_t} \phi_i$, and the number of new tracks as $N_{new} = \sum_{i=1}^{M_t} \nu_i$ or $N_{new} = M_t - N_{det} - N_{fal}$, respectively. Equally, the number of detected tracks can also be obtained as $N_{det} = \sum_{i=1}^{N_{t-1}} \delta_i$ as only bi-unique observation to track assignments are allowed. Strictly speaking, the indicator variable δ_j is redundant. However, it is introduced to ease notations.

As mentioned in the introduction the number of hypotheses grows exponentially over time hence reasoning about all hypotheses is impossible. Thus an approach to maintain only the most accurate or important hypotheses and to prune the worse once is required. To obtain a ranking of the hypotheses and guide the pruning algorithm each hypothesis receives a probability $p(\Omega_l^t \mid \mathcal{Z}^t)$ that is introduced in the following.

The probability $p(\Omega_l^t \mid \mathcal{Z}^t)$ of the hypothesis Ω_l^t can then be calculated from its parent hypothesis $\Omega_{p(l)}^{t-1}$ and the given assignment set $\psi_l(t)$ using Bayes' rule, hence

$$\begin{aligned}
p(\Omega_l^t \mid \mathcal{Z}^t) &= p(\psi_l(t),\ \Omega_{p(l)}^{t-1} \mid \mathcal{Z}(t),\ \mathcal{Z}^{t-1}) \\
&= \eta\ p(\mathcal{Z}(t) \mid \psi_l(t),\ \Omega_{p(l)}^{t-1},\ \mathcal{Z}^{t-1}) \\
&\quad p(\psi_l(t) \mid \Omega_{p(l)}^{t-1},\ \mathcal{Z}^{t-1})\ \ p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}).
\end{aligned} \tag{3.5}$$

The equation consists of four terms. The normalizer η, the measurement likelihood, and the probability of the assignment set. The rightmost term represents the probability of the parent hypothesis. The measurement likelihood and assignment set probability are discussed in more detail in the next sections.

[14] The assignment set is kind of overdetermined by the defined indicator variables as an observation is implicitly marked as new track if it is neither marked as matched nor as false alarm.

3.3.1 Measurement Likelihood

The measurement likelihood (also likelihood function of the current assignment set) specifies how well the sensor data describes the set of assignments given the parent hypothesis and all previous sensor readings. Under the assumption that an observation $\mathbf{z}_i(t)$ associated to an existing track $\mathbf{x}_j(t-1)$ has a Gaussian pdf centered around the measurement prediction $\hat{\mathbf{z}}_j(t)$ with innovation covariance matrix $S_{i,j}(t)$ the measurement likelihood of a single matched observation is $\mathcal{N}(\mathbf{z}_i(t)) := \mathcal{N}(\mathbf{z}_i(t);\, \hat{\mathbf{z}}_j(t),\, S_{i,j}(t))$. Further, the pdf of new tracks and false alarms is assumed to be uniform in the surveillance volume[15] V of the sensor field of view with probability V^{-1}. Thus the measurement likelihood in Eq. 3.5 – the likelihood function of $\psi_l(t)$ – yields

$$\begin{aligned} p(\mathcal{Z}(t) \mid \psi_l(t),\, \Omega_{p(l)}^{t-1},\, \mathcal{Z}^{t-1}) &= \prod_{i=1}^{M_t} \left[\mathcal{N}(\mathbf{z}_i(t))^{\tau_i}\; V^{-\phi_i}\; V^{-\nu_i}\right] \\ &= V^{-(N_{fal}+N_{new})} \prod_{i=1}^{M_t} \mathcal{N}(\mathbf{z}_i(t))^{\tau_i}, \end{aligned} \tag{3.6}$$

where $M_t = N_{det} + N_{new} + N_{fal}$ and the indicator variables τ, ϕ, and ν introduced in Eq. 3.3 have been used for compactness of notation.

3.3.2 Prior Assignment Probability

The prior probability of the assignment set $\psi_l(t)$ is obtained next. Using the indicator variables τ, ϕ, and ν introduced in Eq. 3.3 the prior assignment probability can be rewritten as

$$\begin{aligned} P(\psi_l(t) \mid \Omega_{p(l)}^{t-1},\, \mathcal{Z}^{t-1}) &= P(\psi_l(t),\, \tau[\psi_l(t)],\, \phi[\psi_l(t)],\, \nu[\psi_l(t)] \mid \Omega_{p(l)}^{t-1},\, \mathcal{Z}^{t-1}) \\ &= P\left(\psi_l(t) \mid \tau[\psi_l(t)],\, \phi[\psi_l(t)],\, \nu[\psi_l(t)]\right) \\ &\quad P\left(\tau[\psi_l(t)],\, \phi[\psi_l(t)],\, \nu[\psi_l(t)] \mid \Omega_{p(l)}^{t-1}\right), \end{aligned} \tag{3.7}$$

where the irrelevant conditions above have been dropped. The combinatorics of the current assignment set – $P\left(\psi_l(t) \mid \tau[\psi_l(t)],\, \phi[\psi_l(t)],\, \nu[\psi_l(t)]\right)$ depends on the number of assignment sets that have exactly the same numbers of N_{det} detected tracks, N_{fal} false alarms, and N_{new} new tracks – is given by two aspects. First, the number of possible configurations that partition the number of available observation M_t into a specific number of detected targets N_{det}, false alarms from clutter N_{fal}, and new targets N_{new}, respectively. And second, the number of permutations to assign the N_{det} detections to the N_{t-1} previously existing tracks.

The number of configurations is given by a product of combinations to select N_{det} from M_t observations as detected, N_{fal} from the remaining $M_t - N_{det}$ observations as false alarms, and finally N_{new} from $M_t - N_{det} - N_{fal}$ observations as new tracks. Hence, the number of configurations is given by

$$\begin{aligned} &\begin{pmatrix} M_t \\ N_{det} \end{pmatrix} \begin{pmatrix} M_t - N_{det} \\ N_{fal} \end{pmatrix} \begin{pmatrix} M_t - N_{det} - N_{fal} \\ N_{new} \end{pmatrix} \\ &= \frac{M_t!}{N_{det}!\,(M_t - N_{det})!}\, \frac{(M_t - N_{det})!}{N_{fal}!\, N_{new}!}\, \frac{N_{new}!}{N_{new}!} \\ &= \frac{M_t!}{N_{det}!\, N_{fal}!\, N_{new}!} \end{aligned} \tag{3.8}$$

[15] Although the volume of the sensor field of view varies due to occlusions and sensor noise it is often assumed to be constant over time.

The number of permutations to assign the N_{det} detections to the N_{t-1} previously existing tracks is given by

$$\frac{N_{t-1}!}{(N_{t-1} - N_{det})!} \tag{3.9}$$

Combining Eq. 3.8 and Eq. 3.9 and assuming each of theses events equally likely, the first term of Eq. 3.7 is defined as

$$P\left(\psi_l(t) \mid \tau[\psi_l(t)],\ \phi[\psi_l(t)],\ \nu[\psi_l(t)]\right) \ = \ \frac{N_{det}!\ N_{fal}!\ N_{new}!}{M_t!}\ \frac{(N_{t-1} - N_{det})!}{N_{t-1}!} \tag{3.10}$$

To model the last term of Eq. 3.7 it is assumed that the number of new tracks N_{new} and the number of false alarms N_{fal} both follow a Poisson distribution[16] with expected number of events $\lambda_{new}V$ and $\lambda_{fal}V$ in the observation volume V, respectively. Further, it is assumed that the number of previously known tracks (N_{t-1}) that are detected (N_{det}) is given by a binomial distribution. With these assumptions, the probability of the numbers N_{det}, N_{fal}, and N_{new} given $\Omega_{p(l)}^{t-1}$ is

$$\begin{aligned} P\left(\tau[\psi_l(t)],\ \phi[\psi_l(t)],\ \nu[\psi_l(t)] \mid \Omega_{p(l)}^{t-1}\right) \ &= \ \begin{pmatrix} N_{t-1} \\ N_{det} \end{pmatrix} p_{det}^{N_{det}}\ (1 - p_{det})^{(N_{t-1} - N_{det})} \\ &\quad P_{\lambda_{fal}V}(N_{fal})\ P_{\lambda_{new}V}(N_{new}) \\ &= \frac{N_{t-1}!}{(N_{t-1} - N_{det})!\ N_{det}!}\ p_{det}^{N_{det}}\ (1 - p_{det})^{(N_{t-1} - N_{det})} \\ &\quad \frac{\lambda_{fal}^{N_{fal}} V^{N_{fal}}}{N_{fal}!}\ e^{-\lambda_{fal}V}\ \frac{\lambda_{new}^{N_{new}} V^{N_{new}}}{N_{new}!}\ e^{-\lambda_{new}V}. \end{aligned} \tag{3.11}$$

Substituting Eq. 3.10 and Eq. 3.11 into Eq. 3.7 the prior assignment probability is given as

$$\begin{aligned} p(\psi_l(t) \mid \Omega_{p(l)}^{t-1},\ \mathcal{Z}^{t-1}) \ &= \ \frac{N_{det}!\ N_{fal}!\ N_{new}!}{M_t!}\ \frac{(N_{t-1} - N_{det})!}{N_{t-1}!} \\ &\quad \frac{N_{t-1}!}{(N_{t-1} - N_{det})!\ N_{det}!}\ p_{det}^{N_{det}}\ (1 - p_{det})^{(N_{t-1} - N_{det})} \\ &\quad \frac{\lambda_{fal}^{N_{fal}} V^{N_{fal}}}{N_{fal}!}\ e^{-\lambda_{fal}V}\ \frac{\lambda_{new}^{N_{new}} V^{N_{new}}}{N_{new}!}\ e^{-\lambda_{new}V} \\ &= \ \eta\ p_{det}^{N_{det}}\ (1 - p_{det})^{(N_{t-1} - N_{det})}\ \lambda_{fal}^{N_{fal}} V^{N_{fal}}\ \lambda_{new}^{N_{new}} V^{N_{new}}, \end{aligned} \tag{3.12}$$

where η is a normalizer that combines the constant values $\frac{1}{M_t!}$, $e^{-\lambda_{fal}V}$, and $e^{-\lambda_{new}V}$ that are independent from the specific assignment set $\psi_l(t)$.

[16] The Poisson distribution with parameters $\lambda > 0$ (expected number of events) and n (number of occurrences) is defined as $P_\lambda(n) = \frac{\lambda^n}{n!}e^{-\lambda}$.

3.3.3 Recursive Hypothesis Probability

Finally, by substituting the results of Eq. 3.6 and Eq. 3.12 into Eq. 3.5 the recursive expression of the probability of a hypothesis as it was defined by Reid [1979] yields

$$\boxed{\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t) \;=\; \eta\, \lambda_{fal}^{N_{fal}}\, \lambda_{new}^{N_{new}} \prod_{i=1}^{M_t} \mathcal{N}(\mathbf{z}_i(t))^{\tau_i} \\ p_{det}^{N_{det}}\; (1-p_{det})^{(N-N_{det})}\; p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \end{aligned}} \tag{3.13}$$

3.4 Explicit Deletion Labels

Cox and Hingorani [1996] extended the original formulation of the MHT by introducing *explicit deletion* labels. Simultaneously, occluded tracks are modeled implicitly as tracks that are neither detected nor deleted. Hence, the assignment set $\psi_l(t)$ assigns each observation to one of the observation labels *detected*, *new track*, or *false alarm*, and each track to one of the track labels *detected*, *deleted*, or *neither of them (occluded)* so describing the meaning of observations and the evolution of tracks. The set of indicator variables defined in 3.3 must[17] be extended by

$$\chi_j \;=\; \chi_j[\psi_l(t)] \;=\; \begin{cases} 1 & \text{if track } \mathbf{x}_j(t-1) \text{ from } \Omega_{p(l)}^{t-1} \text{ is deleted in the current frame,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.14}$$

to encode the deleted tracks. Further, using χ the number of deleted tracks $N_{del} = \sum_{j=1}^{N_{t-1}} \chi_j$ can be obtained. Combining the indicators from 3.3 and 3.14 the joint probability of the prior assignment likelihood – as it was defined in Eq. 3.7 – can now be rewritten as

$$\begin{aligned} P(\psi_l(t) \mid \Omega_{p(l)}^{t-1},\; \mathcal{Z}^{t-1}) \;&=\; P\left(\psi_l(t) \mid \tau[\psi_l(t)],\; \phi[\psi_l(t)],\; \nu[\psi_l(t)],\; \chi[\psi_l(t)]\right) \\ &\quad P\left(\tau[\psi_l(t)],\; \phi[\psi_l(t)],\; \nu[\psi_l(t)],\; \chi[\psi_l(t)] \mid \Omega_{p(l)}^{t-1}\right). \end{aligned} \tag{3.15}$$

The combinatorics change equally as the number of possible configurations does not only depend on the number of permutations to assign N_{det} tracks to the N_{t-1} previously known tracks but also on the number of combinations to select N_{del} tracks from the remaining $(N_{t-1} - N_{det})$ tracks as deleted. Thus Eq. 3.9 changes to

$$\frac{N_{t-1}!}{(N_{t-1}-N_{det})!} \begin{pmatrix} N_{t-1}-N_{det} \\ N_{del} \end{pmatrix} \;=\; \frac{N_{t-1}!}{(N_{t-1}-N_{det}-N_{del})!\; N_{del}!}, \tag{3.16}$$

where the first term on the l.h.s yields the number of permutations (already known from Eq. 3.9) and the second term yields the combinations. Combining these new results with Eq. 3.8 the prior probability of an assignment set depending on the combinatorics – as previously Eq. 3.10 – is

$$\begin{aligned} P\left(\psi_l(t) \mid \tau[\psi_l(t)],\; \phi[\psi_l(t)],\; \nu[\psi_l(t)],\; \chi[\psi_l(t)]\right) \;= \\ \frac{N_{det}!\; N_{fal}!\; N_{new}!\; (N_{t-1}-N_{det}-N_{del})!\; N_{del}!}{M_t!\; N_{t-1}!}. \end{aligned} \tag{3.17}$$

[17] The set of indicators needs to be extended by χ_j as the number of deleted tracks N_{del} can not be inferred from the assignments of the observations.

Under the assumption that both the number of detected tracks (N_{det}) and the number of deleted tracks (N_{del}) follow a binomial distribution Eq. 3.11 changes to

$$P\left(\tau[\psi_l(t)],\ \phi[\psi_l(t)],\ \nu[\psi_l(t)],\ \chi[\psi_l(t)]\ \mid\ \Omega_{p(l)}^{t-1}\right) \ =\ P_{\lambda_{fal}V}(N_{fal})\ P_{\lambda_{new}V}(N_{new}) \tag{3.18}$$

$$\begin{pmatrix} N_{t-1} \\ N_{det} \end{pmatrix} p_{det}^{N_{det}}\ (1-p_{det})^{(N_{t-1}-N_{det})} \begin{pmatrix} N_{t-1}-N_{det} \\ N_{del} \end{pmatrix} p_{del}^{N_{del}}\ (1-p_{del})^{(N_{t-1}-N_{det}-N_{del})}$$

$$= \frac{\lambda_{fal}^{N_{fal}} V^{N_{fal}}}{N_{fal}!}\ e^{-\lambda_{fal}V}\ \frac{\lambda_{new}^{N_{new}} V^{N_{new}}}{N_{new}!}\ e^{-\lambda_{new}V}\ \frac{N_{t-1}!\ (N_{t-1}-N_{det}-N_{del})!\ N_{del}!}{N_{det}!}$$

$$p_{det}^{N_{det}}\ (1-p_{det})^{(N_{t-1}-N_{det})}\ p_{del}^{N_{del}}\ (1-p_{del})^{(N_{t-1}-N_{det}-N_{del})},$$

where the number of false alarms and new tracks are still assumed to be Poisson distributed. Knowing the number of configurations and their prior probabilities the results from Eq. 3.17 and Eq. 3.18 can now be combined thus Eq. 3.15 yields

$$\begin{aligned} p(\psi_l(t) \mid \Omega_{p(l)}^{t-1},\ \mathcal{Z}^{t-1}) \ =\ & \eta\ p_{det}^{N_{det}}\ (1-p_{det})^{(N_{t-1}-N_{det})} \\ & p_{del}^{N_{del}}\ (1-p_{del})^{(N_{t-1}-N_{det}-N_{del})} \\ & \lambda_{fal}^{N_{fal}} V^{N_{fal}}\ \lambda_{new}^{N_{new}} V^{N_{new}}. \end{aligned} \tag{3.19}$$

It can be seen, that many terms cancel out making the final expression of the prior assignment probability simple. Substituting Eq. 3.6 and Eq. 3.19 into Eq. 3.5 the probability of hypothesis Ω_l^t given its parent $\Omega_{p(l)}^{t-1}$ and all measurements $\mathcal{Z}^t$ is

$$\boxed{\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t) \ =\ & \eta\ \lambda_{fal}^{N_{fal}}\ \lambda_{new}^{N_{new}} \prod_{i=1}^{M_t} \left(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i}\right)\ p_{det}^{N_{det}}\ (1-p_{det})^{(N-N_{det})} \\ & p_{del}^{N_{del}}\ (1-p_{del})^{(N-N_{det}-N_{del})}\ p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \end{aligned}} \tag{3.20}$$

3.4.1 Criticism

The approach of Cox and Hingorani [1996] models the events of track *detections* and track *deletions* using two independent binomial distributions. This leads to four possible combinations explaining the evolution of tracks. These combinations refer to tracks that are:

(1) **detected** and **not deleted**,

(2) **not detected** but **deleted**,

(3) **not detected** and **not deleted**, (implicitly modeling occlusion events),

(4) both: **detected** and **deleted**.

The first three combinations refer to meaningful explanations, namely (1) a matched track, (2) a deleted track, and (3) an occluded track. However, the fourth point has no meaningful equivalent since detection and deletion exclude each other. In other words, a track can not be marked as detected and deleted at the same time. To resolve that contradiction Arras et al. [2008] models the occurrences of track events using a multinomial distribution. Details are given in the subsequent section 3.5.

		previous tracks			false alarm labels			new track labels		
	$\mathcal{Z}(t) \setminus \Omega_{p(l)}^{t-1}$	$\mathbf{x}_1^{p(l)}(t-1)$	…	$\mathbf{x}_N^{p(l)}(t-1)$	fal_1	…	fal_M	new_1	…	new_M
current observations	$\mathbf{z}_1(t)$	$\mathcal{N}_{1,1}$	…	$\mathcal{N}_{1,N}$	λ_{fal}		0.0	λ_{new}		0.0
	⋮	⋮	⋱	⋮		⋱			⋱	
	$\mathbf{z}_M(t)$	$\mathcal{N}_{M,1}$	…	$\mathcal{N}_{M,N}$	0.0		λ_{fal}	0.0		λ_{new}
deletion labels	del_1	$\widehat{p_{del}}$		0.0						
	⋮		⋱							
	del_N	0.0		$\widehat{p_{del}}$						
occlusion labels	occ_1	$\widehat{p_{occ}}$		0.0						
	⋮		⋱							
	occ_N	0.0		$\widehat{p_{occ}}$						

Figure 3.2: Layout of the assignment matrix used to solve the data association problem in the MHT framework with explicit deletion labels and implicit occlusion labels. Rows denote current observations $\mathbf{z}_1(t), \ldots, \mathbf{z}_M(t)$, deletion labels $del_1, \ldots, del_N$, and implicit occlusion labels $occ_1, \ldots, occ_N$. Columns denote previously existing tracks $\mathbf{x}_1(t-1), \ldots, \mathbf{x}_N(t-1)$, false alarm labels $fal_1, \ldots, fal_M$, and new track labels $new_1, \ldots, new_M$. Matched tracks must be detected, thus $\mathcal{N}_{i,j} = p_{det}\mathcal{N}(\mathbf{z}_i(t))$. Not detected tracks can either be deleted with a probability of $\widehat{p_{del}} = (1 - p_{det})p_{del}$, or not deleted, hence $\widehat{p_{occ}} = (1 - p_{det})(1 - p_{del})$.

3.4.2 Notes on Assignment Generation

In the discussed expressions above the assignment set $\psi_l(t)$ used to generate a child hypothesis Ω_l^t from its parent $\Omega_{p(l)}^{t-1}$ was always assumed to be given and known in advance. The purpose of this section is to give a brief introduction into the generation of these assignment sets. A more detailed description of an efficient algorithm to generate the k best assignments that result in the k best hypotheses is given in section 3.8.

The assignment sets have the task to provide any possible[18] solution of the data association problem. Fortunately, the *best* solution is returned. Unfortunately, in many situation it is even hard to say how *best* is defined making the data association problem so difficult. However, in people tracking the data association problem (in this case also called assignment problem) is the task to assign each track to an observation or to mark it as deleted or occluded. Vice versa, each observation must be assigned to a track or marked as new track or false alarm.

This task can be reformulated using a weighted bipartite graph $\mathcal{G} = (U, V, E)$, with observations, deletions, and occlusion labels being the nodes U on the one hand. Tracks, false alarm, and new track labels are the nodes V on the other side. Last, the edges E connecting a node in U to a node in V are assigned with a weight or cost value to denote the likelihood of such an association event. The information encoded in the bipartite graph $\mathcal{G}$ can be expressed by a matrix $\mathcal{M} \in \mathbb{R}^{O \times T}$, where $O = M_t + 2 \cdot N_{t-1}$ denotes the number of rows and $T = N_{t-1} + 2 \cdot M_t$ the number of columns.

[18] At this point no requirements on the quality of the assignment sets are made. So far, only possible and meaningful associations from tracks to observations are required.

[19] The entries in the lower right block of the assignment matrix $\mathcal{M}$ assign virtual observation labels – new track and false alarm – to virtual track labels – occlusion and deletion – both not referring to real observations $\mathbf{z}_i(t)$ and

The entries in $\mathcal{M}$ correspond to the weights or costs of the edges E. The layout of the assignment matrix is shown in Figure 3.2. Each assignment set provides a complete matching in $\mathcal{G}$ connecting each node in U to exactly one node in V. An arbitrary complete matching can be found using the augmenting path algorithm. The minimum weight bipartite matching, defined as the matching that minimizes the sum of the weight values of the edges in the matching can be found using e.g. the Hungarian method.

Figure 3.2 shows that $\mathcal{M}$ is not of arbitrary shape but consists of several structured blocks. The dense top left block denotes all track to observation assignments and is completely populated in general. The entries refer to the matching likelihoods defined as $\mathcal{N}_{i,j} = p_{det}\mathcal{N}(\mathbf{z}_i(t);\, \hat{\mathbf{z}}_j(t),\, S_{i,j}(t))$. The two other blocks on the left hand side denote deleted and occluded tracks, respectively. These blocks contain only one entry different from zero in each column as each track can only be assigned to one deletion or occlusion label at maximum. Deleted and occluded tracks must not be detected, thus $\widehat{p_{del}} = (1 - p_{det})p_{del}$ and $\widehat{p_{occ}} = (1 - p_{det})(1 - p_{del})$. All other entries are set to zero. The same applies to the top middle and top right block that denote false alarms and new tracks occurring with an expected number of events λ_{new} and λ_{fal}, respectively. Here, each row contains only one entry different from zero as one observation can only be assigned to one false alarm or one new track label. The lower right block corresponds to the edges of the track labels to occlusion labels needed to complete the graph. As none of these assignments must have influence on the tracking result their probability is set to 1.0. Hence, the lower right block is completely populated with ones.

Finally, the probability of an assignment set is simply the product of the weight values[20] of the edges contained in the matching. Obviously, many possible matchings contain edges with zero weights thus their probability drops to zero as well. Such assignment sets are called invalid and are discarded.

3.5 Explicit Occlusion Labels

A further extension of the MHT was proposed by Arras et al. [2008]. Adding explicitly modeled track occlusions the assignment set $\psi_l(t)$ assigns each observation to one of the observation labels *detected*, *new track*, or *false alarm*, and each track to one of the track labels *detected*, *deleted*, or *occluded*. To formally express the tracks that are marked as occluded in the current frame an additional indicator variable needs to be defined. It extends the set of indicator variables introduced in 3.3 and 3.14 and is

$$\omega_j \;=\; \omega_j[\psi_l(t)] \;=\; \begin{cases} 1 & \text{if track } \mathbf{x}_j(t-1) \text{ from } \Omega^{t-1}_{p(l)} \text{ is occluded in the current frame,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.21}$$

Using ω the number of occluded tracks can be obtained as $N_{occ} = \sum_{j=1}^{N_{t-1}} \omega_j$. With the assumption that tracks must either be detected, occluded or declared as deleted it follows that

$$N_{t-1} = N_{det} + N_{occ} + N_{del}. \tag{3.22}$$

tracks $\mathbf{x}_j(t-1)$, respectively. Their likelihoods must not have influence on the final assignment probability and is therefore set to 1.0.

[20] The maximum weighted bipartite matching maximizes the **sum** of the costs but a maximization of the **product** of the corresponding probabilities is required. Therefore, the probabilities are converted into negative log likelihoods serving as costs in the assignment matrix.

With the information encoded in all indicator variables (see 3.3, 3.14, and 3.21) the joint probability of the prior assignment probability as first defined in Eq. 3.7 can be refined to

$$P(\psi_l(t) \mid \Omega_{p(l)}^{t-1},\ \mathcal{Z}^{t-1}) = P\left(\psi_l(t) \mid \tau[\psi_l(t)],\ \phi[\psi_l(t)],\ \nu[\psi_l(t)],\ \chi[\psi_l(t)],\ \omega[\psi_l(t)]\right) \tag{3.23}$$
$$P\left(\tau[\psi_l(t)],\ \phi[\psi_l(t)],\ \nu[\psi_l(t)],\ \chi[\psi_l(t)],\ \omega[\psi_l(t)] \mid \Omega_{p(l)}^{t-1}\right).$$

The number of possible configurations depends on the permutations to assign N_{det} tracks to the N_{t-1} previously existing tracks and the combinations to declare the remaining tracks as occluded or deleted. Given the assumption in Eq. 3.22 latter is given by the multinomial coefficient thus Eq. 3.16 changes into

$$\frac{N_{t-1}!}{(N_{t-1}-N_{det})!} \begin{pmatrix} N_{t-1}-N_{det} \\ N_{occ}\ N_{del} \end{pmatrix} = \frac{N_{t-1}!}{N_{occ}!\ N_{del}!}. \tag{3.24}$$

Combining Eq. 3.8 and Eq. 3.24 the prior probability of an assignment set depending on the combinatorics yields

$$P\left(\psi_l(t) \mid \tau[\psi_l(t)],\ \phi[\psi_l(t)],\ \nu[\psi_l(t)],\ \chi[\psi_l(t)],\ \omega[\psi_l(t)]\right) = \tag{3.25}$$
$$\frac{N_{det}!\ N_{fal}!\ N_{new}!\ N_{occ}!\ N_{del}!}{M_t!\ N_{t-1}!}.$$

As already mentioned, the two independent binomial distributions used to model detection and deletion events are replaced by a common multinomial distribution that models all three events (detections, occlusions, and deletions) jointly, thus Eq. 3.11 changes into

$$P\left(\tau[\psi_l(t)],\ \phi[\psi_l(t)],\ \nu[\psi_l(t)],\ \omega[\psi_l(t)] \mid \Omega_{p(l)}^{t-1}\right) \tag{3.26}$$
$$= \begin{pmatrix} N_{t-1} \\ N_{det}\ N_{occ}\ N_{del} \end{pmatrix} p_{det}^{N_{det}}\ p_{occ}^{N_{occ}}\ p_{del}^{N_{del}}\ P_{\lambda_{fal}V}(N_{fal})\ P_{\lambda_{new}V}(N_{new})$$
$$= \frac{N_{t-1}!}{N_{det}!\ N_{occ}!\ N_{del}!}\ p_{det}^{N_{det}}\ p_{occ}^{N_{occ}}\ p_{del}^{N_{del}}\ \frac{\lambda_{fal}^{N_{fal}} V^{N_{fal}}}{N_{fal}!}\ e^{-\lambda_{fal}V}\ \frac{\lambda_{new}^{N_{new}} V^{N_{new}}}{N_{new}!}\ e^{-\lambda_{new}V},$$

where the number of false alarms N_{fal} and new tracks N_{new} are further modeled with the two Poisson distributions $P_{\lambda_{fal}V}(N_{fal})$ and $P_{\lambda_{new}V}(N_{new})$, respectively. Still assuming that all possible assignments are equally likely, the substitution of Eq. 3.24 and Eq. 3.26 into Eq. 3.23 yields

$$p(\psi_l(t) \mid \Omega_{p(l)}^{t-1},\ \mathcal{Z}^{t-1}) = \eta\ p_{det}^{N_{det}}\ p_{occ}^{N_{occ}}\ p_{del}^{N_{del}}\ \lambda_{fal}^{N_{fal}}\ V^{N_{fal}}\ \lambda_{new}^{N_{new}}\ V^{N_{new}}, \tag{3.27}$$

where the normalizer η is the product of the assignment independent terms $\frac{1}{M_t}$, $e^{-\lambda_{fal}V}$, and $e^{-\lambda_{new}V}$. The final expression of the hypothesis probability – as a result of substituting Eq. 3.6 and Eq. 3.27 into Eq. 3.5 – simplifies to

$$\boxed{\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t) = \eta\ \lambda_{fal}^{N_{fal}}\ \lambda_{new}^{N_{new}} \prod_{i=1}^{M_t} \left(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i}\right) \\ p_{det}^{N_{det}}\ p_{occ}^{N_{occ}}\ p_{del}^{N_{del}}\ p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \end{aligned}} \tag{3.28}$$

		previous tracks			false alarm labels			new track labels		
	$\mathcal{Z}(t) \setminus \Omega_{p(l)}^{t-1}$	$\mathbf{x}_1^{p(l)}(t-1)$	$\dots$	$\mathbf{x}_N^{p(l)}(t-1)$	fal_1	$\dots$	fal_M	new_1	$\dots$	new_M
current observations	$\mathbf{z}_1(t)$	$\mathcal{N}_{1,1}$	$\dots$	$\mathcal{N}_{1,N}$	λ_{fal}		0.0	λ_{new}		0.0
	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\ddots$			$\ddots$	
	$\mathbf{z}_M(t)$	$\mathcal{N}_{M,1}$	$\dots$	$\mathcal{N}_{M,N}$	0.0		λ_{fal}	0.0		λ_{new}
deletion labels	del_1	p_{del}		0.0						
	$\vdots$		$\ddots$							
	del_N	0.0		p_{del}						
occlusion labels	occ_1	p_{occ}		0.0						
	$\vdots$		$\ddots$							
	occ_N	0.0		p_{occ}						

1.0

Figure 3.3: Layout of the assignment matrix used to solve the data association problem in the MHT framework with explicite deletion labels and explicite occlusion labels. See Figure 3.2 for comparison. Matched tracks are detected, thus $\mathcal{N}_{i,j} = p_{det}\mathcal{N}(\mathbf{z}_i(t))$. Detections, occlusions, and deletions are modeled jointly with a multinomial distribution, hence $p_{det} + p_{occ} + p_{del} = 1$.

Adding explicit occlusion labels the assignment matrix introduced in Figure 3.2 needs to be extended. The new layout containing an additional block of tracks assigned to occlusion labels is shown in Figure 3.3. Matching likelihoods are defined as $\mathcal{N}_{i,j} = p_{det}\mathcal{N}(\mathbf{z}_i(t);\, \hat{\mathbf{z}}(j)(t),\, S_(i,j))$.

Preparatory to the next sections Eq. 3.28 will be provided in a different syntax using the indicator variable more extensively. With $N_{det} + N_{fal} + N_{new} = M_t$ and $N_{det} + N_{occ} + N_{del} = N_{t-1}$ it can be rewritten as

$$p(\Omega_l^t \mid \mathcal{Z}^t) = \eta \prod_{i=1}^{M_t} \left(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i} \;\; \lambda_{fal}^{\phi_i} \;\; \lambda_{new}^{\nu_i}\right) \prod_{j=1}^{N_{t-1}} \left(p_{det}^{\delta_j} \;\; p_{occ}^{\omega_j} \;\; p_{del}^{\chi_j}\right) \;\; p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \tag{3.29}$$

The conversion of Eq. 3.28 into Eq. 3.29 becomes obvious when using the indicator variables to count the number of events. It can be shown that $\lambda_{fal}^{N_{fal}} = \prod_{i=1}^{M_t} \lambda_{fal}^{\phi_i}$ with $N_{fal} = \sum_{i=1}^{M_t} \phi_i$ and $p_{occ}^{N_{occ}} = \prod_{j=1}^{N_{t-1}} p_{occ}^{\omega_j}$ with $N_{occ} = \sum_{j=1}^{N_{t-1}} \omega_j$. The same applies for N_{new}, N_{det}, and N_{del}. A benefit of Eq. 3.29 is that it allows to easily introduce spatio-temporal rates of false alarm and new track events (see section 3.6) as well as spatio-temporal probabilities of occlusion and deletion events (see section 3.7).

3.6 Space-Time-Dependent Prior Probabilities

This section introduces explicitly modeled spatio-temporal priors on new track and false alarm events that further extend the MHT framework. Up to this point, occurrences of new track and false alarm events have been modeled using Poisson distributions with fixed rate functions λ_{new} and λ_{fal},

respectively, modeling the expected numbers of events, assumed to be constant and independent of position and time of the observations. It can easily be illustrated, that both assumptions are wrong in practice. People do not appear uniformly in space and time. They typically enter the environment at specific locations like doors, elevators, or corners and appear more often at daytime than at night. The same applies for false alarms, e.g. they appear more often in clutter than in open space. Such simple examples demonstrate that detailed information about the distributions of new track and false alarm events – learned or modeled – should be incorporated into the MHT to refine the hypotheses probabilities.

To model events that occur randomly in time, the non-homogeneous Poisson process is a natural choice. It is a stochastic process that counts the number of events given an expected average number of events in a time interval. The probability distribution of the number of occurrences $N(t)$ follows a Poisson distribution. The Poisson distribution is parametrized with a positive real number λ that encodes the rate at which events occur per time unit. To implement the concept of temporal-dependent events the average expected number of events λ is modeled with a function conditioned on time as the rate parameter may change. The generalized rate function is given as $\lambda(t)$ and the expected number of events occurring within the time interval $(t_s,\, t_e]$ is

$$\lambda_{t_s,t_e} \;=\; \int_{t_s}^{t_e} \lambda(t)\, dt. \tag{3.30}$$

Let $N(t) = (N(t_e) - N(t_s))$ be a discrete random variable to represent the number of events occurring in the time interval $(t_s,\, t_e]$ with rate function λ_{t_s,t_e} then $N(t)$ follows a Poisson distribution with parameter λ_{t_s,t_e}, hence

$$P((N(t_e) - N(t_s)) = n) \;=\; \frac{(\lambda_{t_s,t_e})^n}{n!}\, e^{-\lambda_{t_s,t_e}} \qquad n = 0, 1, \ldots \tag{3.31}$$

A homogeneous Poisson process used previously is a special case of a non-homogeneous process with constant rate function $\lambda(t) = \lambda$.

A Poisson process with spatio-temporal dependent rate function is called the *space-time* Poisson process. It extends the non-homogeneous Poisson process and introduces a spatial dependency on the rate function given as $\lambda(\vec{x}, t)$ with $\vec{x} \in X$ where X is a vector space such as $\mathbb{R}^2$ or $\mathbb{R}^3$. For any subset $S \subset X$ of finite extent (e.g. a spatial region), the number of events occurring inside this region can be modeled as a Poisson process with associated rate function $\lambda_S(t)$ such that

$$\lambda_S(t) = \int\limits_{\vec{x} \in S} \lambda(\vec{x}, t)\, d\vec{x}. \tag{3.32}$$

Is the generalized rate function $\lambda(\vec{x}, t)$ a separable function of time and space, $\lambda(\vec{x}, t) = f(\vec{x})\lambda(t)$ for some function $f(\vec{x})$[21], this decomposition allows to decouple the occurrence of events between time and space. More details on these functions and how they can be modeled or learned is explained in Chapter 5 and Chapter 6, respectively

For the purpose of modeling spatio-temporal dependent occurrences of new track and false alarm events the formally used homogeneous Poisson processes with fixed rate parameters λ_{new} and λ_{fal} are now conditioned on both: the position of the currently investigated observation $\mathbf{z}_i(t)$ and the current time t, thus the spatio-temporal dependent expected number of new track and false alarm events yields $\lambda_{new}(\mathbf{z}_i(t), t)$ and $\lambda_{fal}(\mathbf{z}_i(t), t)$, respectively. Replacing the fixed rates λ_{new} and λ_{fal} in Eq. 3.29 with spatio-temporal rate functions $\lambda_{new}(\mathbf{z}_i(t), t)$ and $\lambda_{fal}(\mathbf{z}_i(t), t)$, respectively, the expression of the

[21] For which can be demanded $\int_X f(\vec{x})\, d\vec{x} = 1$ without loss of generality.

hypothesis probability Eq. 3.5 yields

$$p(\Omega_l^t \mid \mathcal{Z}^t) = \eta \prod_{i=1}^{M_t} \left(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i} \ \lambda_{fal}(\mathbf{z}_i(t), t)^{\phi_i} \ \lambda_{new}(\mathbf{z}_i(t), t)^{\nu_i} \right) \prod_{j=1}^{N_{t-1}} \left(p_{det}^{\delta_j} \ p_{occ}^{\omega_j} \ p_{del}^{\chi_j} \right) p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \tag{3.33}$$

The same idea of spatio-temporal dependent event likelihoods can be applied to improve the models of detections, occlusions, and deletions, respectively. How this is realized is shown in the next section.

3.7 Space-Time-Dependent Target Probabilities

While the last section was concerned with improved models on observation events, in this section refined models for track-specific events are introduced. Comparable to spatio-temporal priors that influence the creation of new tracks spatial and temporal information aids to deal with track continuation, occlusion, and deletion events, respectively. In the previous considerations those track-specific events were assumed to occur uniformly in space and time with constant probabilities. Indeed, the work of Arras et al. [2008] proposes adaptive occlusion probabilities for leg tracks that surely belong to a person. But they distinguish only between these two states (approved or not approved leg tracks) and do not allow the integration of additional information.

However, the assumptions that track events occur uniformly in space and time do not hold in general. Illustrated by examples, occlusion events are more likely to occur close to other people and obstacles than in open space. Especially static obstacles cause frequent occlusions in their shadow. Moreover, in the dark when a camera based tracking system fails to detect people reliably detection events are less likely while occlusions occur more often. Further, the probability of deletion events increases at the borders of the sensor field of view.

To capture spatial and temporal dependencies the formerly used constant probabilities of detection (p_{det}), occlusion (p_{occ}), and deletion (p_{del}) events, respectively, are now modeled as functions conditioned on the position of the currently investigated track $\mathbf{x}_j(t-1)$ and the current time frame t, thus

$$p_{event}\left(\mathbf{x}_j(t-1),\, t\right) \ := \ f_{event}\left(\mathbf{x}_j(t-1),\, t\right). \tag{3.34}$$

Note, that the occurrences of event are modeled jointly using a multinomial distribution, hence the event probabilities must sum up to one for each track ($p_{det}^j + p_{occ}^j + p_{del}^j = 1$). How these functions can be modeled is shown in Chapter 5.

Substituting the spatio-temporal dependent probabilities $p_{det}(\mathbf{x}_j(t-1), t)$, $p_{occ}(\mathbf{x}_j(t-1), t)$, and $p_{del}(\mathbf{x}_j(t-1), t)$ into Eq. 3.33 yields

$$p(\Omega_l^t \mid \mathcal{Z}^t) = \eta \prod_{j=1}^{N_{t-1}} \left(p_{det}(\mathbf{x}_j(t-1), t)^{\delta_j} \ p_{occ}(\mathbf{x}_j(t-1), t)^{\omega_j} \ p_{del}(\mathbf{x}_j(t-1), t)^{\chi_j} \right) \prod_{i=1}^{M_t} \left(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i} \ \lambda_{fal}(\mathbf{z}_i(t), t)^{\phi_i} \ \lambda_{new}(\mathbf{z}_i(t), t)^{\nu_i} \right) p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \tag{3.35}$$

		previous tracks			false alarm labels			new track labels		
	$\mathcal{Z} \setminus \Omega$	$\mathbf{x}_1$	$\dots$	$\mathbf{x}_N$	fal_1	$\dots$	fal_M	new_1	$\dots$	new_M
current observations	$\mathbf{z}_1$	$\mathcal{N}_{1,1}$	$\dots$	$\mathcal{N}_{1,N}$	$\lambda_{fal}(\mathbf{z}_1, t)$		0.0	$\lambda_{new}(\mathbf{z}_1, t)$		0.0
	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\ddots$			$\ddots$	
	$\mathbf{z}_M$	$\mathcal{N}_{M,1}$	$\dots$	$\mathcal{N}_{M,N}$	0.0		$\lambda_{fal}(\mathbf{z}_M, t)$	0.0		$\lambda_{new}(\mathbf{z}_M, t)$
deletion labels	del_1	$p_{del}(\widehat{\mathbf{x}}_1, t)$		0.0						
	$\vdots$		$\ddots$							
	del_N	0.0		$p_{del}(\widehat{\mathbf{x}}_N, t)$						
occlusion labels	occ_1	$p_{occ}(\widehat{\mathbf{x}}_1, t)$		0.0			1.0			
	$\vdots$		$\ddots$							
	occ_N	0.0		$p_{occ}(\widehat{\mathbf{x}}_N, t)$						

Figure 3.4: Layout of the assignment matrix used to solve the data association problem in the MHT framework with spatio-temporal-dependent deletion, occlusion, false alarm, and new track labels. See Figure 3.2 and Figure 3.3 for comparison. The likelihood of the matched tracks is modeled as $\mathcal{N}_{i,j} = p_{det}\mathcal{N}(\mathbf{z}_i(t))$. Occlusions, deletions, new track, and false alarm events are conditioned on observation positions $\mathbf{z}_i$ or predicted track positions $\widehat{\mathbf{x}}_j$, respectively, and on the current time frame t.

In case the motion of people can be estimated in advance (using e.g. Kalman or particle filters) the predicted position $\widehat{\mathbf{x}}_j(t)$ can be used to calculate the place-dependent probabilities of the track events. The Kalman-filter prediction as used in this work is

$$\widehat{\mathbf{x}}_j(t) \;=\; F_t\ \mathbf{x}_j(t-1) \;+\; \omega_t, \tag{3.36}$$

with state transition model F_t applied to the previous state $\mathbf{x}_j(t-1)$ and process noise $\omega_t \sim \mathcal{N}(0, Q_t)$ which is assumed to be drawn from a zero mean Gaussian distribution with covariance matrix Q_t. Employing the predicted positions $\widehat{\mathbf{x}}_j(t)$ at time t the final expression of the hypothesis probability Eq. 3.5 yields

$$\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t) \;= \eta \prod_{j=1}^{N_{t-1}} & \left(p_{det}(\widehat{\mathbf{x}}_j(t), t)^{\delta_j}\ p_{occ}(\widehat{\mathbf{x}}_j(t), t)^{\omega_j}\ p_{del}(\widehat{\mathbf{x}}_j(t), t)^{\chi_j} \right) \\ & \prod_{i=1}^{M_t} \left(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i}\ \lambda_{fal}(\mathbf{z}_i(t), t)^{\phi_i}\ \lambda_{new}(\mathbf{z}_i(t), t)^{\nu_i} \right) p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \end{aligned} \tag{3.37}$$

As proposed in Arras et al. [2008] further information can be utilized to adapt the probabilities of assignment events and to refine the MHT formulation. Chapter 7 introduces learned models from which social relations between walking people can be inferred. These models aid to adapt the occlusion probabilities of people walking in groups and frequently occlude each other.

3.8 Efficient Implementation and Pruning Strategies

This section addresses the efficient implementation of the MHT framework. Especially, the fast and memory efficient generation of the best assignments/hypotheses is explained in detail. As a reminder, the number of hypotheses generated by the MHT grows exponentially over time. Thus, only the k most accurate or important hypotheses are maintained and the worse once are discarded. The formulation of the MHT introduced above provides a concept to measure the quality of each hypothesis based on sensory information and system parameters. But how to find the k best solutions? Brute-force search – that generates all possible permutations, sorts them by their probabilities, and finally selects the k best ones – has to generate $n!$ hypotheses[22] where

$$n \;=\; \max(T, O), \qquad T \;=\; N_{t-1} + 2 \cdot M_t, \qquad O \;=\; M_t + 2 \cdot N_{t-1}, \tag{3.38}$$

where T denotes the total number of track labels $\mathcal{T}$ and O denotes the total number of observation labels $\mathcal{O}$, respectively. This enumeration is unfeasible in practice, thus the k best hypotheses need to be generated directly in a more efficient manner.

As outlined in subsection 3.4.2, the assignment problem can be represented as weighted bipartite graph $\mathcal{G} = (U, V, E)$ in which each node in U represents a track label and each node in V an observation label, respectively. The weighted edges $\varepsilon_{i,j} = (\mathbf{z}_i, \mathbf{x}_j, m_{i,j}, c_{i,j})$ denote that observation z_i can be assigned to track $\mathbf{x}_j$ with assignment probability $m_{i,j}$ and costs $c_{i,j}$. These costs are derived using negative log likelihoods of the assignment probabilities encoded in the assignment matrix $\mathcal{M}$, thus

$$c_{i,j} \propto -\log(m_{i,j}). \tag{3.39}$$

With the use of $\mathcal{G}$ and $\mathcal{M}$, finding the best assignment $\psi = \{\varepsilon_{i,j}\}$ can be reformulated as a classical linear assignment problem (LAP) that minimizes the objective function

$$\sum_{i \in \mathcal{O}} \sum_{j \in \mathcal{T}} c_{i,j}\, \psi_{i,j} \tag{3.40}$$

constraint to

$$\sum_{i \in \mathcal{O}} \psi_{i,j} \;=\; \mathbf{1}, \quad \forall\, j \in \mathcal{T}, \qquad \text{and} \qquad \sum_{j \in \mathcal{T}} \psi_{i,j} \;=\; \mathbf{1}, \quad \forall\, i \in \mathcal{O}, \tag{3.41}$$

where $\psi_{i,j} \in \{0, 1\}$ indicates which edges $\varepsilon_{i,j}$ (association of observation $\mathbf{z}_i$ to track $\mathbf{x}_j$) are included in the assignment set ψ.

The assignment algorithm – also known as Hungarian method – by Munkres [1957] (found to be previously solved by Jacobi [1865]) can be used to find the best solution of the assignment problem in polynomial time with an upper bound of $O(n^3)$. The probability of an assignment $\psi_l(t)$ and thus also of the resulting hypothesis Ω_l^t is

$$\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t) \;&=\; p(\psi_l(t),\; \Omega_{p(l)}^{t-1} \mid \mathcal{Z}^t) \\ &=\; \prod_{\varepsilon_{i,j}} (m_{i,j}) \quad p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \end{aligned} \tag{3.42}$$

[22] Note, that not all of the $n!$ possible permutations yield valid assignment sets with a positive probability. However, the upper bound can not be reduced without further considerations.

Algorithm 2: Murty's single parent k-best hypotheses generation algorithm.

Input : Set of current observations $\mathcal{Z}(t)$.
Parent hypothesis Ω^{t-1} from the previous time step $t-1$.
Number of requested child hypotheses K.

Output : Set $\mathcal{H} = \{\Omega_1^t \ \dots \ \Omega_k^t\}$ of the 1^{st} to the k^{th} best child hypotheses generated from $\Omega_{p(l)}^{t-1}$, with $k \leq K$.

Variables: Set $\mathcal{A}$ of triples $(\mathcal{G}_i,\ \psi_i,\ p(\Omega_i|\mathcal{Z}^t))$ containing a problem description $\mathcal{G}_i$, its solution ψ_i, and the probability $p_i = p(\Omega_i|\mathcal{Z}^t)$ of the resulting hypothesis Ω_i.

$\mathcal{A} \leftarrow \emptyset$;
`/* initialize the bipartite graph` $\mathcal{G}_1$ `*/`
$\mathcal{G}_1 \leftarrow$ *initialize*$(\Omega_{p(l)}^{t-1}, \mathcal{Z}(t))$;
`/* find the best solution` ψ_1 `of` $\mathcal{G}_1$`, e.g. using Hungarian method */`
$(\psi_1, p_1) \leftarrow$ *solve*$(\mathcal{G}_1)$;
$\mathcal{A} \leftarrow \mathcal{A} \cup (\mathcal{G}_1, \psi_1, p_1)$;
`/* main loop */`
for $k \leftarrow 1$ **to** $K \ \wedge \ \mathcal{A} \neq \emptyset$ **do**
  `// get the triple with the highest hypothesis probability`
  $(\mathcal{G}_k,\ \psi_k,\ p_k) = \max_{p_i} (\mathcal{G}_i,\ \psi_i,\ p_i)$;
  `// remove it from` $\mathcal{A}$
  $\mathcal{A} \leftarrow \mathcal{A} \setminus (\mathcal{G}_k, \psi_k, p_k)$;
  `// create the hypothesis and add it to` $\mathcal{H}$
  $\Omega_k^t = \{\psi_k, \Omega^{t-1}\}$;
  $\mathcal{H} \leftarrow \mathcal{H} \cup \Omega_k^t$;
  `// create sub-problems`
  for *each* $\varepsilon_{i,j} \in \psi_k$ **do**
    $\mathcal{G}' = \mathcal{G}_k \setminus \varepsilon_{i,j}$;
    $(\psi', p') \leftarrow$ *solve*$(\mathcal{G}')$;
    `// if` ψ' `is a valid assignment, add the triple` $(\mathcal{G}', \psi', p')$ `to` $\mathcal{A}$
    if ψ' *exists and is valid* **then**
      $\mathcal{A} \leftarrow \mathcal{A} \cup (\mathcal{G}', \psi', p')$;
    end
    `// reduce the dimensionality of the problem` $\mathcal{G}_k$
    for *each* $\varepsilon_{u,v} \in \mathcal{G}_k$ **do**
      if $\mathbf{z}_i = \mathbf{z}_u \oplus \mathbf{x}_j = \mathbf{x}_v$ **then**
        $\mathcal{G}_k \leftarrow \mathcal{G}_k \setminus \varepsilon_{u,v}$;
      end
    end
  end
end
`/* generation of the best` k `hypothesis done, with` $(k = K \vee \mathcal{A} = \emptyset)$ `*/`
return $\mathcal{H}, k$;

Figure 3.5: Murty's single parent k-best hypotheses generation algorithm.

Danchick and Newnam [1993] illustrate how the cost matrix $\mathcal{M}$ (shown in Figure 3.2, 3.3, and 3.4) must be modified and solved repeatedly to compute the k best assignments. Unfortunately, their algorithm must identify and eliminate duplicate assignments and requires, in the worst case, the solution of $k!$ linear assignment problems. Hence its time complexity is $O(k!\,n^3)$.

To generate the k best hypotheses more efficiently, Murty [1968] proposes an algorithm for that the number of linear assignment problems needed to be solved is linear in k. The algorithm (depicted in algorithm 2) operates on a sequence of transformed cost matrices $\mathcal{M}$ as well and solves the assignment problems consecutively. In advantage to Danchick and Newnam it avoids solving duplicates, thereby eliminating the need to compare and delete duplicate hypotheses. However, each of the k best solutions is partitioned, in the worst case, into $O(n)$ new subproblems. This creates up to $O(k\,n)$ assignment problems in total inserted into an priority queue, hence its complexity is $O(k\,n^4)$. See also Miller et al. [1997] for more details.

Pedersen et al. [2008] showed that consecutive assignment subproblems generated within Murty's algorithm can be re-solved based on previous solutions. Employing Jonker and Volgenant [1987] LAP algorithm and using Dijkstra's method to find the shortest path, the solution of a subproblem can be found in $O(n^2)$. Thus, the time complexity of their approach is $O(k\,n^3)$, which is equal to the best-known time complexity for ranking the k best assignments.

3.8.1 Murty's Algorithm to Find the k best Assignments

In this section a brief description of Murty's assignment algorithm to find the best k hypotheses in polynomial time is given. Its pseudo code is presented in algorithm 2.

The method requires a parent hypothesis Ω^{t-1} from the previous time step and a set of current observations $\mathcal{Z}(t)$ as input[23]. Based on these information, line 2 starts with initializing the unconstrained bipartite assignment graph $\mathcal{G}_1$ and constructs the assignment cost matrix $\mathcal{M}$ which encodes the costs of all possible and impossible assignments between track and observation labels. The costs of possible assignments are derived from the assignment probabilities using negative log likelihoods while impossible assignments receive infinite costs. Thereafter, in line 3 the unconstrained solution (ψ_1, p_1) to the LAP in which all matrix entries $c_{i,j}$ may participate is calculated – using e.g. the Hungarian method – and added to the set $\mathcal{A}$ of triples ($\mathcal{G}_k$, ψ_k, $p_k := p(\Omega_k|\mathcal{Z}^t)$) containing a problem description, its solution, and the probability of the resulting hypotheses. Latter can be found by summing the negative log likelihoods of all edges $\varepsilon_{i,j}$ specified by ψ_k and the negative log likelihood of the parent hypothesis Ω^{t-1}, hence

$$-\log(p(\Omega_k^t \mid \mathcal{Z}^t)) \;=\; \sum_{\varepsilon_{i,j}\in\psi_k} (-\log(m_{i,j})) \;-\; \log(p(\Omega^{t-1} \mid \mathcal{Z}^{t-1})). \tag{3.43}$$

In the main loop of the algorithm (line 5 to line 22) further subproblems are generated by partitioning the currently best available assignment problem $\mathcal{G}_k$ found by e.g. linear search in line 6. Beforehand, the k^{th} best hypothesis is created in line 8 and added to the set of hypotheses $\mathcal{H}$ returned at the end. Murty's approach to generate the sub-problems has two advantages, namely:

(1) The set of valid solutions for any one of the subproblems does not intersect with the set of solutions for any other problem. In other words, there are no duplicate problems.

(2) The union of the sets of valid solutions for all the subproblems is exactly the set of solutions for problem $\mathcal{G}_k$, minus its solution ψ_k.

[23] The input set of observations and tracks (contained in the parent hypothesis) are allowed to be empty. Is, for example, the set of observations empty all tracks can still be assigned to occlusion and deletion labels.

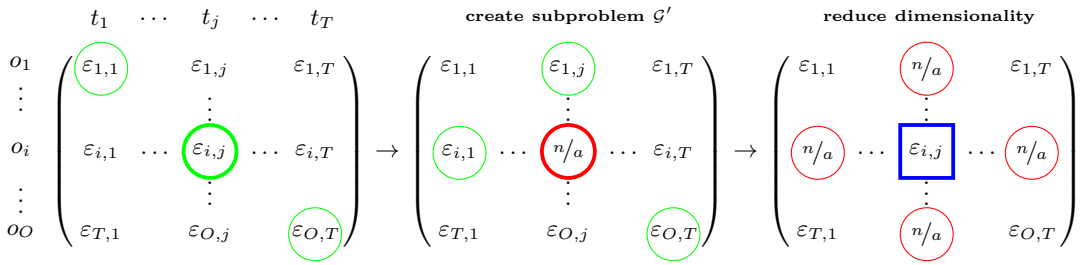


Figure 3.6: Schematic illustration of Murty's k best hypotheses algorithm. *Left:* Given an assignment problem $\mathcal{G}$, shown by its assignment matrix $\mathcal{M}$, and its best solution ψ, indicated by green circles, an edge $\varepsilon_{i,j} \in \psi$ is selected. *Middle:* The edge $\varepsilon_{i,j}$ is removed from the bipartite graph (red circle), thus $\mathcal{G}' = \mathcal{G} \setminus \varepsilon_{i,j}$ for that ψ' is the best solution, consequently ψ' is the 2^{nd} best solution of $\mathcal{G}$. *Right:* All edges connected to o_i or t_j (marked in red) are removed from further subproblems $\mathcal{G}'' = \mathcal{G}\,(U,\, V,\, E \setminus \{\varepsilon_{u,v}\})$, with $(i = u \oplus j = v)$, while $\varepsilon_{i,j} \in \psi''$ (blue box) is fixed in their solutions. This decreases the dimensionality by 1.

The procedure to iteratively derive the subproblems from a given problem description $\mathcal{G}_k$ is illustrated in Figure 3.6. Shown in line 10, each edge $\varepsilon_{i,j}$ contained in the solution ψ_k is removed from $\mathcal{G}_k$, successively, creating new subproblems $\mathcal{G}' = \mathcal{G}_k \setminus \varepsilon_{i,j}$. If the solution ψ' of $\mathcal{G}'$ is valid, indicated by $p' > 0$, a new triple $(\mathcal{G}', \psi', p')$ is added to $\mathcal{A}$ to be later selected according to its probability. However, all edges connected to o_i or t_j are removed from the graph $\mathcal{G}_k$ except the edge $\varepsilon_{i,j}$ itself (see line 16). Latter is kept fixed for all following solutions. Each entry in row i and column j has to be removed as visualized in Figure 3.6(*right*). This reduces the dimensionality of the assignment problem for the next iteration by 1. This is repeated until all edges in ψ_k have been removed from $\mathcal{G}_k$.

3.8.2 Multi Parent Variant of Murty's Algorithm

Murty's algorithm presented above provides an efficient method to find the k best child hypotheses of a single parent $\Omega^{t-1}_{p(l)}$. But the MHT holds multiple parents $\{\Omega^{t-1}_0, \Omega^{t-1}_1, \dots\}$ that emerged from their previous time step $t-2$, thus an adaptation to the assignment generation algorithm is required. The naïve approach is to apply algorithm 2 on all parents, independently, thereby generating multiple sets $\mathcal{H}_i$ each containing the k best child hypotheses of the parent Ω^{t-1}_i. Thereafter, all children need to be sorted with respect to their probability to find and maintain only the global k best ones. The worse once would be rejected. This procedure is very inefficient as it has to solve many assignment problems, unnecessarily. Its time complexity is $O(k^2\, n^3 + k^4)$, where the former term denotes the generation of $k \cdot k$ child hypotheses and the latter one the sorting of the $l = k^2$ child hypotheses that needs $O(l^2)$ time.

In algorithm 3 an extended approach to Murty's algorithm is presented. Accepting multiple parent hypotheses as input it generates the global k best children, jointly. It takes advantage of a shared pool $\mathcal{A}$ of cached solutions used to define the subproblems (line 7 in alg. 3) and inserts all initial, unconstrained solutions into $\mathcal{A}$. The most expensive part of the algorithm is the initialization procedure that has to find the best solution to all given parent hypotheses. No matter what method is used, this takes $O(k\, n^3)$ time. Fortunately, the solutions of the subproblems can be calculated more efficiently using the re-optimization method of Pedersen et al. [2008], which yields $O(k\, n^3)$ as well. Thus the overall time complexity is $O(k\, n^3)$.

Algorithm 3: The multi parent k-best hypotheses generation algorithm.

Input : Set of current observations $\mathcal{Z}(t)$.
Set $\mathcal{H}_{t-1} = \left\{\Omega_1^{t-1}, \ldots, \Omega_{n_{t-1}}^{t-1}\right\}$ of parent hypotheses with $n_{t-1} \leq K$.
Number of requested child hypotheses K.

Output : Set $\mathcal{H}_t = \{\Omega_1^t, \ldots, \Omega_k^t\}$ of the k best child hypotheses generated from $\mathcal{H}_{t-1}$ with $k \leq K$.

Variables: Set $\mathcal{A}$ of triples $(\mathcal{G}_i,\ \psi_i,\ p(\Omega_i|\mathcal{Z}^t))$ containing a problem description $\mathcal{G}_i$, its solution ψ_i, and the probability $p_i = p(\Omega_i|\mathcal{Z}^t)$ of the resulting hypothesis Ω_i.

1 $\mathcal{A} \leftarrow \emptyset$;
`/* initialize all problem descriptions` $\mathcal{G}_i$ `and find their best solutions` ψ_i `*/`
2 **for** $i \leftarrow 1$ **to** n_{t-1} **do**
3 $\quad \mathcal{G}_i \leftarrow initialize(\Omega_i^{t-1}, \mathcal{Z}(t))$;
4 $\quad (\psi_i, p_i) \leftarrow solve(\mathcal{G}_i)$;
5 $\quad \mathcal{A} \leftarrow \mathcal{A} \cup (\mathcal{G}_i, \psi_i, p_i)$;
6 **end**
`/* main loop, similar to algorithm 2 (line 5 - line 22) */`
7 **for** $k \leftarrow 1$ **to** $K \ \wedge \ \mathcal{A} \neq \emptyset$ **do**
8 $\quad (\mathcal{G}_k,\ \psi_k,\ p_k) = \max_{p_i} (\mathcal{G}_i,\ \psi_i,\ p_i)$;
9 $\quad \mathcal{A} \leftarrow \mathcal{A} \setminus (\mathcal{G}_k, \psi_k, p_k)$;
$\quad$ `// create the child hypothesis using its parent` $\Omega_{p(k)}^{t-1}$ `and add it to` $\mathcal{H}$
10 $\quad \Omega_k^t = \{\psi_k, \Omega_{p(k)}^{t-1}\}$;
11 $\quad \mathcal{H} \leftarrow \mathcal{H} \cup \Omega_k^t$;
12 $\quad$ **for** *each* $\varepsilon_{i,j} \in \psi_k$ **do**
13 $\quad\quad \mathcal{G}' = \mathcal{G}_k \setminus \varepsilon_{i,j}$;
14 $\quad\quad (\psi', p') \leftarrow solve(\mathcal{G}')$;
15 $\quad\quad$ **if** ψ' *exists and is valid* **then**
16 $\quad\quad\quad \mathcal{A} \leftarrow \mathcal{A} \cup (\mathcal{G}', \psi', p')$;
17 $\quad\quad$ **end**
18 $\quad\quad$ **for** *each* $\varepsilon_{u,v} \in \mathcal{G}_k$ **do**
19 $\quad\quad\quad$ **if** $\mathbf{z}_i = \mathbf{z}_u \oplus \mathbf{x}_j = \mathbf{x}_v$ **then**
20 $\quad\quad\quad\quad \mathcal{G}_k \leftarrow \mathcal{G}_k \setminus \varepsilon_{u,v}$;
21 $\quad\quad\quad$ **end**
22 $\quad\quad$ **end**
23 $\quad$ **end**
24 **end**
`/* generation of the best` k `hypothesis done, with` $(k = K \vee \mathcal{A} = \emptyset)$ `*/`
25 **return** $\mathcal{H}, k$;

Figure 3.7: The multi parent k-best hypotheses generation algorithm.

3.8.3 Further Pruning Strategies

Pruning is essential to any practical implementation of the MHT algorithm. Next to the k best approach that limits the number of nodes in the hypotheses tree to k for every time step t two other pruning strategies are frequently applied.

The "N-scan-back" algorithm proposed by Kurien [1990] assumes that any ambiguity occurring during the tracking process at time t is resolved at the time $t+N$. In other words, N defines the number of frames in a time window considered to look ahead into the future and explore the evolution of the world in order to resolve the ambiguities. The strategy is briefly explained as follows. Let $\mathcal{H}^t = \{\Omega_1^t, \ldots, \Omega_k^t\}$, $\mathcal{H}^{t+N} = \{\Omega_1^{t+N}, \ldots, \Omega_k^{t+N}\}$ be the hypotheses at time t and $t+N$, respectively. For each $\Omega_i^t \in \mathcal{H}^t$ the probability mass of all hypotheses $\Omega_j^{t+N} \in \mathcal{H}^{t+N}$ arising from the Ω_i^t is calculated. Whichever hypothesis receives the highest probability is retained. All others are pruned. This strategy results is an irrevocable decision regarding the assignments of observations to tracks. Consequently, below the decision time t there is a tree of depth N with $N \cdot k$ hypotheses at maximum. Above the decision time the tree has degenerated into a simple list of hypotheses. A visualization of a hypotheses tree with a maximum number of hypotheses at each frame of $k = 5$ and a scan-back depth of $N = 30$ is shown in Figure 1.

The second strategy, called "ratio pruning", defines a lower threshold on the ratio on the worst and best hypotheses probabilities. In case of little or no ambiguity theirs is no need to consider all k hypotheses in particular. Moreover, in combination with N-scan-back pruning hypotheses have only a finite number of iterations to prove their correctness. If their children fail to collect the largest proportion of the probability mass they are deleted after N iterations. Therefore, ratio pruning determines a lower threshold p_{min} on the probabilities that prevents hypotheses from being considered if their probability drops below that threshold. Two variants of ratio pruning are known. A global strategy that defines the threshold based on the best hypotheses, thus $p_{min} = \theta\, p(\Omega_1^t)$. And a local one that defines parent-dependent thresholds, thus $p_{min}^l = \theta\, p(\Omega_{1,l}^t)$, where $\Omega_{1,l}^t$ is the best hypothesis emerged from parent Ω_l^{t-1}. However, ratio pruning has to be employed carefully. In most situations the correct hypothesis receives a very low probability for the first time – e.g. when a track needs to be deleted – making the definition of a appropriate threshold θ tough.

3.8.4 Memory Efficient Data Structures and Run-time Experiments

This section proposes a memory efficient data structure to represent the assignment matrix $\mathcal{M}$ and presents experimental results on the run-times needed to generate the best hypothesis using different assignment algorithms.

While the previous sections focus on an efficient generation of the k best hypotheses the memory consumption was neglected so far. But it can be critical as a large number of assignment matrices need to be stored in the pool $\mathcal{A}$ of cached assignment problems (line 7 in alg. 3). To reduce the memory consumption a sparse matrix representation is introduced. As shown in Figure 3.2, 3.3, and 3.4 the assignment matrices are not of arbitrary shape but consists of six blocks. Only the upper left block is dense in general and represents the $N_{t-1} \times M_t$ possible assignments of tracks to observations. The lower right block is dense as well, indeed, but contains fixed entries of 1.0 not needed to be kept in memory. The remaining four block are diagonal matrices representing $2 \cdot N_{t-1}$ track labels for *occlusions* and *deletions* and $2 \cdot M_t$ observation labels for *new tracks* and *false alarms*, respectively. In summary, instead of maintaining the full assignment matrix $\mathcal{M}$ with all entries that scales squared in the number of tracks and observations (its memory complexity is $O(2N^2+2M^2+5NM)$[24]) only the relevant entries are represented. The sparse representation scales

[24] Time indices have been removed for the sake of readability.

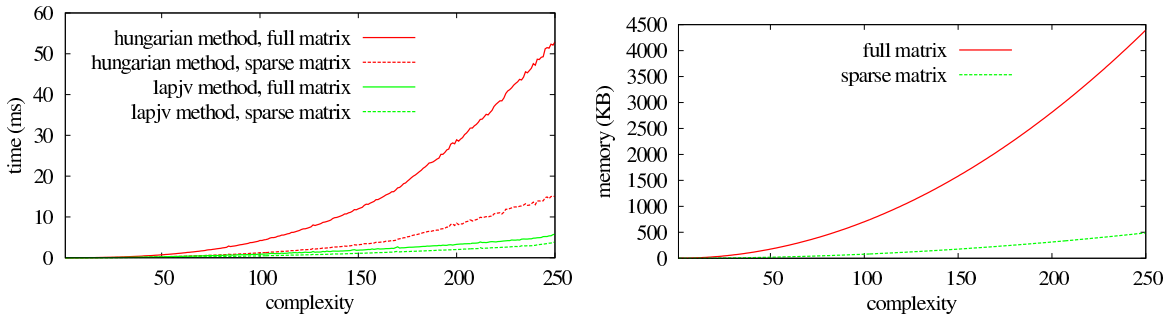


Figure 3.8: Comparison of run-time and memory consumption to find the best solution to a linear assignment problem of given complexity. *Left:* Runtime of the Hungarian method (shown in red) and the linear assignment solver by Jonker and Volgenant [1987] (called lapjv, shown in green) for general matrices (solid lines) and optimized for the sparse assignment matrix representation (dashed lines). *Right:* Memory consumption of the full (solid red line) and the proposed sparse assignment matrix (dashed green line).

linear with the number of tracks and observations and yields $O(2N + 2M + NM)$. Unfortunately, the memory reduction comes at the expense of an additional index check when accessing the specific entries of the matrix.

An experimental evaluation shows that the sparse matrix also improves the run-times of the assignment algorithms. Average run-times to solve the linear assignment problem of given complexity and the memory consumption are shown in Figure 3.8. The method proposed by Jonker and Volgenant [1987] using the sparse matrix representation outperforms all other approaches.

3.9 Conclusions

In this chapter the theoretical background on multi-hypothesis tracking is provided and the mathematical formulations of the hypotheses probabilities is derived. After introducing the original formulation and reviewing the advances made in the past two further extensions of the MHT are introduced. The first extension allows to consider spatio-temporal target priors while the second addresses spatio-temporal target probabilities. Both extensions are fundamental to integrate spatial, temporal, and especially social information into the MHT. How such information is learned, modeled, and used to guide the hypotheses generation is inspected in the remainder of this thesis.

Besides the theoretical background, several implementation strategies are outlined in this section. An adaptation of Murty's assignment generation algorithm allows to generate the k-best child hypotheses of multiple parents jointly. Furthermore, an efficient re-optimization approach enables to find those assignment sets faster based on previous solutions. In addition, a memory efficient representation of the assignment matrices needed in the assignment generation algorithm is proposed. Experimental evaluations showed that efficient re-optimization and smart memory management reduces the runtime of the system dramatically.

The multi-parent assignment generation proposed in algorithm 3 is suitable for parallelization, thus future work includes the analysis of tracking accuracy using thousands of hypotheses. The benefit of such massive parallelization is shown in Ganapathi et al. [2010]. Their system is able to estimate the pose and configuration of a human body in a stream of depth images efficiently evaluating more than 50,000 body configuration candidates per second using a GPU-accelerated implementation.

Part II

Social and Spatio-Temporal Constraints: Model-Based Approaches

4 Human Motion Prediction from Social Forces

For many tasks in populated environments, robots need to keep track of current and future motion states of people. Most approaches to people tracking make weak assumptions on human motion such as constant velocity or constant acceleration. But even over a short period, human behavior is more complex and influenced by factors such as the intended goal, other people, objects in the environment, and social rules. This motivates the use of more sophisticated motion models for people tracking especially since humans frequently undergo lengthy occlusion events.

In this chapter, computational models developed in the cognitive and social science communities are considered that describe individual and collective pedestrian dynamics for tasks such as crowd behavior analysis. In particular, a model based on a social force concept is integrated into a multi-hypothesis target tracker. It is shown how the refined motion predictions translate into more informed probability distributions over hypotheses and finally into a more robust tracking behavior and better occlusion handling. In experiments in indoor and outdoor environments with data from a laser range finder, the social force model leads to more accurate tracking. Measured using the CLEAR MOT metrics and compared to a baseline tracker that assumes constant motion the presented approach reduces the number of data association errors by 72% and the number of false positive targets by up to 55%.

This chapter is structured as follows. Introduction and related work are presented in section 4.1 and section 4.2, respectively. Section section 4.3 introduces the theory of the social force model followed by section 4.4 describing how the model can be employed to compute refined motion predictions. The integration of these motion predictions into the MHT framework is presented in section 4.5. In section 4.6 the experimental results are presented followed by the conclusions in section 4.7.

4.1 Introduction

People tracking is a key technology for mobile robots to be safely and efficiently deployed in populated environments. Most related work on people tracking like Arras et al. [2008], Cui et al. [2005], Fod et al. [2002], Schulz et al. [2003] make weak assumptions on the motion of humans and employ either the Brownian model or the constant velocity motion model. The former makes no assumptions about the target dynamics, the latter assumes linear target motion and constant velocity. Both models predict the future states of people merely based on the history of their past states.

However, human motion is more complex and follows non-random, non-linear patterns. People are usually driven by an inner motivation towards some goal, they are influenced by obstacles and other people along their path, and follow social rules. In other words, human motion is influenced by *inner motivation*, *social rules* in the presence of other people, and the *physical* and *social* constraints of the environment. Social forces offer a concept well suited to model all these aspects in a sound and common framework. However, a detailed description on how these influences on human motion can be modeled using social forces is presented in the following sections.

4.2 Related Work

Better motion models that predict the short-term and long-term goals of people and that employ these predictions to support the tracking algorithm have been proposed. The short-term prediction refers to a one-step (or N-step) ahead prediction while the long-term prediction estimated the final destination of the tracked people.

In Foka and Trahanias [2002] robot movements are controlled by a Partially Observable Markov Decision Process (POMDP). Short term predictions are obtained by a Polynomial Neural Network (PNN) trained off-line with an evolutionary method. Potential final destinations are defined manually, in advance, on a map representation of the environment. Hence, the long-term prediction refers to the prediction of the destinations a person is going to approach. Utilizing a persons tangent vector and field-of-view the probabilities of the destination can be calculated. Finally, short- and long-term predictions, are integrating into the reward function of the POMDP.

In Bruce and Gordon [2004] common destinations are learned on training trajectories of people recorded in the same environment. These trajectories are cluster by hand into groups that roughly follow the same behavior and end at the same locations. Goals are identified as the end points of clustered trajectories. Human motion is then predicted along paths computed by a Markov decision process (MDP) planner from the actual location of the person being tracked to the estimated goal location.

The approach of Vasquez et al. [2009] learns motion patterns and goals incrementally using Growing HMMs (GHMMs). Learning is performed on-line on complete sequences of observations assuming that the last observations corresponds to the persons goal. The structure of the GHMM is estimated using a topological map representing the goal positions in the environment. The parameters are learned employing an adapted incremental EM approach. Motion prediction is performed by propagating the persons state estimate multiple time steps into the future.

Maximum entropy Inverse Reinforcement Learning (IRL) is employed in Ziebart et al. [2009] to model the goal-directed trajectories of people. To this end a softened version of the value iteration algorithm is utilized to obtain distributions over actions and trajectories. This allows to infer the posterior distribution of destinations. Future trajectories of people are then predicted by computing the conditional probabilities of any path continuing their motion.

Liao et al. [2003] extract a Voronoi graph from a map of the environment and constrain the states of people to lie on the edges of that graph. The motion of people is then predicted along those edges following the topological shape of the environment. For tracking particle filters are employed.

Bennewitz et al. [2005] learns typical motion patterns that people follow in an environment. The approach collects trajectories of people with multiple statically mounted laser scanners and combines them to motion patterns using EM clustering. From each pattern a HMM is derived that enables a mobile robot to predict the long-term motion of people towards a goal as well as their short-term motion over the next few time steps.

On a bigger scale Ashbrook and Starner [2003] and Liao et al. [2007] track people using global positioning system (GPS). Both approaches extract significant locations from user traces by detecting places where the GPS signal is lost, which is equivalent with people moving indoors. The latter extends the definition of goal locations to be those locations where people typically spend extended periods of time, indoors or outdoors.

In Chapter 6 an approach to learn spatial priors on human motion in an environment using a non-homogeneous spatial Poisson process is proposed. The process is learned by observing people, leading to spatial distributions of where people usually walk, appear or disappear. A place-dependent motion model is then derived from the Poisson process using a sample-based approach.

Models for pedestrian dynamics have also been developed and applied in communities such as

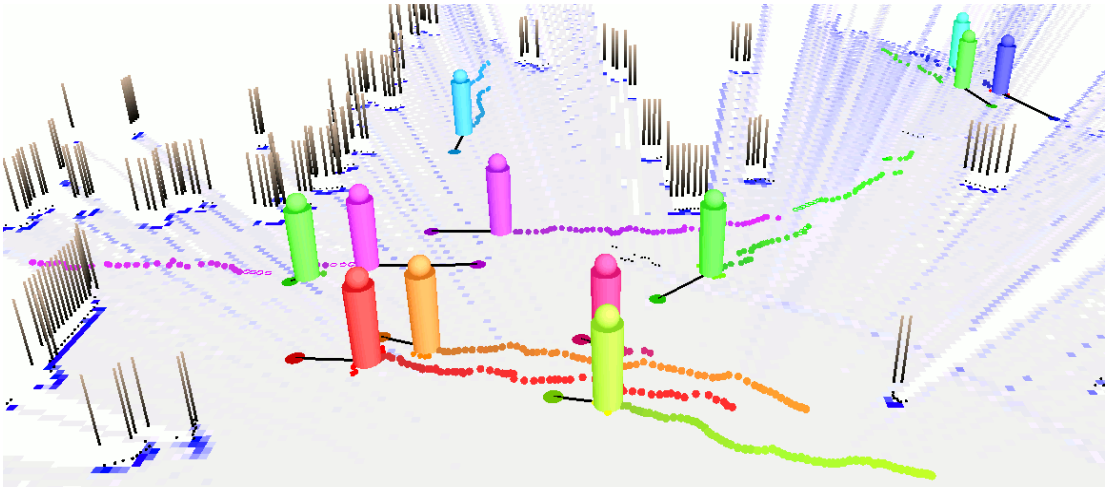


Figure 4.1: Twelve of 150 tracks from the outdoor experiment. The cylinders show the estimated positions of the pedestrians, the colored dots illustrate their past trajectories. The circles connected with lines are the virtual goals. The occupancy grid map is overlaid where darker blue cells indicate higher occupancy probability. Vertical lines visualize static objects.

quantitative sociology or spatial cognition. They are used for crowd simulation, evacuation dynamics, or building design (Figure 4.2). Schadschneider et al. [2009] present a taxonomy of various models. A first group of approaches employ fluid-dynamic and gas-kinetic models like in Hoogendoorn and Bovy [2000] in which people are considered particles with their motion being described by fluid-dynamic equations. These approaches are typically deterministic and force-based. This is unlike methods based on cellular automata that are discrete, rule-based dynamical models. They discretize space into cells that can be occupied by at most one person. The dynamics is usually described by a set of rules specifying the probability of moving to the neighboring cells. An extension of this approach is the floor field approach Burstedde et al. [2001], where the transition probability of cells are not fixed but vary dynamically. Ali and Shah [2008] combine a floor field approach with an evacuation model to improve people tracking. However, given their discrete nature, such motion models cannot be readily applied within a probabilistic tracker that requires proper error propagation in the state prediction step. Robin et al. [2009] developed a pedestrian model and its calibration methodology based on random utility theory and a discrete-choice model. Their model can be included in the category of cellular-based models, as it adopted a context-dependent cellular system that is varied according to situations. A drawback is that utilizing this model requires considerable effort in setting up the context-dependent cellular system and computing the probability that each alternative cell will be chosen.

The social force model proposed by Helbing et al. [2002] is a deterministic continuum model in which interactions between pedestrians are described using the concept of social forces or social fields. These forces model different aspects of motion behaviors, such as the motivation of people to reach a goal, the repulsive effect of walls and other people as well as physical constraints.

In this chapter a people tracker is combined with a pedestrian dynamics model for the purpose of more realistic human motion predictions. Among the existing methods, the social force model is choosen as it is a simple yet powerful approach with justifications from social psychology. Its building block, the social force, is well explained by psychological and social insights. State-of-the-

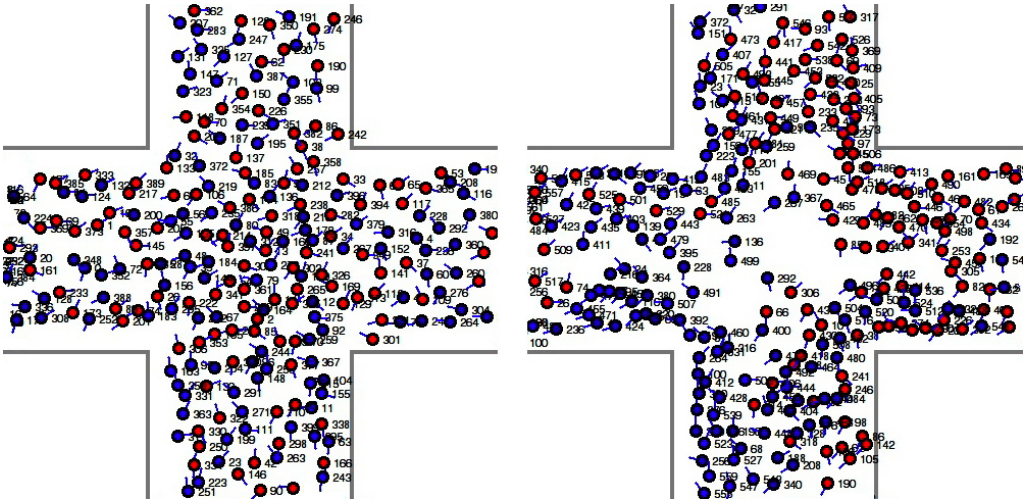


Figure 4.2: Example simulation of crowd behavior in a crossing using the social force model. Two groups of pedestrians (shown as red and blue circles) try to traverse the crossing to reach the opposite corridor. Over time, lanes in a twirl-like pattern emerge in a bottom-up fashion as the most efficient motion strategy (left $t = 0$, right $t = 2$ min).

art of human motion prediction for tracking is extended by two points. First, a single out-of-the-box model is employed that is able to coherently describe several aspects of pedestrian dynamics such as intention towards a goal, constraints from the environments and from other people. This is in contrast to previous work that uses combinations of methods, such as goal learning and planning Bruce and Gordon [2004], Voronoi graph extraction from a map Liao et al. [2003] or EM clustering and HMMs Bennewitz et al. [2005]. The model requires no learning step. Secondly, the proposed model deals with inter-people relations for motion prediction. During an occlusion event of two persons approaching each other, for instance, the model will not predict their future states into the other person (or into an object). This is an important aspect, as shown in the experiments, that has not been previously considered.

Independently developed from this work, Pellegrini et al. [2009] recently proposed the social force model for motion prediction in the context of visual people tracking. In their approach, the person velocity is also accounted for in the energy potential. This is achieved by estimating the closest future distance in the space-time trajectory of targets and use this distance as an additional potential. This enables the system to plan ahead to some extent but comes at the expense of the loss of a closed-form solution. Unlike Pellegrini et al. [2009] this work considers additional forces from the environment by maintaining a short-term environment model and computing different repulsion forces and physical constraints from static obstacles. Subsequently, their contribution are compared with those from inter-person influences only.

4.3 The Social Force Model

The social force model introduced by Helbing et al. [2002, 2000] is a computational model in which the interactions between pedestrians are described by using the concept of a social force. It is based on the idea that changes in behavior can be explained in terms of social fields or forces. Applied to pedestrians, the social force model accounts for the influence of the environment and other people

and describes how the intended direction of motion changes as a function of these influences. The model does not cover cases of multiple options, when people have to actively decide. Game theoretic approaches can be applied in such situations.

4.3.1 Personal Intentions

Formally, the models assumes that a pedestrian p_i with mass m_i likes to move with a certain *desired* or *intended velocity* $\widehat{\mathrm{v}}_i$ in an *intended direction* $\widehat{\boldsymbol{e}}_i$. In case of a deviation from the desired velocity vector $\widehat{\boldsymbol{v}}_i$ defined by the intended velocity and direction

$$\widehat{\boldsymbol{v}}_i = \widehat{\mathrm{v}}_i \, \widehat{\boldsymbol{e}}_i \tag{4.1}$$

due to necessary detours or evasive maneuvers the person tends to adapt his or her velocity $\boldsymbol{v}_i$ to approach $\widehat{\boldsymbol{v}}_i$ within a so called *relaxation time* τ_i. This change of velocity can be described by an acceleration term and is modeled by the personal motivation force

$$\mathbf{F}_i^{\text{pers}} \;=\; m_i \frac{\widehat{\mathrm{v}}_i \, \widehat{\boldsymbol{e}}_i - \boldsymbol{v}_i}{\tau_i}. \tag{4.2}$$

The relaxation time is the time interval needed to reach the intended velocity and the intended direction.

4.3.2 Interaction Forces

In the presence of other people or objects in the environment, a pedestrian might not be able to keep the intended direction and velocity. In the social force model, repulsive effects from these influences are described by an interaction force $\mathbf{F}_i^{\text{int}}$. This force prevents humans from walking along their intended direction and is modeled as a sum of forces either introduced by other individuals p_j or by static obstacles denoted by subscript o

$$\mathbf{F}_i^{\text{int}} \;=\; \sum_{j \in \mathcal{P} \setminus \{i\}} \mathbf{f}_{i,j}^{\text{int}} \;+\; \sum_{o \in \mathcal{O}} \mathbf{f}_{i,o}^{\text{int}} \tag{4.3}$$

with $\mathcal{P} = \{p_i\}_{i=1}^{N_p}$ being the set of all people and $\mathcal{O}$ the static objects of the environment. These forces decrease proportional to the distance of their sources and are modeled as

$$\mathbf{f}_{i,k}^{\text{int}} \;=\; a_k \, \exp\left(\frac{r_{i,k} - d_{i,k}}{b_k}\right) \boldsymbol{n}_{i,k} \tag{4.4}$$

where $k \in \mathcal{P} \cup \mathcal{O}$ is either a person or an object of the environment, a_k specifies the magnitude and b_k the range of the force. In order to calculate the Euclidean distance between p_i and entity k, pedestrians and objects are assumed to be of circular shape with radii r_i and r_k, respectively. Then, distance $d_{i,k}$ is given by the Euclidean distance between the centers, and $r_{i,k}$ is the sum of their radii. The term $\boldsymbol{n}_{i,k}$ is the normalized vector pointing from k to p_i which describes the direction of the force.

Given the limited field of view of humans, influences might not be isotropic. This is formally expressed by scaling the forces with an anisotropic factor

$$\mathbf{f}_{i,k}^{\text{int}} \;=\; a_k \, \exp\left(\frac{r_{i,k} - d_{i,k}}{b_k}\right) \left(\lambda + (1-\lambda)\frac{1 + \cos\left(\varphi_{i,k}\right)}{2}\right) \boldsymbol{n}_{i,k} \tag{4.5}$$

where λ defines the strength of the anisotropic factor (see Figure 4.3 for details) and

$$\cos\left(\varphi_{i,k}\right) \;=\; -\,\boldsymbol{n}_{i,k} \cdot \widehat{\boldsymbol{e}}_i \tag{4.6}$$

the deviation of the exerted force from the center of the persons field of view.

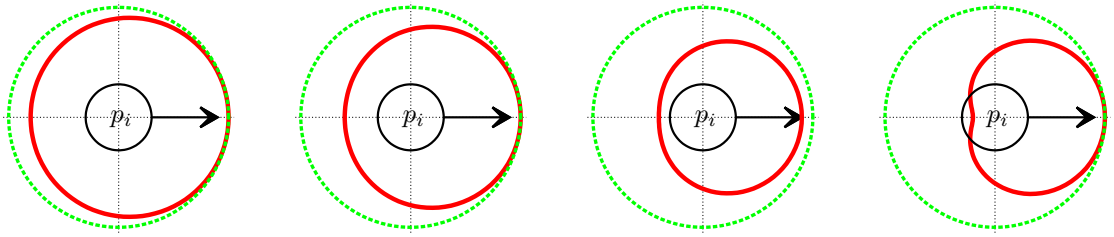


Figure 4.3: Visualization of the anisotropic factor with different strength parameter λ. Person p_i is located in the center of each plot heading right. The shape of the red plot indicates the impact strength of the exerted forces w.r.t. the impact angle. The strength parameter λ varies from left to right from 0.8, 0.6, and 0.4 to 0.2, respectively.

4.3.3 Environmental Constraints

Human motion is not only influenced by personal motivation and reactive behavior towards obstacles or other people but is physically constrained by the environment Helbing et al. [2000]. Hard constraints restrict the motion and thereby define the walkable area of the environment. Therefore, the social force model introduces a physical force $\mathbf{F}_i^{\text{phys}}$ onto pedestrian p_i described as

$$\mathbf{F}_i^{\text{phys}} = \sum_{j \in \mathcal{P} \setminus \{i\}} \mathbf{f}_{i,j}^{\text{phys}} + \sum_{o \in \mathcal{O}} \mathbf{f}_{i,o}^{\text{phys}} \tag{4.7}$$

$$\mathbf{f}_{i,k}^{\text{phys}} = c_k \; g(r_{i,k} - d_{i,k}) \; \boldsymbol{n}_{i,k}, \tag{4.8}$$

where c_k represents the magnitude of the exerted force. To make the physical forces a real contact force where the circular shapes of p_i and k overlap, the function g is defined as $g(x) = x$ if $x \geq 0$ and 0 otherwise.

Finally, human motion is explained by the superposition of all exerted forces. Accordingly, the force $\mathbf{F}_i$ changing the motion of individual p_i

$$\mathbf{F}_i = \mathbf{F}_i^{\text{pers}} + \mathbf{F}_i^{\text{int}} + \mathbf{F}_i^{\text{phys}}. \tag{4.9}$$

Using $\mathbf{F}_i$, the basic equation of motion for a pedestrian is then of the general form

$$\frac{d}{dt} \boldsymbol{v}_i = \frac{\mathbf{F}_i}{m_i} \tag{4.10}$$

and describes the movements of p_i over time. An illustration of all forces is shown in Figure 4.4. The physical force $\mathbf{f}_{k,o}^{\text{phys}}$ that the wall exerts onto person p_k is shown. This avoids motion predictions through walls. A superposition of different forces onto pedestrian p_i is also shown. The person wants to keep his or her intended velocity through the motivation $\mathbf{F}_i^{\text{pers}}$ but is also influenced by $\mathbf{f}_{i,j}^{\text{int}}$ from person p_j and by $\mathbf{f}_{i,o}^{\text{int}}$ from the wall. This results in the superimposed force $\mathbf{F}_i$ used to adapt the velocity of p_i.

4.4 Motion Prediction Using Social Forces

Good motion models are particularly important for people tracking as people frequently undergo lengthy occlusion events during interaction with each other or with the environment. In this section, it is shown how the social force model can be combined with a Kalman filter based tracker to result in a more realistic prediction model of human motion.

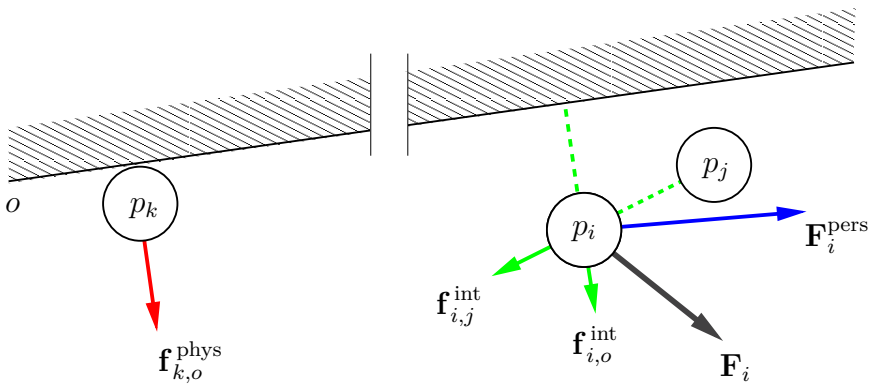


Figure 4.4: Forces in the social force model shown as colored arrows. The pedestrian p_k is only affected by the physical force $\mathbf{f}^{\text{phys}}_{k,o}$ emitted by the static part o of the environment. The force is shown as red arrow. The pedestrian p_i is both affected by the wall o and an other person p_j. The interaction forces emitted by the wall $\mathbf{f}^{\text{int}}_{i,o}$ and the other person $\mathbf{f}^{\text{int}}_{i,j}$ are shown in green. The personal motivation $\mathbf{F}^{\text{pers}}_i$ of p_i is shown as blue arrow. The superposition of all forces exerted to p_i is shown as the black arrow $\mathbf{F}_i$.

Let $\mathbf{x}_t = (\ x_t \ \ y_t \ \ \dot{x}_t \ \ \dot{y}_t\ )^T = (\ \boldsymbol{x}_t \ \ \boldsymbol{v}_t\ )^T$ be the state of a pedestrian p_i at time t and Σ_t its 4×4 covariance matrix estimate. The term $\boldsymbol{x}_t$ represents the position and $\boldsymbol{v}_t$ the velocity of the pedestrian in Cartesian space. The *continuous white noise acceleration model* – often called *constant velocity motion model* as it describes an object that moved with constant velocity – is then defined as

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \ = \ \mathcal{N}(\mathbf{x}_t;\, A_t\, \mathbf{x}_{t-1}, A_t\, \Sigma_{t-1}\, A_t^T + Q_t), \tag{4.11}$$

with A_t being the constant velocity state transition matrix and Q_t the process noise Matrix defined as

$$A_t \ = \ \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad Q_t \ = \ \begin{pmatrix} 1/3\Delta t^3 & 0 & 1/2\Delta t^2 & 0 \\ 0 & 1/3\Delta t^3 & 0 & 1/2\Delta t^2 \\ 1/2\Delta t^2 & 0 & \Delta t & 0 \\ 0 & 1/2\Delta t^2 & 0 & \Delta t \end{pmatrix} \tilde{\sigma}. \tag{4.12}$$

The entries of Q_t represent the acceleration capabilities of a human modeled with the continuous-time process noise intensity $\tilde{\sigma}$. For more details on the continuous white noise acceleration model see [Bar-Shalom et al., 2002, p. 269–270].

This model is extended by considering how the state of the pedestrian at a generic time t is influenced by its previous state, other people $\mathcal{P}$ and static obstacles $\mathcal{O}$. A discrete time approximation of Eq. 4.10 within a fixed interval of time Δt[25] is used to obtain $\mathbf{x}_t = \xi(\mathbf{x}_{t-1}, \mathcal{P}, \mathcal{O})$, where

$$\xi(\mathbf{x}_{t-1}, \mathcal{P}, \mathcal{O}) \ = \ \begin{bmatrix} \boldsymbol{x}_{t-1} + \boldsymbol{v}_{t-1}\, \Delta t + \frac{1}{2} \frac{\mathbf{F}}{m} \Delta t^2 \\ \boldsymbol{v}_{t-1} + \frac{\mathbf{F}}{m} \Delta t \end{bmatrix} \tag{4.13}$$

describes how the motion of a pedestrian evolves over time. The change in motion is calculated according to the pedestrian's intended velocity, reactive behavior from interaction forces and physical

[25] Usually, Δt is the cycle time between t and $t-1$.

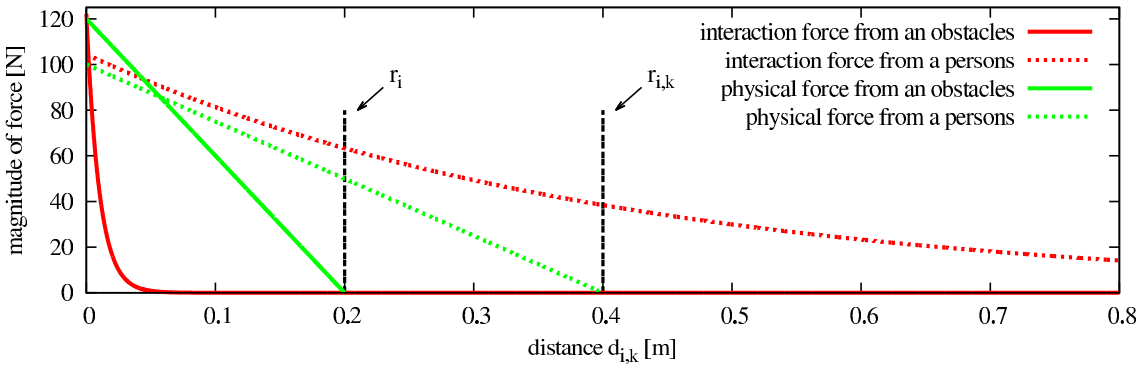


Figure 4.5: Typical functions for the exerted social forces. The x-axis shows the distance from person p_i to an object o or a person p_k in meter and the y-axis shows the magnitude of the forces in Newton. The radius of p_i is $r_i = 0.2\ m$ and the sum of the radii of p_i and p_k is $r_{i,k} = 0.4\ m$. Interaction and physical forces are shown in red and green, respectively.

constraints from the environment, according to Eq. 4.9. Assuming that the motion is affected by Gaussian noise with zero mean and covariance matrix Q yields

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathcal{P}, \mathcal{O}) \;=\; \mathcal{N}\left(\mathbf{x}_t;\; \xi(\mathbf{x}_{t-1}, \mathcal{P}, \mathcal{O}), J_\xi\, \Sigma_{t-1}\, J_\xi^T + Q\right), \tag{4.14}$$

where

$$J_\xi \;=\; \frac{\partial \xi(\cdot)}{\partial \mathbf{x}} \tag{4.15}$$

is the Jacobian of $\xi(\cdot)$ evaluated at $\mathbf{x}_{t-1}$.

Without external stimuli and deviation from the intended direction and preferred velocity, $\mathbf{F}$ is zero, no social forces are applied to the pedestrian and the social force model falls back onto the constant velocity motion model. The computation of the individual forces is explained in the following sections.

4.4.1 Estimating Short-Term Intentions

The social force model is based on the assumption that each pedestrian has an intended direction of motion and a preferred velocity, both as a result of a higher level goal. However, estimating this goal from the observation of a tracked person involves the problem of intention or activity recognition, an issue beyond the scope of this work. Since the focus is on short-term human motion prediction, the weak assumption is made that subjects will continue to pursue his or her short-term intention keeping the current velocity vector.

In doing so, the concept of *virtual goals* is introduced. A virtual goal is defined as the hypothetical position that a person would reach if he or she moved by keeping the current velocity. Virtual goals also hold the expected time $t_\mathbf{g}$ until which the person would attain the goal. In other words, the position of the individual p_i is projected into the future for a short time interval (proportional to the tracker cycle time) and denoted *goal ahead time* $\Delta g = k\Delta t$. The location of the virtual goal $\mathbf{g}$ and the expected goal time $t_\mathbf{g}$ are defined as

$$\mathbf{g}_t \;=\; \boldsymbol{x}_{t_o} + \boldsymbol{v}_{t_o}\,(t - t_o + \Delta g), \tag{4.16}$$

$$t_\mathbf{g} \;= t + \Delta g, \tag{4.17}$$

where t_o denotes the time when the person was last observed and t the actual time. Note that the virtual goal is not fixed but moves along the last estimate of the velocity vector as time goes by. This behavior is needed as the duration of an occlusion event is not known in advance and therefore the virtual goals has to be adjusted on-the-fly.

Once $\mathbf{g}_t$ is estimated, the intended direction $\widehat{\boldsymbol{e}}_t$ and velocity $\widehat{\mathrm{v}}_t$ can be calculated using the offset between $\mathbf{x}_t$ and $\mathbf{g}_t$ and the expected goal time $t_{\mathbf{g}}$ as

$$\widehat{\boldsymbol{e}}_t \;=\; \frac{\mathbf{g}_t - \boldsymbol{x}_t}{\|\mathbf{g}_t - \boldsymbol{x}_t\|} \quad \text{and} \qquad \widehat{\mathrm{v}}_t \;=\; \frac{\|\mathbf{g}_t - \boldsymbol{x}_t\|}{t_{\mathbf{g}} - t}. \tag{4.18}$$

For long term motion prediction the assumption of a virtual target destination can be generalized to goals that can either be learned by observing motion trajectories like in Bennewitz et al. [2005] and Vasquez et al. [2009] or calculated from map representations as done in Bruce and Gordon [2004] ar the spatial affordance map proposed in Chapter 6.

4.4.2 Estimating Social Interactions

While moving towards their virtual goal, the movements of the person are affected by the surrounding environment according to Eq. 4.3. The position $\boldsymbol{x}_p$ of all $p \in \mathcal{P}$ is provided by the tracking system, where each existing track is considered to be a separate individual. As for the static objects, it is assumed that no representation of the environment is known in advance, thus it is estimated using an occupancy grid framework. The grid is built by using the most recent laser observations (the last 30 in the experiments) and discarding points that are detected as people from the detection algorithm. When the occupancy probability of a cell is above a predefined threshold, it is consider to be occupied and inserted in the set of static obstacles $\mathcal{O}$. The center of this cell is then used as its position $\boldsymbol{x}_o$.

Once the surrounding people are known and the environment is learned, the interaction force exerted from $k \in \mathcal{P} \cup \mathcal{O}$ can be calculated following Eq. 4.5, where the normalized vector pointing from k to p_i is $\boldsymbol{n}_{i,k} = \frac{\boldsymbol{x}_i - \boldsymbol{x}_k}{\|\boldsymbol{x}_i - \boldsymbol{x}_k\|}$. Since the laser range finder is sensing the surfaces of the obstacles, no additional tolerance is needed. Hence, they are assumed to have no radius, i.e. $r_k = 0$, and the sum of the radii $r_{i,k}$ is set to the radius of the individual p_i.

The grid is meant to be a short-term memory. In case of a mobile sensor, subsequent laser observations need to be registered into the common reference frame of the grid, either by using odometry, scan matching or map-based localization.

4.4.3 Estimating Physical Forces

Physical forces according to Eq. 4.7 model the close contact interaction between two rigid bodies and express the principle that two different bodies cannot occupy the same space. The position of a person $\boldsymbol{x}_p$ and of a static object $\boldsymbol{x}_o$ and the direction of the force $\boldsymbol{n}_{i,k}$ are obtained in the same way as for the interaction forces described above. Physical forces are effective only when a physical contact is present (model by the function $g(\cdot)$ in Eq. 4.8).

The choice of having an occupancy grid instead of raw data as static obstacles has the advantage of constant force density. Using raw data, the density would vary over different parts of the environment, making the system behave differently. With an occupancy grid, it is possible to adjust and control the density of static obstacles by the cell size.

4.5 Integration into the Multi-Hypothesis Tracking Framework

The integration of the social force based motion prediction into the multi-hypothesis tracking framework is explained in this section. The MHT approach described in section 3.5, based on the work of Arras et al. [2008] is adapted. Briefly, Multi-Hypothesis Tracking hypothesizes about the state of the world by considering all statistically feasible assignments between observations and tracks and all possible interpretations of observations as false alarms or new track and tracks as matched, occluded or obsolete. A hypothesis Ω_l^t is one possible set of assignments and interpretations at time t.

The probability of each hypothesis (see Eq. 3.28) depends among others on the likelihood of track-to-observation assignments. These likelihoods are modeled using Gaussian distributions

$$\mathcal{N}(\mathbf{z}_i(t)) \;:=\; \mathcal{N}\left(\mathbf{z}_i(t);\, \hat{\mathbf{z}}_j(t),\, S\right), \tag{4.19}$$

centered around the position of the observation $\mathbf{z}_i(t)$. The measurement prediction (or predicted position of the track j) is $\hat{\mathbf{z}}_j(t)$ and S the innovation covariance matrix. The measurement prediction $\hat{\mathbf{z}}_j(t)$ is now conditioned on the surrounding people $\mathcal{P}$ and static obstacles $\mathcal{O}$ and calculated using Eq. 4.14. In more detail

$$\hat{\mathbf{z}}_j(t) \;=\; H\, p\left(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathcal{P}, \mathcal{O}\right), \tag{4.20}$$

where $H = \begin{Bmatrix} 1\ 0\ 0\ 0 \\ 0\ 1\ 0\ 0 \end{Bmatrix}$ is the measurement Jacobian.

As the motion of an individual track is now influenced by other tracks in a hypothesis Ω_i^t, motion predictions cannot be done on a per-track basis but must be done on a per-hypothesis basis. The track tree proposed by Kurien Kurien [1990] is a data structure that holds tracks from all hypothesis to exploit the fact that frequently, several hypothesis share the same tracks. This optimization technique has to be abandoned with the social force model leading to increased memory requirements of our implementation.

4.6 Experiments

To experimentally evaluate the social force model and analyze its behavior, it is integrate within an MHT-based people tracker. Note that although the MHT is used as a tracking approach, the model is general and can be integrated in any existing tracking method. During the prediction of a single pedestrian the state estimates of the other pedestrians are assumed to be constant.

The tracking performance between the constant velocity motion model, the social force model, and the social force model where only inter-human influences are considered is compared using the CLEAR MOT metrics intruduced by Bernardin and Stiefelhagen [2008]. The metric counts three numbers with respect to ground truth : misses (missing tracks that should exist at a ground truth position, FN), false positives (tracks that should not exist, FP), and mismatches (track identifier switches, ID). The latter value quantifies the ability to deal with occlusion events that typically occur when tracking people. From these numbers the tracking accuracy is determined: MOTA (average number of a correct tracking output). The numbers are calculated based on the best hypotheses returned by the people tracker. Most important is the number of identifier switches as this number indicates the ability to deal with ambiguities that occur when people walk in groups and in case of lengthy occlusion events. Especially when tracking people in 2D range data ambiguities are very likely as targets are of identical appearance.

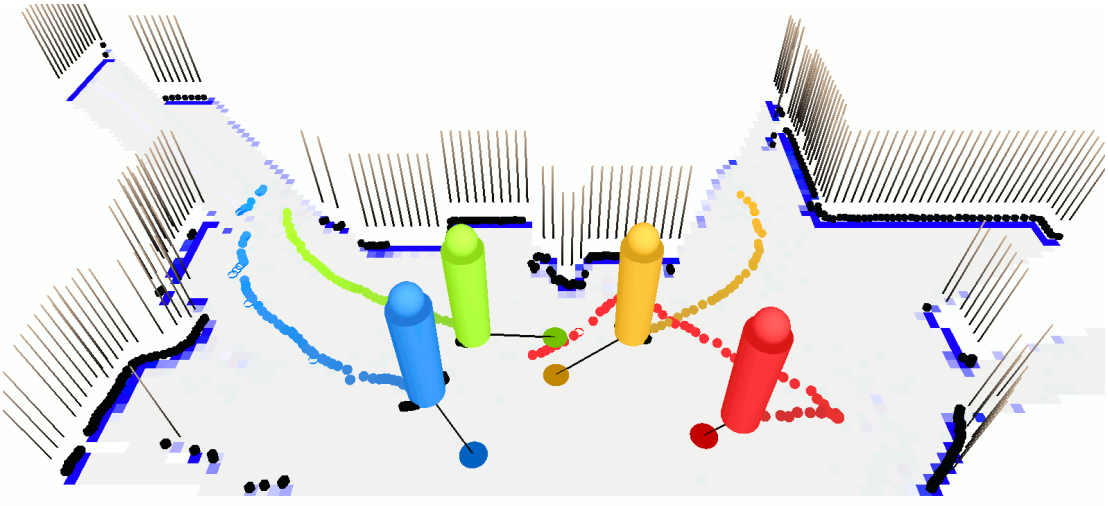
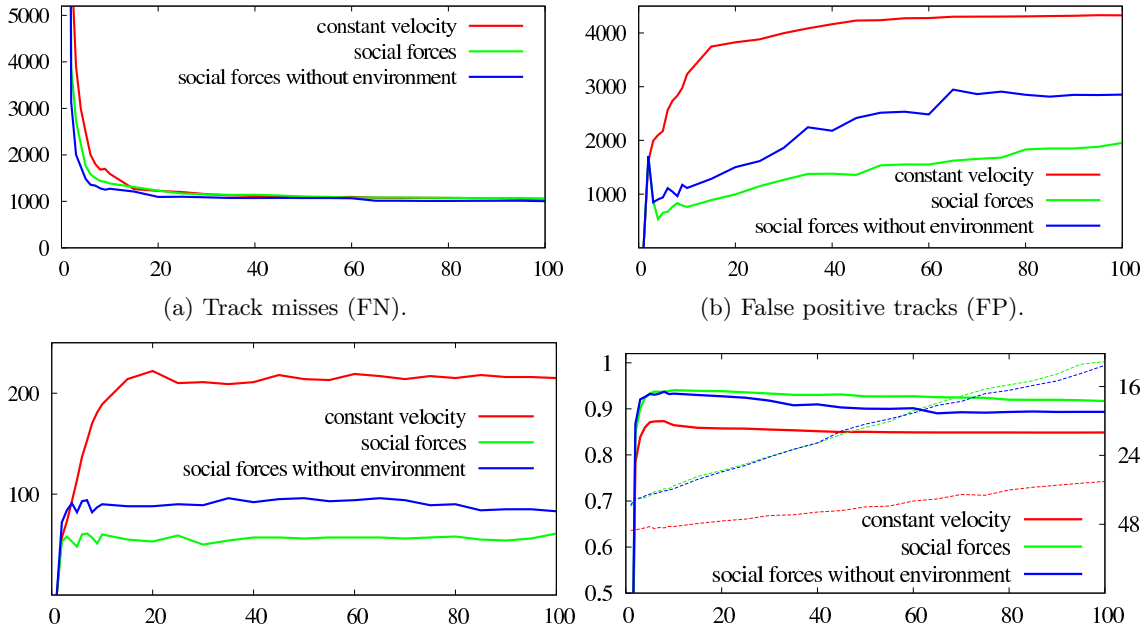


Figure 4.6: Four of 135 tracks from the indoor experiment. The cylinders show the estimated positions of the pedestrians, the colored dots illustrate their past trajectories. The circles connected with lines illustrate the virtual goals. The occupancy grid map is overlaid where darker blue cells indicate higher occupancy probability. Vertical lines visualize static objects.

Three data sets have been collected: the first one in a laboratory (Figure 4.6), a second one in the city center of Freiburg (Figure 4.1), and the third one at the Freiburg main station. The data sets have been recorded at 12Hz using a fixed SICK laser range finder with an angular resolution of 0.5 degree mounted at $\sim$0.85 meter above the floor. The k-best pruning strategy is employed to limit the maximum number of hypotheses generated by the MHT at every step to N_{Hyp}. The generation of the k best hypotheses is done using the multi-parent variant (see algorithm 3 for details) of the pruning algorithm proposed by Murty [1968]. In order to show the evolution of the error as a function of the computational effort, N_{Hyp} is varied from 1 to 100 in the experiments.

The social force model needs to be calibrated to reflect the observed behavior of pedestrians. Ko et al. [2012] propose an automatic calibration framwork that adopts maximum likelihood estimation to learn the parameters of the social force model on statistical information like speed, density, and the flow rate of the observed pedestrian trajectories. Since their calibration results differ with the density of the observed pedestrian crowds the parameters used in the following experiments have been determined experimentally from a set of calibration trials. Some of the person-dependent parameters are not directly observable with a laser range finder. Therefore, they are assumed to be fixed and identical for all individuals. More specifically, each person is defined to have a radius of 0.2 meter and a mass of 80 kg. The anisotropic blending factor of Eq. 4.5 is defined to be 0.5, for penalizing forces coming from the back of persons. The goal ahead time is set to 60 tracking cycles (that is 5 seconds) and the relaxation time to 0.5 seconds. The parameters for the social forces exerted by obstacles are $a = 100\,N$ and $b = 0.01\,m$, respectively. Social forces between persons are modeled with $a = 70\,N$ and $b = 0.4\,m$. Physical forces from obstacles have a magnitude of $c = 600\,{}^{N}/_{m}$ where the magnitude of the physical forces between persons is $c = 250\,{}^{N}/_{m}$. The functions are plotted in Figure 4.5.

The parameters of the MHT, that are in detail the probabilities of detections, occlusions, deletions and the fixed rates for false alarms and new tracks, respectively, have been learned from another



(a) Track misses (FN).

(b) False positive tracks (FP).

(c) Identifier switches (ID).

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

Figure 4.7: CLEAR MOT analysis of the indoor data set. Total number of (a) track misses, (b) false positive tracks, (c) identifier switches, and (d) multi object tracking accuracy (MOTA) and average cycle times as functions of the number of used hypotheses N_{Hyp} varied from 1 to 100. Red lines show the results using the constant velocity motion model, blue indicates inter-human influences only, and green the social force based motion prediction. With $N_{Hyp} = 100$ the number of FP and ID decrease by 54.9% and 71.9%, respectively, while the number of FN is almost constant. MOTA increases from 84.8% to 91.7%.

training data set with 95 tracks over 28,242 frames. In detail, $p_{det} = 0.7$, $p_{occ} = 0.27$, $p_{del} = 0.03$, $\lambda_{new} = 0.0002$, and $\lambda_{fal} = 0.005$, respectively. To detect people the cascade of spatial informed detectors in Chapter 2 is used. Each stage employs 50 decision stumps as weak learners. The detector has also been learned on a separate training set and is evaluated in section 2.5.

4.6.1 Indoor Environment

The indoor data set has been conducted in a laboratory and consists of 38,994 frames with a total number of 135 people entering and leaving the sensor field of view. All participants were instructed to show usual human behavior. They walk at different speed, stay in front of desks and tables, and interact with each other. Both, detections and data association have been labeled by hand to yield ground truth information. The results of the CLEAR MOT metrics analysing the MHT employing either the constant velocity model, the social force model, or social forces without environmental influences for motion prediction are presented in Figure 4.7.

The results show a clear improvement of the MHT using the social force motion model over the regular approach using constant velocity motion prediction. The number of false positive tracks and identifier switches decreases dramatically while the number of track misses remains almost the same. Consequently, the overall accuracy (MOTA) increases increses from 84.9% to 91.7%. In detail, using 100 hypotheses the number of false positives decreases from 4328 to 1954 (improvement of 54.9%),

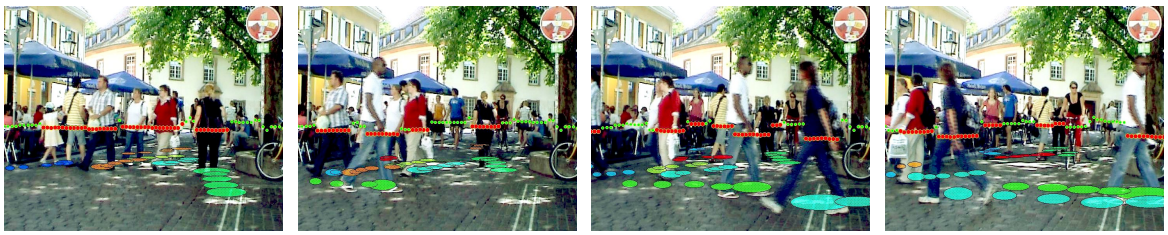


Figure 4.8: Four images of the outdoor experiments carried out in the city center of Freiburg. The information of the laser range finder and the tracking system are projected onto a recorded video sequence. The images from left to right show the tracking results at timesteps $t = 335, 353, 365$ and 381. Laser range measurements are shown as small green dots (background) or small red dots (detected pedestrians). The traces of the observed pedestrians are drawn with colored ellipses.

identifier switches drop from 217 to 61 (improvement of 71.9%), misses remain at 1055 and 1057, respectively. The results of the approach using the social force model with inter-human influences only (no environmental constraints are respected) are in between of the baseline and the proposed approach. Detailed information can be found in Table 4.1.

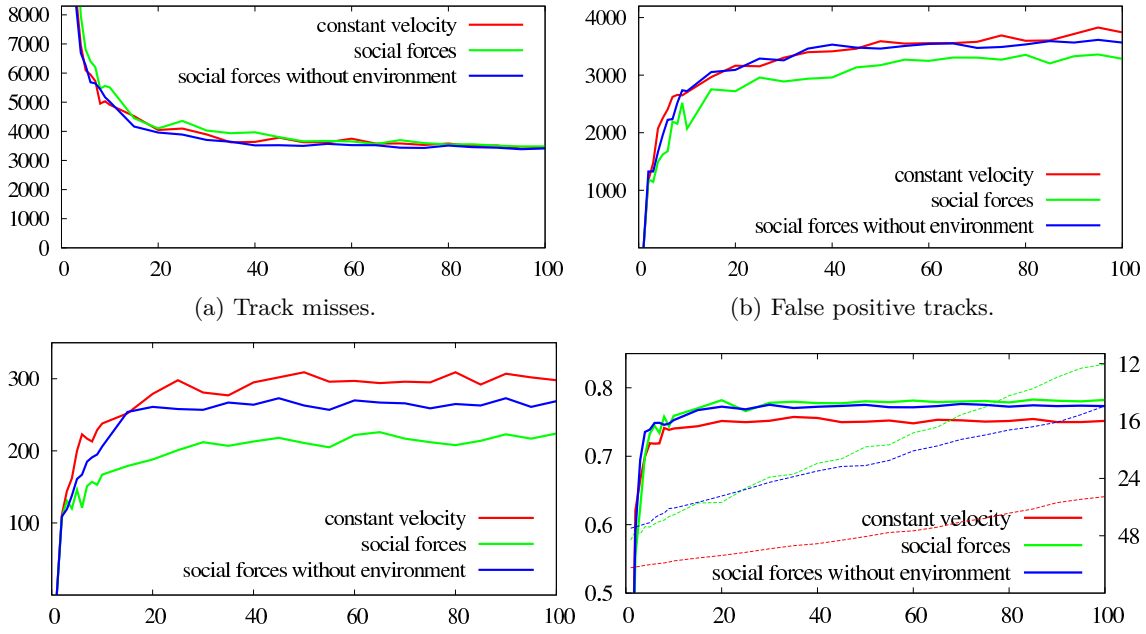
The improvements are explained by two facts. First, during lengthy occlusion events the tracker is able to avoid predictions into other people or walls leading to a more likely confirmations of tracks when they reappeared. This reduces the number of false positives as no second track of a reappearing person is generated. Therby also no identifier switch occurs. Second, ambiguities in case of interactions are avoided. While the constant velocity motion model predicts people independently and neglects other people's positions the social force model respects them and maintains the distances and spatial relations of people in groups. The insight of this behavior is that better motion predictions lead to smaller innovations in the Kalman filter which translates to higher likelihoods as now the observations are better explained by the predictions. This leads also to less identifier switches of these people and less false positives as no "ghost" tracks are created.

The tracking improvement comes to the expense of a lower run-time. While the baseline approach is able to track people in indoor environments with 29.5 Hz (assuming data is available, immediately), the run-time of the social force approach drops to 14.2 Hz. The reasons are two fold. The social force based motion prediction conditioned on other people and the environment requires more computational effort. But the main reason is the loss of the memory and run-time efficient track-trees proposed by Kurien [1990]. Predicting people independently tracks can be shared in multiple hypotheses, thus their motion needs to be calculated only once. Is the motion of people is influenced by other tracks, predictions must be done on a per-hypothesis basis.

4.6.2 Outdoor Environments

The second experiment has been carried out in the city center of Freiburg during a regular workday. The data set consists of 55,475 frames during 25 minutes. 10,000 frames with 162 persons have been labeled by hand, again to determine the ground truth detections and data associations. Occlusions are even more likely in such an scenario as more people move in a larger space with many obstacles like trees and trash cans. The resulting tracking accuracies are shown in Figure 4.9 and Table 4.1.

While the improvements by the social force model are not that significant anymore the results of the outdoor experiment are comparable to those of the the indoor experiment. Especially the number of identifier switches is reduced by 24.8% which indicates the ability to deal with lengthy occlusion



(a) Track misses.

(b) False positive tracks.

(c) Identifier switches.

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

Figure 4.9: CLEAR MOT analysis of the Freiburg city center data set. With 100 hypotheses the number of track misses is almost constant (+1.3%) while the number of false positives and identifier switches decrease by 12.3% and 24.8%, respectively. The tracking accuracy (MOTA) increases from 75.1% to 78.2%. The presented approach can still be applied in real time. But the average run-times decrease from 28.6 to 12.1 Hz.

events and ambiguities that typically occur when people walk in groups. Also the number of false positives decreases (−12.3%) showing that a socially constraint motion prediction avoids the creation of "ghost" tracks. The lower improvements are explained by a smaller ratio of people interacting with each other and their environment. Furthermore, the contribution of the forces exerted by static obstacles is rather small. Applying forces from interacting people only, the number of ID switches decreases by 9.7% and the number of false positives by 4.8% in comparison to the constant velocity motion model. This result is explained by the public scene with lots of open space where interactions with the environment occur less often.

The third experiment was conducted at the Freiburg main station shortly after the arrival of a train such that many of the people passing the monitored space carry luggage. The data sets consist of 33,204 frames during 15 minutes. 6,000 frames with 160 persons have been annotated, again to determine ground truth information. The tracking results are shown in Figure 4.10 and Table 4.1.

Similar to the results on the Freiburg city center data set the number of identifier switches (−25.6%) and false positives (−28.3%) decreases significantly. Different than in the previous experiments, the number of missed tracks increases by 6.8%. This unexpected result is explained by carried luggage that is sometimes treated as part of the static environment. In these situations, people appear to melt together with the environment causing strong physical forces. As the social force model prevents people from walking into the limitations of the environment people with huge luggage are sometimes not tracked correctly. If the environmental forces are deactivated this behavior does not occur and the number of misses decreases by 4.9% in comparison to the baseline method. However, additional information on the map or a detector of the luggage could help to resolve that conflict.

(a) Track misses.

(b) False positive tracks.

(c) Identifier switches.

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

Figure 4.10: CLEAR MOT analysis of the Freiburg main station data set. Using 100 hypotheses the number misses increases by 6.8% while the number of false positives and mismatches decrease by 28.3% and 25.6%, respectively. The accuracy (MOTA) increases from 79.4% to 82.1%. The average run-time drops from 27.6 to 12.1 Hz.

The overall influence of the environment is slightly bigger than in the inner city but smaller in comparison to the indoor environment. This is due to the solid structurs of the station hall that limit the walkable area and resamble an indoor environment. On the other hand, the main station provides more open space than a laboratory.

4.7 Conclusions

In this chapter human motion prediction that incorporates previously observed states and accounts for personal interests, the influence of other people, and the constraints of the environment is addressed.

The presented social force based human motion prediction overcomes the rather simple assumptions typically made in related work. Usually, people are assumed to walk with constant velocity into a constant direction. Moreover, they are treated to behave independently without any interferences or interactions to other people or the environment. Hence, the accuracy of the predicted behavior is limited. However, the proposed approach extends state-of-the-art by taking into account these influences for human motion prediction using a model that incorporates personal interest as well as for environmental and social constraints. This is achieved with a single coherent model able to describe complex motion with a sound mathematical formulation. It is based on a general and analytical model which easily allow to perform error propagation, an important property for tracking algorithm such as the Kalman filter. The model does not require to learn a set of goals or human motion patterns, and can share the same set of parameters over different environments and situations.

Data set	Approach	FN	FP	ID	MOTA	*Hz*
laboratory	baseline	1055	4328	217	84.8%	29.5
	without env.	1004 (-4.8%)	2853 (-34.1%)	83 (-61.8%)	89.4%	14.5
	social forces	1057 (+0.2%)	1954 (-54.9%)	61 (-71.9%)	91.7%	14.2
Freiburg city center	baseline	3440	3744	298	75.1%	28.6
	without env.	3412 (-0.8%)	3566 (-4.8%)	269 (-9.7%)	77.3%	14.7
	social forces	3484 (+1.3%)	3284 (-12.3%)	224 (-24.8%)	78.2%	12.1
Freiburg main station	baseline	3006	3327	473	79.4%	27.6
	without env.	2868 (-4.9%)	3114 (-7.3%)	401 (-15.2%)	80.7%	16.1
	social forces	3213 (+6.8%)	2386 (-28.3%)	352 (-25.6%)	82.1%	12.1

Table 4.1: CLEAR MOT results of all data sets using $N_{Hyp} = 100$ hypotheses. Employing the social force based motion prediction the number of identifier switches (ID) can be increased dramatically. This improvement comes to the expense of a lower frame-rate.

Further, it is presented how the social force model can be incorporated into any Kalman filter-based tracker. The integration into the multi-hypothesis tracking framework is discussed in detail.

The experiments, carried out in indoor and outdoor environments, demonstrate that the presented approach reduces the number of data association error by up to 72%. Especially, in case of interacting people and lengthy occlusion events social forces aid to predict people correctly and to resolve ambiguities. Further, the model prevents people from being predicted into the solid structures of the environment reducing the number of false positive tracks by up to 55%. It was found that the contribution of the interaction forces from the environment is rather small with respect to the forces from other people. This outcome is noteworthy and underlines the importance of considering other people in the prediction of human motion.

In future research, friction forces and attractive forces expressed within the social force model should be analyzed to improve motion prediction for people walking in groups. Moreover, long term motion prediction can be achieved by human activity recognition that provides a set of global and local goal locations as well as adjusted force parameters.

5 Modeling Place Dependent Prior and Target Probabilities

People tracking is a key component for robots that operate in populated everyday environments. In previous works different target filtering, motion prediction, and data association techniques have been employed for the purpose of people tracking. These techniques typically rely on a set of simple and generic assumptions on target behavior, detector characteristics, environmental constraints, and social rules, respectively.

This chapter focuses on these assumptions rather than the tracking approach itself and show that with informed models, people tracking can be made substantially more accurate without compromising efficiency. Concretely, more accurate, human-specific models for the occurrence of new tracks, false alarms, track occlusions, and track deletions are presented. In the experiments with large-scale outdoor data sets collected with a laser range finder, the models and combinations thereof are experimentally compared using a multi-hypothesis baseline tracker and the CLEAR MOT metrics. The results show how some models selectively improve tracking performance at the expense of other measures. The final combination is then able to resolve these trade-offs, leading to a reduction of data association errors by more than a factor of two.

This chapter is structured as follows. The next section provides a brief introduction followed by a review of related work in section 5.2. Sections 5.3 and 5.4 introduce the proposed observation and target-specific models, respectively. In section 5.5 a short description of the integration of these models into the MHT framework is provided. Experimental results are presented in section 5.6 followed by the conclusions in section 5.7.

5.1 Introduction

As robots enter domains in which they interact and cooperate closely with humans, people tracking is becoming a key technology for many research and application areas in robotics and related fields.

The task has been addressed with a variety of general-purpose target tracking techniques that employ different filtering, motion prediction, and data association schemes. Typically, these systems make generic assumptions about the target behavior, the detector characteristics, and (less often) environmental constraints. But for people as targets, some of these assumptions are overly simplistic and ignore important information that is available. For example, new track events are usually assumed to be uniformly distributed over the sensor field of view occuring with a constant probability. But the way how people move is often given by static environmental constraints that can be modeled. Indoors, for instance, doors or convex corners are typical places where people appear. The same place-dependency applies for the behavior of the detector. Regions of clutter and complex background produce false alarms more likely than open space, making a spatially uniform model a poor approximation. Modeling occlusion and deletion events with uniform and constant probabilities are further examples of poor assumptions. Clearly, former are more likely to occur in the shadow of static obstacles or other people. Latter can be interpreted as the counterpart of new track events and occur more often at the border of the sensor field of view when people are leaving.

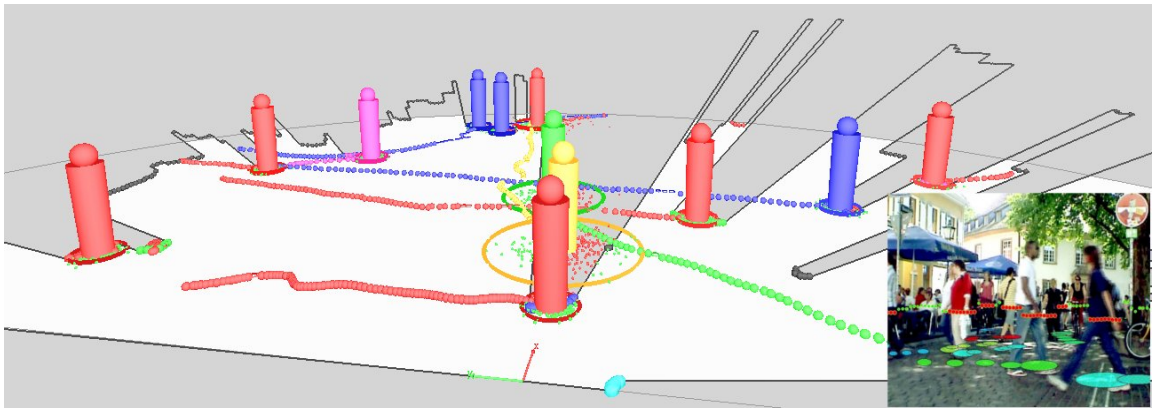


Figure 5.1: Twelve of 162 people observed during the Freiburg city center experiment. The cylinders show the estimated positions of the pedestrians, the bigger colored dots illustrate their past trajectories. The black outer contour shows the border of the visible space indicated by white ground. The occluded area is shown in gray. Further, small dots denote position samples that are either visible (shown in green) or invisible (colored in red).

The techniques that have been employed for people tracking in previous works include Kalman filter (KF) based motion estimates and nearest-neighbor data association as proposed by Kluge et al. [2001] and Fod et al. [2002]. Motion prediction using particle filters and a sample-based variant of the Joint Probabilistic Data Association filter (JPDAF) are employed in Schulz et al. [2003]. A KF-based tracker in which the data association is formulated as a Minimum Description Length problem and solved using Quadratic Boolean Programming proposed by Leibe et al. [2008]. As in the works of Taylor and Kleeman [2004], Mucientes and Burgard [2006], and Arras et al. [2008] the KF-based multi-hypothesis tracking framework is applied in this thesis as well.

Tracking people becomes particularly challenging if the targets are identical in appearance which is typically the case for tracking using radar or laser range finders. With a good, target-specific appearance model, many hard tracking problems such as dealing with occlusions and interactions of tracks, becomes much easier to cope with. For this reason, visual tracking systems, where rich appearance information is available, can achieve good results with nearest-neighbor data association filters as in Breitenstein et al. [2009] using a set of independent particle filters. However, in this chapter, as targets are detected using a 2D laser scanner, they are assumed to have identical appearance.

5.2 Related Work

Every tracking system needs to mark observations as detected, new tracks, or false alarms and has to deal with detected, occluded, and deleted tracks. The people tracking literature is reviewed with respect to how these tracking events have been modeled.

In the work of Schulz et al. [2003] a local occlusion grid introduced by Moravec and Elfes [1985] and Elfes [1989] is proposed to determine the probability of tracks being occluded or observations being missed in a sample-based JPDAF framework. This is implemented by an additional label that represents situations in which an object has not been detected due to occlusions and detection failures. Taylor and Kleeman [2004] track the legs of a single person in laser range data using the MHT framework. Based on the relative positions of the legs to one another and to the sensor, the occlusion probability is approximated with a piecewise linear model. Therby, implicitly modeling

the fact that legs frequently occlude each other. Following this idea Arras et al. [2008] reformulate the MHT expressions to make the occlusion probability an explicit parameter. Then they track multiple people by separately tracking legs using Kalman filters and adapt the occlusion probabilities of those tracks that are recognized to belong to a person. However, a correcly adapted occlusion probability relies on the recognition of people given the set of leg tracks. Latter is based on a set of simple heuristics. In Katz et al. [2008] probabilistic occlusion checking is used to improve the robustness of a motion detection algorithm. The occlusion probability of each track is computed by a sample-based visibility check. A set of particles is drawn from the position estimate of the tracks. The visibility of each particle is verified using ray-tracing. A similar model has also been used by Mucientes and Burgard [2006]. For the purpose of vision-based multi-person tracking Ess et al. [2009b] model occlusions in a occlusion grid map, keeping tracks alive that are known to be hidden by other tracks and static objects as a hard decision. With the exception of Arras et al. [2008], all these works compute the final occlusion probability on a per-track basis by a geometric visibility test that determines if or how far a track is 'in the shadow' of other tracks or static objects. In Taylor and Kleeman [2004] this is realized using a piecewise linear model, all others use samples to this end.

For track terminations or deletions, one can assume a constant deletion probability as in the regular MHT approach by Reid [1979]. Counting the number of consecutive non-confirmations or non-detections of a track and deleting it when it exceed a threshold has been done e.g. in the work of Breitenstein et al. [2009]. Following the approach of Cox and Hingorani [1996] that increase the deletion probability of each track using an exponential function Mucientes and Burgard [2006] track clusters of people in laser range data, modeling the probability of deleting a track from a cluster with an exponential decay function. This simulates the decrease in probability of detecting a target that could not be assigned to an observation in several consecutive frames. In Schulz et al. [2003], the same principle was realized using a discounted average weight of the particles that automatically decreases when tracks are no longer confirmed. Weak tracks in this sense are then deleted if a mismatch with the number of observations is encountered. In Lin et al. [2004] and Wieneke and Willett [2008] the track score based on a likelihood ratio of deleting or not deleting the track is computed. If this score falls below a given threshold the track is deleted. For the purpose of face tracking Duffner and Odobez [2011] employ a HMM to estimate a hidden status variable indicating that a face is visible or not. The HMM state evolves based on the confidence of the face detector and the previous trajectory, the observation likelihood of the mean state, and the maximum of the variances in x and y direction of the particle distribution used to track the face.

Duffner and Odobez [2011] use a similar approach to decide when to add new targets. Here, the HMM is based on the face detector output and a long-term memory of the positions of tracked faces. The arrival of new tracks is modeled by Schulz et al. [2003] using a Poisson process with constant rate over time and a uniform distribution over space. The same assumptions are taken in the regular MHT approach. A simple form of place-dependency has been realized by Breitenstein et al. [2009], a visual surveillance scenario with a static camera, where a predefined area around the border of the image has been manually put to describe the region where new tracks may appear. It is assumed that no new tracks arrive in the center of the image.

The false alarms model that is employed in most related works, e.g. in Khan et al. [2006], predicts spurious measurements uniformly over the sensor field of view. This is also the assumption in the original MHT and its extensions. Another usual strategy is to detect false alarms directly during the detection step. In the context of body part based people detection in video surveillance data Corvee and Bremond [2010] declare observations as false alarms if the number of detected body parts is below a specific threshold.

In Chapter 6 an approach on learning place place-dependent Poisson rate functions that describe

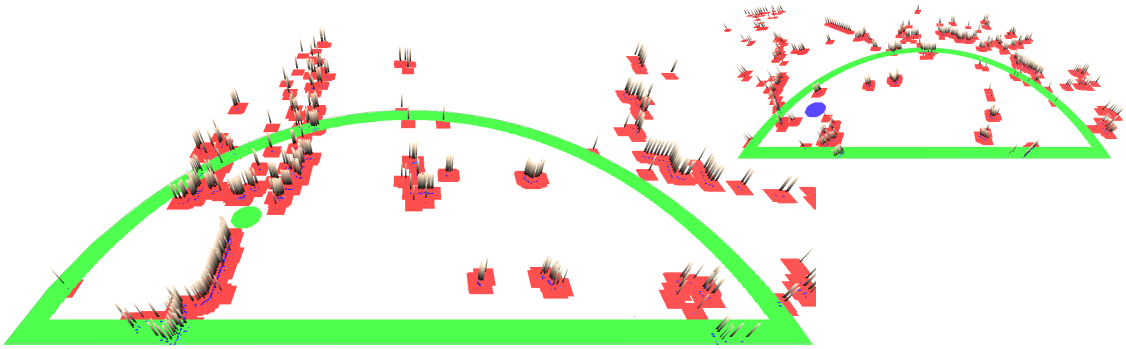


Figure 5.2: Models of the Freiburg city center (*left*) and Freiburg main station (*right*) environments learned from annotated raw data using an occupancy grid map. Static obstacles (like buildings, plants, and chairs in front of a restaurant) are illustrated by dark horizontal lines. Cells shown in green indicate the area V_{new}^{high} of increased new track probability. Red cells denote the area V_{fal}^{high} of increased false alarm probability, respectively. The blue area in to top right image denotes an escalator but is on purpose not marked with increased new track probability.

the occurrence of false alarms and new tracks is proposed. This approach overcomes the assumption that these events are uniformly distributed in space (and time) and is able to model that people typically appear at specific locations in the environment and that false alarms occur more likely in places with clutter. Furthermore, no manual annotation of the environment or simplistic assumptions as in Breitenstein et al. [2009] are required.

In this section, the approach of predefined areas of false alarms and new tracks is combined with a sample-based occlusion model that incorporates geometric information from the scene, and a deletion model that assumes exponentially distributed interarrival times of observations. The state of the art where these models have been considered only in isolation is extended by a systematic experimental review of the effects of each model and their combinations. large-scale experiments with challenging data sets, collected in the city center of Freiburg and its main station have been carried out. Futhermore, the different models are compared using the CLEAR MOT metric proposed by Bernardin and Stiefelhagen [2008]. The insights gained in this section are valid for people tracking in general regardless the sensor modality, the filtering approach, or the space in which targets are represented.

5.3 Observation-Specific Models

During tracking, there are situations where a sensor observation cannot be explained by any of the currently existing tracks. The observation is therefore either spurious (a false alarm or false positive) or a new target that entered the sensor field of view. It is typically impossible to determine which of the two interpretations is correct from a single scan or image. Instead, probabilities for both events can be computed, and if the tracking framework is able to integrate data association and interpretations over time, decisions can be taken in a delayed fashion. In order to compute probabilities, models are required that predict how often and where new tracks and false alarms occur. As mentioned in the previous section, the general assumption is that new tracks and false alarms occur both uniformly over the sensor field of view at rates that follow a Poisson distribution. In this section, environment-specific models are addressed that aid to reason about observations

to be declared as false alarms from clutter or new tracks. These models have to be created for each environment, specifically, to provide correct and useful place-dependent information. How such information can be learned from human observations is properties in Chapter 6.

5.3.1 New Track Model

The assumptions that new tracks occur uniformly over the sensor field of view and at constant Poisson rates may be valid for traditional setups in target tracking in which airborne targets are sensed by an upwards looking radar or setups that do not use a target detector. For people, however, this model does not account for the place-dependent character of human behavior and the place-dependent character of visual or range-based people detectors. People typically appear, disappear, walk and stand at specific locations that correspond, for instance, to doors, elevators, or convex corners. Furthermore, people detectors have limited field of view and limited range such that people typically are detected at the border of the surveillance area and at specific distances, respectively.

To implement place-dependency new track events are modeled using a spatial Poisson process (see Merzbach and Nualart [1986]) that predicts their probability using a spatial rate function $\lambda_{new}(\vec{x})$ with $\vec{x} \in X$ where X is a vector space such as $\mathbb{R}^2$ or $\mathbb{R}^3$. For any subset $S \subset X$ of finite extent (e.g. a spatial region), the number of events occurring inside this region can be modeled as a Poisson process with associated rate function λ_S such that

$$\lambda_S = \int_{\vec{x} \in S} \lambda_{new}(\vec{x})\, d\vec{x}. \tag{5.1}$$

The formally used homogeneous Poisson processes with fixed rate parameter λ_{new} is now conditioned on the position of the currently investigated observation $\mathbf{z}_i(t)$, thus the place-dependent expected number of new track events yields $\lambda_{new}(\mathbf{z}_i(t))$ with

$$\lambda_{new}(\mathbf{z}_i(t) \mid V_{new}) \;=\; \begin{cases} \lambda_{new}^{high} & \text{if } \mathbf{z}_i(t) \in V_{new}^{high}, \\ \lambda_{new}^{low} & \text{otherwise.} \end{cases} \tag{5.2}$$

The probability $p_{new}(\mathbf{z}_i(t))$ has a sound physical interpretation as

$$p_{new}(\mathbf{z}_i(t) \mid V_{new}) \;=\; \lambda_{new}(\mathbf{z}_i(t) \mid V_{new}) \cdot \begin{cases} V_{new}^{high} \\ V_{new}^{low} \end{cases} = \begin{cases} \lambda_{new}^{high}\, V_{new}^{high} & \text{if } \mathbf{z}_i(t) \in V_{new}^{high}, \\ \lambda_{new}^{low}\, V_{new}^{low} & \text{otherwise.} \end{cases} \tag{5.3}$$

and is modeled by the average rates of events per volume multiplied by the observation volume V_{new}^{high} or V_{new}^{low}, respectively. The volumes can be approximated by discretizing the environment into a bi-dimensional grid, where each cell $c_{i,j}$ represents a local homogeneous Poisson process with adapted but fixed rates λ_{new}^{high} or λ_{new}^{low}, respectively. Using a grid approximation the place-dependent new track rate $\lambda_{new}(\mathbf{z}_i(t))$ is no longer multiplied by the entire volume but multiplied with the volume of the cells V_c to yield the new track probability, hence

$$p_{new}(\mathbf{z}_i(t) \mid V_{new}) \;=\; \lambda_{new}(\mathbf{z}_i(t))\; V_c. \tag{5.4}$$

The volumes $V_{new} = \left\{V_{new}^{high}, V_{new}^{low}\right\}$ are modeled using an approach similar to Breitenstein et al. [2009]. A predefined area around the border of the monitored area and at the maximum detection range of the applied people detector (see Chapter 2) is marked automatically to describe the region where new tracks may appear. Additionally, environment-specific locations – like doors, corners,

and passages – in which new tracks occur more often are annotated manually. Furthermore, it is assumed that no new tracks arrive in the center of the surveillance area, thus $V_{new}^{low} = V \setminus V_{new}^{high}$. Nevertheless, to model uncertainty and imperfect information $\lambda_{new}^{low} > 0$ is set to a small positive value. An example of the automatically and manually annotated areas of increased new track rates (shown in green) is given in Figure 5.2.

5.3.2 False Alarm Model

As mentioned in the previous section, people are assumed to appear at specific locations in the environment. The same assumption is true for the occurrence of false positive events as people detectors are more prone to false positives in areas of background clutter and at locations of objects with a target-like appearance, leading to systematic misdetections. Therefore, a second spatial Poisson process is employed to model place-dependent false alarm events. The same theory as introduced above applies, thus

$$\lambda_{fal}(\mathbf{z}_i(t) \mid V_{fal}) = \begin{cases} \lambda_{fal}^{high} & \text{if } \mathbf{z}_i(t) \in V_{fal}^{high}, \\ \lambda_{fal}^{low} & \text{otherwise,} \end{cases} \tag{5.5}$$

$$p_{fal}(\mathbf{z}_i(t) \mid V_{fal}) = \lambda_{fal}(\mathbf{z}_i(t))\ V_c,$$

where V_{fal}^{high} is the area of increased false alarm probability occurring with an adapted but fixed Poisson rate of λ_{fal}^{high}. On the other hand, it is assumed that $V_{fal}^{low} = V \setminus V_{fal}^{high}$ denotes the area in which false alarm event occur less often. The lower false alarm Poisson rate is denoted by $\lambda_{fal}^{low} > 0$ which is set to a small positive value. Again, an approximation using a piecewise homogeneous Poisson process modeled with a bi-dimensional grid with cells of volume V_c can be applied.

The volumes $V_{fal} = \left\{V_{fal}^{high}, V_{fal}^{low}\right\}$ are modeled based on the information of an occupancy grid map learned from annotated ground truth data. The occupancy grid map introduced by Elfes [1989] provides a probabilistic tessellated representation of spatial information. Using a multidimensional tessellation of the space into 2D cells, each cell stores a probabilistic estimate of its state by counting the numbers of being hit (#hits) and missed (#miss) by the beams of the laser range finder. Traced cells are calculated using the ray tracing algorithm of Bresenham [1965] and hit cells are defined by the points in which an laser beams ends. Those laser end points that have been manually marked to belong to a person are filtered out. The occupancy probability of a cell c is defined as $p(c) = {}^{\#\text{hits}}/_{(\#\text{hits}+\#\text{miss})}$. An optimal estimate of the state of a cell c is given by the maximum a posteriori (MAP) decision rule

$$p(c = \text{OCC}) = \begin{cases} 1 & \text{if } \#\text{hits} > \#\text{miss}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.6}$$

From the learned occupancy map containing the static parts of the environment the area containing background clutter causing an increase false alarm probability can be defined. Therefore, each cell in the neighborhood of 0.5 meter of the occupied cells is marked. A visualization of the occupancy grid map and the area of background clutter is shown in Figure 5.2. Static part of the environment are illustrated by dark horizontal lines. Cells with increased false alarm probability λ_{fal}^{high} are shown in green.

5.4 Target-Specific Models

This section introduces the track-specific models considered in this work. People can disappear from the sensor field of view, either because they are behind other people or objects or because they left the sensor field of view. While those two situations can look similar – since in both cases the person is not visible – they differ in their information content. In the first case, the person is still in the surveillance area and its state must be maintained by the tracker and the person re-associated to the correct track when he/she reappears from the occlusion. In the second case, the person is outside the monitored area and the corresponding track should be removed from the tracking system. Again, it is typically impossible to determine which of the two interpretations is correct. Therefore, probabilities for both events – track occlusion or deletion – can be computed and integrate into the data association framework. The models applied to compute these probabilities are introduced in the next subsections.

5.4.1 Occlusion Model

When an existing track cannot be confirmed by an observation, the system has to decide if the track is still in the monitored area or not. Such situations can be distinguished into four cases: occlusion, interaction, missed detection, and track termination. This subsection deals with a model for the former three cases, track terminations, however, are considered in the next subsection.

Occlusion events occur when (far apart) people or static objects occult another person. Interactions are situations in which close people interact with each other, potentially changing their behavior, and appear as a single group and observation. These two events are different from detection failures that typically happen with a probability that does not depend on the past, while occlusions and interaction usually occur in an interval of time. In fact, while missed detection can be handled well by data association techniques with delayed decision taking such as MHT framework or the Markov chain methods presented by Oh et al. [2004], lengthy interactions and occlusions are notoriously challenging when targets are identical in appearance as it is the case in laser based people tracking.

The model proposed in this section aims to explain occlusions by the geometry of the scene, i.e. when people are hidden by other people or static objects. Therefore, two aspects need to be considered. First, how can the visibility of the scene be encoded in some representation V_{vis}. As presented in the related work, other authors have approached this with either simple visibility checks as in Katz et al. [2008] and Taylor and Kleeman [2004] or more complex occlusion grids or maps as in Schulz et al. [2003] and Ess et al. [2009b]. In this work, the visible space $V_{vis} \subset V$ of the surveillance area V is defined by the contour derived from the laser end points of the current laser reading $\mathcal{B} = \{b_1, b_2, \dots\}$. Each beam b_i corresponds to a tuple $b_i = (\phi_i, \rho_i)$ that defines a point in a 2D plane and it can be assumed that it observes a triangle of visible space given by the position of the scanner, $(\phi_i - \Delta_\alpha, \rho_i)$, and $(\phi_i + \Delta_\alpha, \rho_i)$ where $2\Delta_\alpha$ is the angular resolution of the sensor. The union of all triangles is then assumed to define the free space V_{vis}.

The second aspect is the knowledge about the current target position. Unlike Ess et al. [2009b] where only the first moments in a non-probabilistic manner are considered, the occlusion probability should also depend on the uncertainty of the expected target position. Thus, the targets are predicted given their past location $\mathbf{x}_{t-1}$ according using the motion model $p(\mathbf{x}_t|\mathbf{x}_{t-1})$.

Following a sample-based approach similar to Mucientes and Burgard [2006] and Katz et al. [2008], the occlusion probability p_{occ} for a track using its predicted position $\widehat{\mathbf{x}}_j(t)$ is determined as

$$p_{occ}(\widehat{\mathbf{x}}_j(t) \mid V_{vis}) \;\approx\; \frac{1}{N}\sum_{i=1}^{N} p_{occ}(x^{(i)} \mid V_{vis}), \tag{5.7}$$

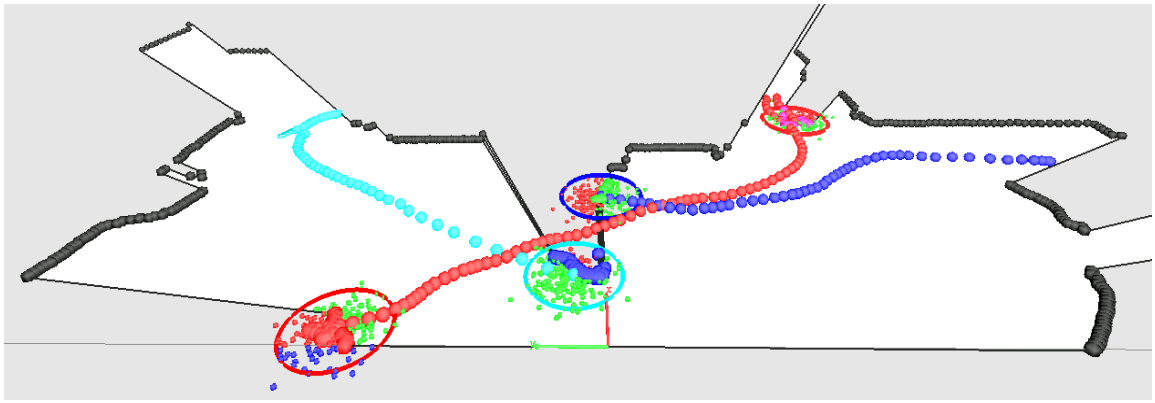


Figure 5.3: Visualization of the occlusion model for an example indoor scene. Black dots mark the laser end points connected by black lines that indicate the border of the visible area V_{vis}. State uncertainties and trajectories of four persons are shown with colored circles and dots. The small dots are particles drawn from the state predictions $\widehat{\mathbf{x}}_j(t)$. They are colored in green when they are inside the visible area and red when they are occluded. Blue particles fall outside the sensor field of view which is limited to 180 degrees. The occlusion probabilities of the tracks are (from left to right) 0.42, 0.16, 0.69, and 0.47.

where N is the number of samples $x^{(j)}$ drawn from the predicted state $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ of the track. The probability $p_{occ}(x^{(i)} \mid V_{vis})$ is determined using a simple visibility check and defined as

$$p_{occ}(x^{(i)} \mid V_{vis}) \;=\; \begin{cases} 1 & \text{if } \; x^{(i)} \in V_{vis}, \\ 1/2 & \text{if } \; x^{(i)} \notin V^{(26)}, \\ 0 & \text{otherwise, that is } \; x^{(i)} \in V \setminus V_{vis}. \end{cases} \tag{5.8}$$

The occlusion probability $p_{occ}(\widehat{\mathbf{x}}_j(t) \mid V_{vis})$ can deal with both occlusion events and interactions of people in groups. However, a sequence of missed detections is less well modeled. In such a case, the laser end points belonging to people the detector failed to classify correctly, erroneously regard a track as static objects. To overcome this fact, the final occlusion probability is modeled as a mixture of a uniform distribution (modeling misdetections as in Schulz et al. [2003]) and the occlusion model described above, hence

$$p_{occ}(\widehat{\mathbf{x}}_j(t) \mid V_{vis}) \;:=\; p_{occ}(\widehat{\mathbf{x}}_j(t) \mid V_{vis}) + (1 - p_{det}). \tag{5.9}$$

Figure 5.3 shows an example scene and the behavior of the occlusion model.

5.4.2 Deletion Model

When targets disappear from the sensor field of view, their tracks need to be declared as obsolete and removed from the tracking system. Otherwise they inflate the system and unnecessarily increase the level of data association ambiguity[27]. As discussed in section 5.2, the common approaches are

[26] Samples outside the surveillance area V are assumed to have a uniform probability over the states of being occluded or visible, thus $p_{occ}(x^{(i)} \notin V \mid V_{vis}) = 1/2$.

[27] The data association ambiguity increases dramatically over time as the uncertainty in the position estimates of unseen people grows quickly.

either a constant deletion probability as in Reid [1979] or to update some score for tracks that have not been confirmed through a sequence of steps and delete them if a specific threshold is exceeded. Latter is applied, e.g. in the works of Schulz et al. [2003], Lin et al. [2004], Mucientes and Burgard [2006], and Breitenstein et al. [2009]. While both approaches have been shown to be practical in the past, these models consider track deletions in isolation and not jointly with track occlusion events although they both try to explain non-detections of existing tracks. Therefore, the only alternative reasons for tracks being not confirmed in such an isolated model are missed detections. Similar to the approach in Mucientes and Burgard [2006], in this work the deletion probability of a track is modeled with an exponential function to simulate the decay in the probability of detecting it when it has not been matched for several consecutive iterations. More formally, let $(t - t_0)$ be the number of consecutive timesteps that a track $\mathbf{x}_j(t) \in V$ inside the surveillance area has not been observed. Its deletion probability $p^V_{del}(\mathbf{x}_j(t))$ is defined as

$$p^V_{del}(\mathbf{x}_j(t)) \;=\; 1 - \exp\left(-\frac{(t-t_0)}{\lambda^V_{del}}\right), \tag{5.10}$$

where λ^V_{del} is the speed of the decay process. The theoretical insight of this model is that Eq. 5.10 represents the cumulative density function of an exponential distribution with parameter $1/\lambda^V_{del}$. This exponential distribution thus represents the probability distribution of the interarrival times of observations – following a Poisson process model for observations. The deletion probability is then the natural result for the probability of not having observed the track after a certain duration.

In case, a person is leaving the sensor field of view V it can obviously not be detected anymore. To reflect that fact, the deletion probability of the corresponding track $\mathbf{x}_j(t) \notin V$ is adapted and set to a constant high value $p^{\notin V}_{del}$. Depending on its predicted position $\widehat{\mathbf{x}}_j(t)$ the final place-dependent deletion probability $p_{del}(\widehat{\mathbf{x}}_j(t) \mid V)$ is defined as

$$p_{del}(\widehat{\mathbf{x}}_j(t) \mid V) \;=\; \begin{cases} p^V_{del}(\mathbf{x}_j(t)) & \text{if } \widehat{\mathbf{x}}_j(t) \in V, \\ p^{\notin V}_{del} & \text{otherwise.} \end{cases} \tag{5.11}$$

The position check $\widehat{\mathbf{x}}_j(t) \in V$ is performed based on the borders of the sensor field of view and the maximum range of the people detector similar to the adapted new track rate presented in subsection 5.3.1.

5.5 Integration into the Multi-Hypothesis Tracker

In this section, the integration of the proposed place-dependent observation and track-specific models into the MHT tracking framework is presented. To be noted, the proposed models are valid for people tracking in general and can be integrated into any probabilistic target tracking framework regardless the sensor modality, the filtering approach, or the space in which targets are represented.

A detailed introduction into place-dependent people tracking using the MHT framework is presented in section 3.7 and briefly summarized in the following. The MHT algorithm hypothesizes about the state of the world by considering all statistically feasible assignments between observations and tracks and all possible interpretations of observations as *false alarms* or *new tracks* and tracks as *matched*, *occluded* or *deleted*. Thereby, the MHT handles the entire life-cycle of tracks from creation and confirmation to occlusion and deletion.

Formally, let Ω^t_l be the l^{th} hypothesis at time t and $\Omega^{t-1}_{p(l)}$ the parent hypothesis from which Ω^t_l was derived. Let $\mathcal{Z}(t) = \{\mathbf{z}_i(t)\}_{i=1}^{M_t}$ be the set of M_t observations which in this case is the set of detected people in the 2D laser range data. Let further $\psi_l(t)$ denote a set of assignments which

associates predicted tracks to observations in $\mathcal{Z}(t)$ and let $\mathcal{Z}^t = \{\mathcal{Z}(0), \ldots, \mathcal{Z}(t)\}$ be the set of all laser readings up to time t. Starting from a hypothesis of the previous time step $\Omega_{p(l)}^{t-1}$, and a new set of observations $\mathcal{Z}(t)$, there are many possible assignment sets $\psi_l(t)$, each giving birth to a child hypothesis that branches off the parent. An example hypothesis tree is shown in Figure 1.

As the number of hypotheses grows exponentially over time reasoning about all hypotheses is impossible. Thus an approach to maintain only the most important hypotheses and to prune the worse once is needed. To obtain the best hypotheses and guide the pruning algorithm each hypothesis receives a probability $p(\Omega_l^t \mid \mathcal{Z}^t)$ that is recursively calculated as the product of a normalizer η, a measurement likelihood, an assignment set probability, and the probability of the parent hypothesis known from the previous iteration, hence

$$p(\Omega_l^t \mid \mathcal{Z}^t) \;=\; \eta \;\; p(\mathcal{Z}(t) \mid \psi_l(t)) \;\; p(\psi_l(t) \mid \Omega_{p(l)}^{t-1}, \mathcal{Z}^{t-1}) \;\; p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}), \tag{5.12}$$

where the last term is known from the previous iteration. More details on the measurement likelihood and the assignment set probability can be found in subsection 3.3.1 and subsection 3.3.2, respectively.

In the following, assignment set dependent indicator variables are used to mark detected observations matched to tracks with $\tau_i \in \{0, 1\}$ (and tracks matched to observations with δ_j, respectively), false alarms with ϕ_i, and new track events with ν_i. Occluded tracks are indicated using ω_j and deleted tracks with χ_j. Further, with the numbers of observation and track-specific events defined as $N_{det} + N_{fal} + N_{new} = M_t$ and $N_{det} + N_{occ} + N_{del} = N_{t-1}$ Eq. 5.12 can be written as

$$\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t) \;= \eta \prod_{i=1}^{M_t} \Big(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i} \;\; \lambda_{fal}^{\phi_i} \;\; \lambda_{new}^{\nu_i} \Big) \\ \prod_{j=1}^{N_{t-1}} \Big(p_{det}^{\delta_j} \;\; p_{occ}^{\omega_j} \;\; p_{del}^{\chi_j} \Big) \;\; p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}), \end{aligned} \tag{5.13}$$

where $\mathcal{N}(\mathbf{z}_i(t)) := \mathcal{N}(\mathbf{z}_i(t);\, \hat{\mathbf{z}}_j(t),\, S_{i,j}(t))$ denotes the measurement likelihood of a single matched observation $\mathbf{z}_i(t)$ associated to an existing track $\mathbf{x}_j(t-1)$ assumed to be a Gaussian pdf centered around the measurement prediction $\hat{\mathbf{z}}_j(t)$ with innovation covariance matrix $S_{i,j}(t)$. Furthermore, the numbers of new tracks N_{new} and false alarms N_{fal} are assumed to follow Poisson distributions with constant expected numbers of events λ_{new} and λ_{fal} in the observation volume V, respectively. The occurrence of N_{det}, N_{occ}, and N_{del} track detection, occlusion, and deletion events is modeled jointly using a multinomial distribution with constant parameters p_{det}, p_{occ}, and p_{del}, respectively.

The integration of the presented place-dependent models is particularly simple in the case of the MHT. The fixed rates λ_{new} and λ_{fal} are substituted by the learned and normalized rate functions $\lambda_{new}(\mathbf{z}_i(t) \mid V_{new})$ and $\lambda_{fal}(\mathbf{z}_i(t) \mid V_{fal})$ from Eq. 5.2 and Eq. 5.5 where $\mathbf{z}_i(t)$ denotes the position of observation i at time t. Further, the track dependent terms p_{occ} and p_{del} are replaced by the place dependent probabilities $p_{occ}(\widehat{\mathbf{x}}_j(t) \mid V_{vis})$ and $p_{del}(\widehat{\mathbf{x}}_j(t) \mid V)$ defined in Eq. 5.9 and Eq. 5.11 with the Kalman-filtered[28] prediction $\widehat{\mathbf{x}}_j(t)$ calculated as

$$\widehat{\mathbf{x}}_j(t) \;=\; F_t \; \mathbf{x}_j(t-1) \;+\; \omega_t, \tag{5.14}$$

where F_t denotes the state transition model applied to the previous state $\mathbf{x}_j(t-1)$ and $\omega_t \sim \mathcal{N}(0, Q_t)$ the process noise drawn from a zero mean Gaussian distribution with covariance matrix Q_t.

The required environment-specific information of the surveillance area V, the regions of increased and decreased new track and false alarm rates V_{new} and V_{fal}, respectively, as well as the currently

[28] Theoretically, any motion prediction method can be used. Experiments showed that even Brownian motion yield comparable results w.r.t the Kalman filtered predictions used in this work.

visible part of V denoted as V_{vis} is combined into

$$\mathbb{V}(t) \;=\; \Big\{V(t),\, V_{new}(t),\, V_{fal}(t),\, V_{vis}(t)\Big\}, \tag{5.15}$$

where t indicates the current time frame as the mentioned volumes and areas might change over time. Additionally, $\mathbb{V}^t$ defines the sequence $\{\mathbb{V}(0), \ldots, \mathbb{V}(t)\}$ of all environmental information from the beginning of the tracking process.

Using the proposed models the hypotheses probability $p(\Omega_l^t \mid \mathcal{Z}^t)$ is now conditioned on the information encoded in $\mathbb{V}$ (Eq. 5.15), thus Eq. 5.13 must be rewritten as

$$\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t, \mathbb{V}^t) \;= \eta \prod_{j=1}^{N_{t-1}} & \Big(p_{det}^{\delta_j} \;\; p_{occ}(\widehat{\mathbf{x}}_j(t) \mid \mathbb{V}(t))^{\omega_j} \;\; p_{del}(\widehat{\mathbf{x}}_j(t) \mid \mathbb{V}(t))^{\chi_j}\Big) \\ & \prod_{i=1}^{M_t} \Big(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i} \;\; \lambda_{fal}(\mathbf{z}_i(t) \mid \mathbb{V}(t))^{\phi_i} \;\; \lambda_{new}(\mathbf{z}_i(t) \mid \mathbb{V}(t))^{\nu_i}\Big) \\ & p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}, \mathbb{V}^{t-1}). \end{aligned} \tag{5.16}$$

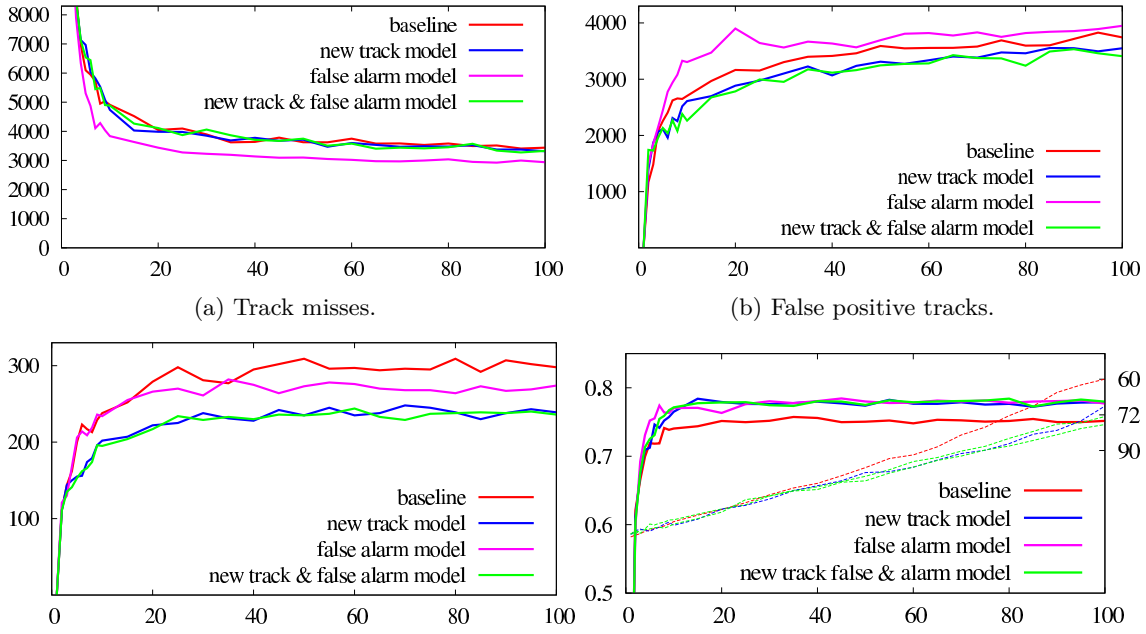
The probabilities $p_{occ}(\widehat{\mathbf{x}}_j(t) \mid \mathbb{V}(t))$ and $p_{del}(\widehat{\mathbf{x}}_j(t) \mid \mathbb{V}(t))$ are calculated with the occlusion and deletion models described above have to be normalized jointly with p_{det} to sum up to one. This is done for each track independently, hence $p_{det} + p_{occ}(\widehat{\mathbf{x}}_j(t) \mid \mathbb{V}(t)) + p_{del}(\widehat{\mathbf{x}}_j(t) \mid \mathbb{V}(t)) = 1, \;\forall\; \mathbf{x}_j(t)$.

5.6 Experiments

The experiments were carried out on two large, unscripted outdoor data set collected in the city center and at the main station of Freiburg during a regular work day. Former consists of 55,475 frames recorded over 25 minutes. The latter has 33,204 frames recorded during 15 minutes. The sensor used for collecting the data is a fixed laser range finder with an angular resolution of 0.5 degree, mounted at a height of $\sim$0.85 meter. The data was collected at fairly busy places that are used by individuals, couples, groups of people, bicycles, cars, people in wheelchairs, subjects on skates and person-shaped static obstacles that all undergo countless occlusions (see also Figure 5.1 and Figure 5.2). By manually annotating 10,000 frames with 162 person tracks of the Freiburg city center data set and 6,000 frames with 160 persons of the Freiburg main station data set to determine the ground truth detections and ground truth data associations were obtained.

A fixed parameter MHT based on the approach of Arras et al. [2008] serves as baseline. The parameters for detections, occlusions, deletions and the fixed rates for false alarms and new tracks have been learned from another training data set with 95 tracks over 28,242 frames. In detail, $p_{det} = 0.7$, $p_{occ} = 0.27$, $p_{del} = 0.03$, $\lambda_{new} = 0.0002$, and $\lambda_{fal} = 0.005$, respectively. As people detector the place-dependent cascade of specialized boosted features classifiers presented in Chapter 2 and based on the approach of Arras et al. [2007] is employed. Shortly, a set of geometrical and statistical features is computed for each group of laser end points and classified by a cascade of spatial informed classifiers. The classifier has also been learned from a separate training set. An experimental evaluation of its accuracy is presented in section 2.5.

The environmental information $\mathbb{V}(t)$ is represented using a two dimensional grid with 10 cm cell resolution. The increased new track and false alarm rates are set to $\lambda_{new}^{high} = 2 \cdot \lambda_{new}$ and $\lambda_{fal}^{high} = 2 \cdot \lambda_{fal}$ while the decreased rates are $\lambda_{new}^{low} = 1/2 \cdot \lambda_{new}$ and $\lambda_{fal}^{low} = 1/2 \cdot \lambda_{fal}$, respectively. The occlusion model uses 200 samples per track. The decay parameter of the deletion model for tracks inside the surveillance area V is $\lambda_{del}^{V} = 30$ and the deletion probability of targets outside the sensor field of view



(a) Track misses.

(b) False positive tracks.

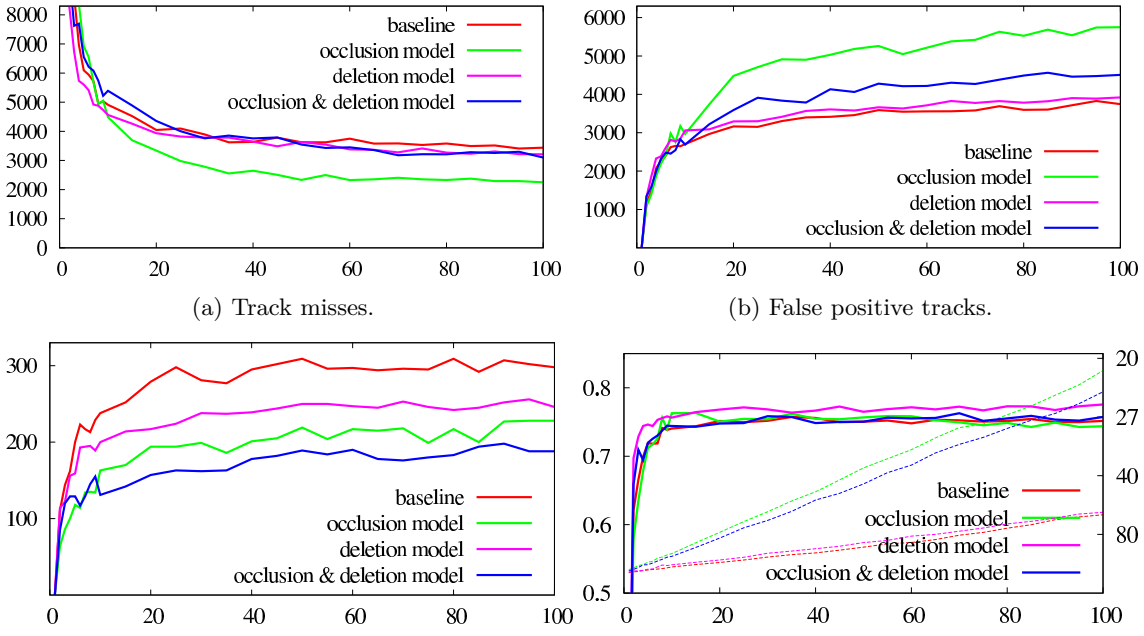
(c) Identifier switches.

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

Figure 5.4: Results of the CLEAR MOT analysis of the Freiburg city center data set inspecting the influence of the proposed place-dependent new track model, false alarm model, and their combination. With 100 hypotheses the tracking accuracy (MOTA) increases from 75.1% to ∼78.0%. The approaches can be applied in real time.

is $p_{del}^{\notin V} = 0.9$. All experiments are conducted with varying number of $N_{hyp} \in [1, \ldots, 100]$ hypotheses to verify the behavior w.r.t. the computational effort.

To compare the impact of the presented models onto the tracking performance first the individual models are tested against the baseline tracker and then the combinations that makes sense. The accuracy of the resulting strategies is then measured using the CLEAR MOT metrics proposed by Bernardin and Stiefelhagen [2008]. The metric counts three numbers with respect to the ground truth information that are incremented at each frame: misses (missing tracks that should exist at a ground truth position, FN), false positives (tracks that should not exist, FP), and mismatches (track identifier switches, ID). Especially the latter value is interesting as it quantifies the ability to deal with occlusion events that typically occur when tracking people. From these numbers, two further values are determined: MOTP (average metric distance between estimated targets and ground truth) and MOTA (average number of times of a correct tracking output with respect to the ground truth). MOTP is ignored as it is based on a precise metric ground truth of target positions which is not available in the data. Note that, for people tracking, the three error types, FN, FP, and ID, are not equally important. The key challenge of a people tracker, according to experience, is to maintain the identity of tracks through occlusions, misdetections, interactions and maneuvers. Delayed track termination of people that leave the field of view or delayed track creation are, compared to this, less relevant aspects. An overview of the comparisons is given in Table 5.1 which contains the CLEAR MOT values and the average cycle time in *Hz* for the baseline tracker, the isolated models and their combination. A more detailed analysis of the tracking behavior in the two investigated environments applying the proposed models in given in the following three subsections.



(a) Track misses.

(b) False positive tracks.

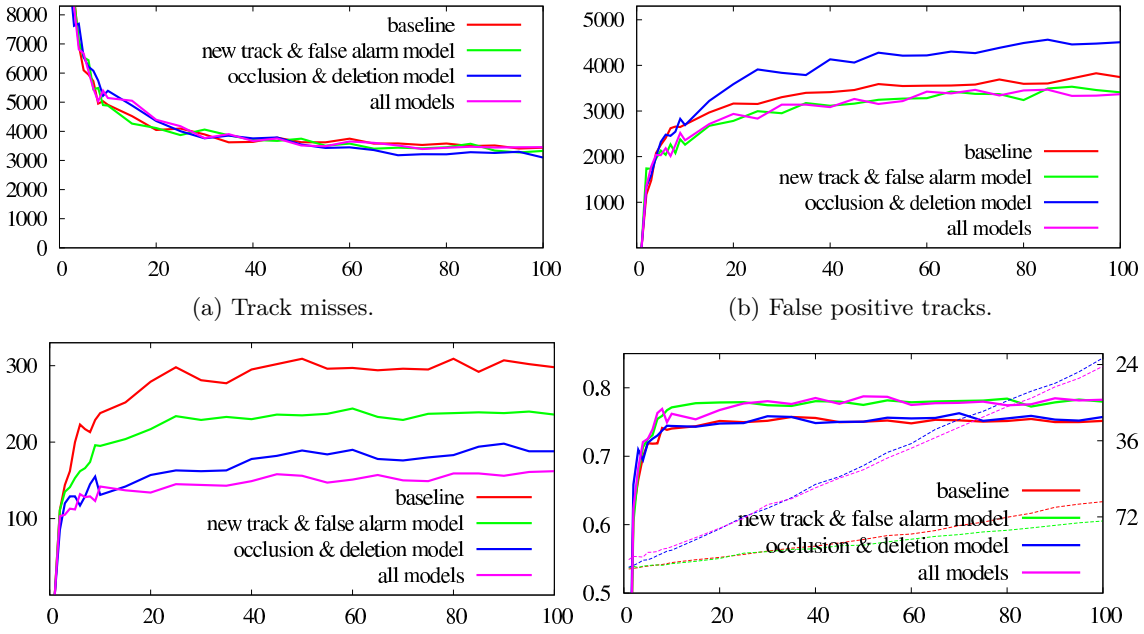
(c) Identifier switches (ID).

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

Figure 5.5: CLEAR MOT results of the Freiburg city center data set analyzing the occlusion, deletion model, and their combination. With 100 hypotheses the occlusion model (shown in green) causes 34.6% fewer track misses at the expense of 53.6% more false positives. However, the number of identifier switches decrease by 23.5%. The approach can be applied in real time but due to sampling the average run-time decreases to 21.1 Hz. The deletion model (purple line) reduces the ID by 17.4% with an average run-time of 58.1 Hz.

5.6.1 New Track and False Alarm Models

In this section the influence of the proposed new track (see subsection 5.3.1) and false alarm (see subsection 5.3.2) models and their combination on the tracking behavior is analyzed in detail. Numerical results of the CLEAR MOT metrics calculated on the Freiburg city center data set are presented in Figure 5.4. Both models reason about the observations that are not assigned to any of the existing tracks. While the former modifies the expected average number of new track events the latter adapts the false alarm rate. These place-dependent rate functions enable the tracker to bring down the values of missed tracks (FN decreases by 3.6% and 14.5%, respectively) as the track creations when targets enter the field of view are supported. They also improve the number of mismatches (-19.8% and -8.1%) as during data association, the system can take better, place-dependent decisions on track creations e.g. after lengthy occlusion events. Using the new track model, the number of false positives is reduced (-5.2%) as observations from cluttered inside the sensor field of view create new tracks less often. The false alarm model implements a form of background learning and is also able to filter systematic misdetections from background clutter. Unfortunately, due to a lower false alarm rate in the free space wrongly detected objects like bicycles and luggage are interpreted as new targets increasing the FP by 5.5%. The combination of both models is able to resolve that issue and to improve the results on all aspects. Especially the number of identifier switches is reduced by 20.8% showing that both models complement each other and contribute to a more accurate tracking behavior.



(a) Track misses.

(b) False positive tracks.

(c) Identifier switches.

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

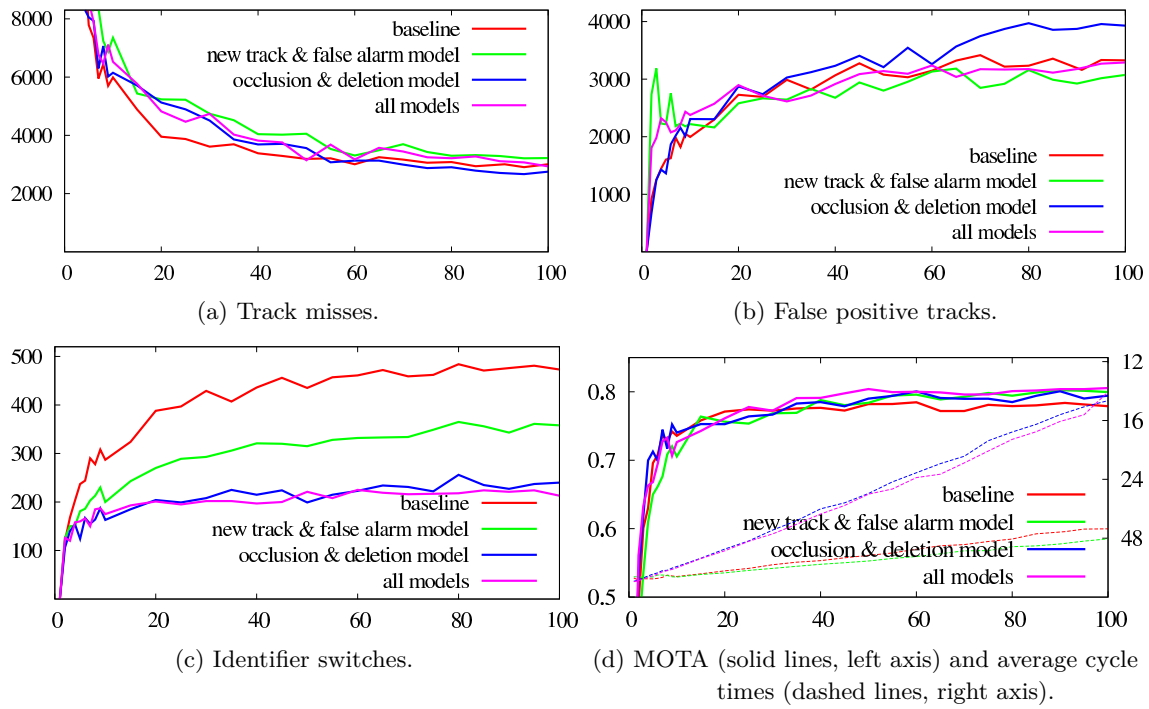
Figure 5.6: CLEAR MOT result of the Freiburg city center data set using the combination of the new track and false alarm model (green line), the combination of the occlusion and deletion model (blue line), and all models (purple line). While the former two improve the tracking results on some aspects only the combination of all models resolves that trade off. With 100 hypotheses the number of identifier switches decreases by 45.6% and the tracking accuracy (MOTA) increases from 75.1% to 78.3%. The presented approach can be applied in real time with an average run-times of 24.2 Hz.

As the new track model requires a manual annotation of the locations where people appear frequently a second experiment investigates its behavior when such information is not available or wrongly modeled. For this purpose the environment model V_{new}^{high} of the Freiburg main station consists of the border of the sensor field of view only. An escalator inside the visible space (shown as blue circle in Figure 5.2) where people appear frequently was intentionally not annotated. If people use the escalator to enter the surveillance area multiple consecutive observations are required to initialize a track, thus the number of FN increases by 5.1%. The new track and false alarm model cause no additional costs as the required environmental information is computed in advance. Much better, the average frame-rate increases as fewer wrong tracks need to be maintained and the data association ambiguity is reduced.

Chapter 6 proposed an approach to learn and encode human-specific spatial priors on new track and false alarm event from observations. A comparison between the modeling and learning approach is presented in section 6.6.

5.6.2 Occlusion and Deletion Models

The occlusion and deletion models explain tracks not assigned to any of the available observations due to missing detections from detector failures or occlusions. The results of the CLEAR MOT metrics analyzing the tracking behavior on the Freiburg city center data set (presented in Figure 5.5)



(a) Track misses.

(b) False positive tracks.

(c) Identifier switches.

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

Figure 5.7: CLEAR MOT analysis of the Freiburg main station data set. Using the combination of all models (shown as purple line) and 100 hypotheses the number of identifier switches decrease by 54.9%. The approach can be applied in real time but due to a sampling strategy used to model occlusion events that occur often in this environment the average run-time drops to 13.8 Hz.

show that the models are able to fill such detection gaps and reduce the number of misses (FN) by 34.6% and 6.7%, respectively, in comparison to the baseline that wrongly terminates many of these tracks. Thereby, they also reduce the number of mismatches (ID) by 23.7% and 17.4% since the baseline incorrectly recreates new tracks if a track was lost during an occlusion. However, this comes at the expense of a higher number of false positives as incorrect detections (e.g. trees) are retained more persistently[29]. Especially, when using the occlusion model the number of FP increases by 53.6%. Furthermore, using this model the tracks of people leaving the sensor field of view are not delete immediately. The combination of both models trades off the ability to bridge gaps and to delete tracks properly, thus an improvement in FN by 9.9% is achieved. On the other hand, the decline in FP is still 20.4%. However, the number of identifier switches (ID) can be decreased by 36.9%. The analysis of the Freiburg main station data set confirms these results.

The improvements in the tracking accuracy come at the expense of a lower frame-rate. In particular, the sampling strategy to estimate the occlusion probability causes large additional costs[30]. Furthermore, if more tracks are maintained the data association ambiguity is increased.

[29] The numbers of misses (FN) and false positives (FP) are counted at every frame, thus an increase by e.g. 12 occurrences corresponds to 1 second of missing or wrongly tracking a target as the employed laser range finder provides the data with 12 Hz.

[30] The MHT generates multiple hypotheses of how the state of the world including all tracks evolves over time. Thereby, also the occlusion probabilities of tracks in free space with a perfectly matching observation need to calculated. This is what makes the occlusion model so expensive.

Data Set	Model(s)	FN	FP	ID	MOTA	*Hz*
Freiburg city center	baseline[31]	3440	3744	298	75.1%	59.9
	new track	3317 (-3.6%)	3550 (-5.2%)	239 (-19.8%)	77.9%	72.6
	false alarm	2940 (-14.5%)	3949 (+5.5%)	274 (-8.1%)	77.7%	68.5
	occlusion	2251 (-34.6%)	5751 (+53.6%)	228 (-23.5%)	74.4%	21.1
	deletion	3211 (-6.7%)	3919 (+4.7%)	246 (-17.4%)	77.6%	58.1
	new + false	3328 (-3.3%)	**3409** (-8.9%)	236 (-20.8%)	77.9%	76.1
	occ + del	**3100** (-9.9%)	4507 (+20.4%)	188 (-36.9%)	75.7%	23.3
	all models	3367 (-2.1%)	3419 (-8.7%)	**162** (-45.6%)	**78.3%**	24.2
Freiburg main station	baseline	3006	3327	473	79.4%	41.4
	new track	3160 (+5.1%)	3189 (-4.1%)	351 (-25.8%)	79.5%	49.6
	false alarm	2343 (-22.1%)	3764 (+13.1%)	397 (-16.1%)	80.4%	45.3
	occlusion	1565 (-47.9%)	5189 (+55.9%)	284 (-39.9%)	78.8%	13.1
	deletion	2864 (-4.7%)	3716 (+11.7%)	339 (-28.3%)	78.5%	34.7
	new + false	3219 (+7.1%)	**3074** (-7.6%)	358 (-24.3%)	79.9%	48.5
	occ + del	**2754** (-8.4%)	3929 (+18.1%)	240 (-49.3%)	79.4%	14.4
	all models	2934 (-2.4%)	3290 (-1.1%)	**213** (-54.9%)	**80.5%**	13.8

Table 5.1: CLEAR MOT results of all data sets using $N_{Hyp} = 100$ hypotheses. Employing place dependent models for the events of new tracks, false alarms, occlusions, and deletions, respectively, the number of identifier switches (ID) can be increased dramatically. The most accurate tracking result is achieved using the combination of all models. This improvement comes at the expense of a lower frame-rate.

5.6.3 Combination of all Models

The isolated models inspected in the previous subsections are only able to make selective improvements, trading off the different performance aspects. In the following, the tracking performance using the combination of all proposed place-dependent models is analyzed. The results are presented in Figure 5.6 and Figure 5.7. An overview of all models is given in Table 5.1.

The combination of all models is able to resolve the trade-offs of the single models and reduces all errors w.r.t. the baseline approach. On the other hand, the numbers of missed tracks (FN) and false positives (FP) are higher compared to the combination of the specialized models. However, the most relevant figure, the number of identifier switches (ID), is reduced by 45.6% on the Freiburg city center and 54.9% on the Freiburg main station data set, respectively. These improvements show that simple, human-specific models for the occurrence of new tracks, false alarms, track occlusions, and track deletions are able to support a probabilistic people tracking framework leading to an increased tracking performance by a factor of two over the baseline.

[31] The baseline approach uses none of the proposed observation and track specific models. Instead, fixed parameters for the Poisson rates of new track and false alarm events and fixed probabilities of detection, deletion, and occlusion events are learned from a training data set.

This encouraging result comes at the expense of additional computational costs. Especially, the occlusion model, that employs sampling, is expensive. The computational effort of all other models is negligible. This is particularly true for the new track and false alarm model that replace fixed Poisson rates by place-dependent functions, simply realized by a lookup into a grid. The cycle time differences in Table 5.1 are due to the behavior differences of the tracking system caused by the models. For instance, more false positive tracks inflate the system and raise the level of data association ambiguity, which in turn, leads to a slower tracker. However, a frame rate above 12.0 Hz is achieved, thus the presented approach runs in real time[32].

5.7 Conclusions

In this chapter informed place-dependent target and detector models for the task of people tracking are presented and compared. The models overcome the rather generic assumptions made in related work and have been shown to significantly improve the tracking performance.

To model place-dependent new track, false alarm, occlusion, and deletion events additional environmental information is required. Approaches to model these information using an occupancy grid maps and a sampling strategy is provided. Furthermore, the integration into the MHT framework that is applied for people tracking in this work is presented. However, the presented approaches can be integrated into any probabilistic target tracking framework regardless the sensor modality, the filtering approach, or the space in which targets are represented.

In the experiments using two large-scale outdoor data sets and a the CLEAR MOT tracking performance metric, the impact of the individual models and their combinations is evaluated systematically. All models reduce the number of identifier switches, which quantifies the ability of the tracking system to deal with occlusion events. It was found that the combined application of all models performs best as it is able to resolve the trade-offs introduced by some of the models applied in isolation. The combination leads to an improvement in terms of track identity confusions – the aspect that is most relevant for people tracking – by a factor of two. Reminding, this has been achieved by integrating a set of rather easy-to-use models leaving the much more complex filtering, data association or target detector machineries unaltered.

[32] To reduce the computational complexity of the applied multiple hypotheses data association the number of generated hypotheses N_{Hyp} can be reduced. This leads to a linear speed up but comes at the expense of a slightly lower tracking accuracy.

Part III

Social and Spatio-Temporal Constraints: Learning-Based Approaches

6 Learning Spatial Affordances

People detection and tracking is important in many scenarios where robots and humans work and live together. But unlike targets in traditional tracking problems, people typically move and act under the constraints of the environment. Probabilities and frequencies at which people appear, disappear, walk or stand are not uniform but vary over space making human behavior strongly place-dependent.

In this chapter a model to learn and encode these spatial priors on human behavior is presented. Furthermore, it is shown how this model can be incorporated into a people tracking system to improve the tracking accuracy. Concretely, a non-homogeneous spatial Poisson process is learned that improves data association in a multi-hypothesis target tracker through more informed probability distributions over hypotheses. Further a place-dependent motion model is presented. Based on the spatial priors motion predictions follow the space usage patterns that people take and that are described by the learned spatial Poisson process.

Large-scale experiments in different indoor and outdoor environments using 2D laser range data, demonstrate how both extensions lead to a more accurate tracking behavior in terms of track losses. Moreover, the number of data association errors decreases by more than 30%. The extended tracker is also slightly more efficient than the baseline approach. The system runs in real-time on a typical desktop computer.

This chapter is structured as follows. A short introduction is given in section 6.1 followed by a review of related work in section 6.2. Section 6.3 introduces the theory of the spatial affordance map and expressions for learning its parameters. Section 6.4 describes how the map is used to improve data association from refined probability distributions over hypotheses, while section 6.5 presents the theory of the place-dependent motion model. In section 6.6 the experimental results are presented and section 6.7 concludes the chapter.

6.1 Introduction

As robots are entering domains in which they interact and cooperate closely with humans, people tracking becomes a key technology for areas such as human-robot interaction, human activity understanding or intelligent cars. In contrast to most air- and waterborne targets, people typically move and act under environmental and social constraints. These constraints vary over time and space making possible motion and action patterns strongly place- and time-dependent. Examples for place-dependent motion include walls that restrict the walkable area of an environment or a cooking stove that constraints the activity of cooking to a specific location.

In this chapter human spatio-temporal behavior is learned for the purpose of improved people tracking. By learning a spatio-temporal model that represents activity events in a global reference frame and on large time scales, the robot acquires place- and time-dependent priors on human activities and behaviors. It will be demonstrated, how such priors can be used to better hypothesize about the state of people in the world, and how place-dependent predictions of human motion that reflect how people are actually using space can be made. Concretely, in this work a non-homogeneous spatial Poisson process is propsed to learn and represent the spatially varying distribution over relevant human activity events. This representation, called *spatial affordance map*, holds space- and

time-dependent Poisson rates for the occurrence of track events such as the creation of new tracks, the confirmation of known targets, or detection failures marked as false alarms. The map is incorporated into a multi-hypothesis tracking (MHT) framework to derive refined probability distribution over hypotheses.

6.2 Related Work

In most related work on laser-based people tracking like Kluge et al. [2001], Fod et al. [2002], Kleinhagenbrock et al. [2002], Schulz et al. [2003], Topp and Christensen [2005], Cui et al. [2005], and Mucientes and Burgard [2006], a person is represented as a single state that encodes torso position and velocities. People are extracted from range data as single blobs or found by merging nearby point clusters that correspond to legs. People tracking has also been addressed as a leg tracking problem where people are represented by the states of two legs. Either a single augmented state is employed as in Cui et al. [2006a] or a high-level person track to which two low-level leg tracks are associated like in Taylor and Kleeman [2004] and Arras et al. [2008].

Different data association approaches have been used to address laser-based people tracking. The nearest neighbor filter and variations thereof are typically employed in earlier works by Kluge et al., Fod et al., and Kleinhagenbrock et al.. A sample-based joint probabilistic data association filter (JPDAF) has been presented by Schulz et al. and adopted by Topp and Christensen [2005]. And in Khan et al. [2006] a Markov chain Monte Carlo (MCMC)-based auxiliary variable particle filter is proposed. This work employs multi-hypothesis tracking (MHT) that has already been used by Taylor and Kleeman, Mucientes and Burgard and Arras et al. and is an attractive choice as it belongs to the most general data association techniques. The method generates joint compatible assignments, integrates them over time, and is able to deal with track creation, confirmation, occlusion, and deletion events in a probabilistically consistent way. Other multi-target data association techniques such as the global nearest neighbor filter, the track splitting filter, the JPDAF, or the probabilistic multi-hypothesis tracking (PMHT) by Streit and Luginbuhl [1995] are suboptimal in nature as they simplify the problem in one or the other way as explained in Bar-Shalom and Li [1995] and Blackman [2004]. For these reasons, the MHT has become a widely accepted tool in the target tracking community as pointed out by Blackman, especially for problems with low to medium number of targets.

For people tracking, however, the original MHT approach according to Reid [1979] and Cox and Hingorani [1996] relies on statistical assumptions that are overly simplified and do not account for place-dependent target behavior. In detail, the approach assumes new tracks and false alarms being uniformly distributed in the sensor field of view with fixed Poisson rates. While this might be acceptable in settings for which the approach has been originally developed (using, e.g., radar or underwater sonar), it does not account for the non-random usage of an environment by people. Human subjects appear, disappear, walk or stand at specific locations that correspond, for instance, to doors, elevators, entrances, or convex corners. False alarms are also more likely to arise in areas with cluttered backgrounds rather than in open spaces.

A simple form of place-dependency has been realized in Breitenstein et al. [2009], a visual surveillance scenario with a static camera, where a frame around the border of the image has been manually annotated to describe the area where new tracks are allowed to appear. In the center of the image, no new tracks are assumed to arrive.

In Chapter 5 a similar approach is taken to derive the regions of increased new track probability by inspecting the border of the sensor field of view. In contrast to Breitenstein et al. new tracks in the center of the monitored area are modeled with lower likelihood and are not forbidden completely.

Furthermore, regions of increased false alarm rate are predicted from a map of the environment modeling that detection failures are more likely in area close to static obstacles that cause clutter.

In this chapter, prior work is extended by incorporating learned distributions over track interpretation events in order to support data association and show how a non-homogeneous spatio-temporal Poisson process can be used to seamlessly extend the MHT approach for this purpose.

For motion prediction of people, most researchers employ the Brownian motion model and the constant velocity motion model. The former makes no assumptions about the target dynamics, the latter assumes linear target motion. Better motion have been proposed by the following authors.

Bruce and Gordon [2004] learn goal locations in the environment from people trajectories obtained by a laser-based tracker. Goals are found as end points of clustered trajectories. Human motion is then predicted along paths that a planner generates from the location of people being tracked to the goal locations. The performance of the tracker was improved in comparison to a Brownian motion model. Liao et al. [2003] extract a Voronoi graph from a map of the environment and represent the state of people being on edges of that graph. This allows them to predict motion of people along the edges that follow the topological shape of the environment. The approach of Vasquez et al. [2009] learns motion patterns and goals incrementally using Growing HMMs (GHMMs). Learning is performed on-line on complete sequences of observations assuming that the last observations corresponds to the persons goal. Maximum entropy Inverse Reinforcement Learning (IRL) is employed by Ziebart et al. [2009] to model the goal-directed trajectories of people. Future trajectories of people are predicted by computing conditional probabilities of any path continuing their motion.

With maneuvering targets, a single model can be insufficient to represent the target's motion. Multiple model based approaches in which different models run in parallel and describe different aspects of the target behavior are a widely accepted technique to deal with maneuvering targets, in particular the Interacting Multiple Model (IMM) algorithm (for a survey, see Mazor et al. [1998]). Different target motion models are also studied by Kwok and Fox [2005]. The approach is based on a Rao-Blackwellized particle filter to model the potential interactions between a target and its environment. The authors define a discrete set of different target motion models from which the filter draws samples. Then, conditioned on the model, the target is tracked using Kalman filters.

In Chapter 4 the social force model, a computational model developed in the cognitive and social science communities, is employed to describe individual and collective pedestrian dynamics. Inner motivation, physical constraints, and social rules are integrated into a sound and common mathematical framework. The motion model presented in Chapter 7 applies on-line learning of geometric relations between people walking in groups. Extending the approach of Mallick et al. [2011] to multiple interaction partners and motion prediction in 2D the model accounts for the fact that people in groups try to maintain their spatial organization.

The motion model presented in this chapter extends prior work by integration learned spatio-temporal information. Opposed to Liao et al., Kwok and Fox and IMM related methods, it does not rely on predefined motion models but applies learning for this task of acquiring place-dependent motion models. In Liao et al. and Vasquez et al., the positions of people are projected onto graphs which are topologically correct but metrically poor models for human motion. While sufficient for the purpose of their work, there is no insight why people should move on a limited set of nodes, particularly in open spaces whose topology is not well defined. The presented approach, by contrast, tracks the actual position of people and predicts their motion according to metric, place-dependent models. Opposed to Bruce and Gordon where motion prediction is done along paths that a planner plans to a set of goal locations, the presented learning approach predicts motion along the trajectories that people are actually following.

6.3 Spatial Affordance Map

The problem of learning a spatio-temporal model of human behavior is posed as a parameter estimation problem of a non-homogeneous spatio-temporal (or space-time) Poisson process. The resulting model, called *spatial affordance map*, is a global long-term representation of human activity events in the environment. The name lends itself to the concept of affordances as it considers the possible sets of human actions and motions as a result from environmental constraints. An affordance is a resource or support that an object (the environment) offers an agent (a human) for action. This section describes the theory and how learning of affordances in the spatial affordance map is implemented.

A Poisson distribution is a discrete distribution to compute the probability of a certain number of events n given an expected average number of events λ over time or space, and is defined as

$$P_\lambda(n) \;=\; \frac{\lambda^n}{n!}\, e^{-\lambda}. \tag{6.1}$$

The parameter of the distribution is the positive real number λ defining the rate at which the events occur per time or volume units. To model events that occur randomly in time or space, the Poisson distribution is a natural choice.

Based on the assumption that events in time occur independently of one another, a *Poisson process* can deal with distributions of time intervals between events. Concretely, let $N(t)$ be a discrete random variable to represent the number of events occurring up to time t with rate λ, $N(t)$ follows a Poisson distribution with parameter λt, thus

$$P(N(t) = n) \;=\; \frac{(\lambda t)^n}{n!}\, e^{-\lambda t}, \qquad \text{with } n = 0, 1, \ldots \tag{6.2}$$

In the general case of a non-homogeneous Poisson process, the rate parameter λ may change over time, thus the generalized rate function is given as $\lambda(t)$[33] and the expected number of events in the time between t_s and t_e is given as

$$\lambda_{t_s,t_e} \;=\; \int_{t_s}^{t_e} \lambda(t)\, dt. \tag{6.3}$$

The discrete random variable $N(t) = (N(t_e) - N(t_s))$ representing the number of events occurring in the time interval $(t_s, t_e]$ with rate function λ_{t_s,t_e} follows a Poisson distribution with parameter λ_{t_s,t_e}, hence

$$P((N(t_e) - N(t_s)) = n) \;=\; \frac{(\lambda_{t_s,t_e})^n}{n!}\, e^{-\lambda_{t_s,t_e}} \qquad n = 0, 1, \ldots \tag{6.4}$$

The *spatio-temporal* Poisson process introduces a spatial dependency on the rate function given as $\lambda(\vec{x}, t)$ with $\vec{x} \in X$ where X is a vector space such as $\mathbb{R}^2$ or $\mathbb{R}^3$. For any subset $S \subset X$ of finite extent (e.g. a spatial region), the number of events occurring inside this region can be modeled as a Poisson process with associated rate function $\lambda_S(t)$, such that

$$\lambda_S(t) \;=\; \int_{\vec{x} \in S} \lambda(\vec{x}, t)\, d\vec{x}. \tag{6.5}$$

In the case that events occur independently in space and time the generalized rate function $\lambda(\vec{x}, t)$ is a separable function (of time and space) and can be expressed as a product of two functions, hence

$$\lambda(\vec{x}, t) \;=\; f(\vec{x})\, \lambda(t), \tag{6.6}$$

[33] A homogeneous Poisson process is a special case of a non-homogeneous process with constant rate $\lambda(t) = \lambda$.

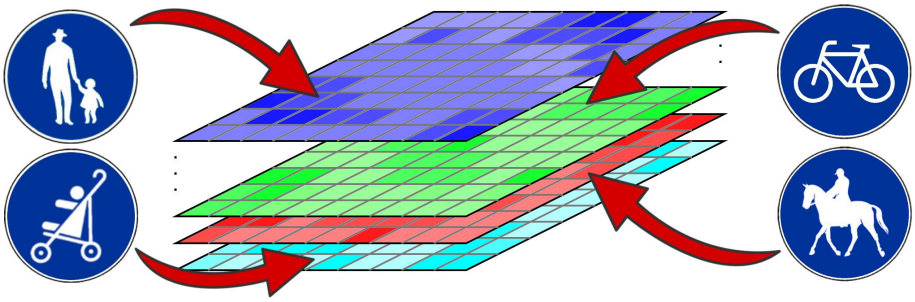


Figure 6.1: Visualization of the multi-layered spatial affordance map encoding the probability distributions and rate functions of various human activity events. Oppose to the illustrated activities in this work the occurrences of typical tracking events like new tracks, track continuations, and false alarms are learned.

for some spatial-dependent function $f(\vec{x})$ which can be constraint to

$$\int_X f(\vec{x})\, d\vec{x} \;=\; 1, \tag{6.7}$$

without loss of generality. This particular decomposition allows to decouple the occurrence of events between time and space. Given Eq. 6.7, $\lambda(t)$ defines the occurrence rate of events in time, while $f(\vec{x})$ can be interpreted as a probability distribution on where the event occurs in space.

Learning the spatio-temporal distribution of events in an environment is equivalent to learn the generalized rate function $\lambda(\vec{x}, t)$. However, learning the full continuous function is a highly expensive process. For this reason, an approximation of the non-homogeneous spatio-temporal Poisson process using a piecewise homogeneous spatial Poisson process is taken. The approximation is performed by discretizing the environment into a bi-dimensional grid, where each cell represents a local homogeneous Poisson process with a fixed rate over time and space,

$$P_{ij}(n) \;=\; \frac{(\lambda_{ij})^n}{n!}\, e^{-\lambda_{ij}} \qquad k = 0, 1, \ldots \tag{6.8}$$

where λ_{ij} is assumed to be constant over time. Finally, the spatial affordance map is the generalized rate function $\lambda(\vec{x}, t)$ using a grid approximation,

$$\lambda(\vec{x}, t) \;\simeq\; \sum_{(i,j)\in X} \lambda_{ij}\, \mathbf{1}_{ij}(\vec{x}) \tag{6.9}$$

with $\mathbf{1}_{ij}(\vec{x})$ being the indicator function defined as $\mathbf{1}_{ij}(\vec{x}) = 1$ if $\vec{x} \in \text{cell}_{ij}$ and $\mathbf{1}_{ij}(\vec{x}) = 0$ if $\vec{x} \notin \text{cell}_{ij}$. The type of approximation is not imperative and goes without loss of generality. Other space tessellation techniques such as graphs, quadtrees or arbitrary regions of homogeneous Poisson rates can equally be used. Subdivision of space into regions of fixed Poisson rates has the interesting properties that the preferable decomposition in Eq. 6.6 holds and that properties of the environment can be inferred instantly without computing expensive integrals.

Each type of human activity event can be used to learn its own probability distribution and rate functions in the map. Therefore, the map as a representation with multiple layers, one for every type of event. For the purpose of this work, the map has three layers, one for new tracks, for matched tracks, and for false alarms. The first layer represents the distribution and rates of people appearing in the environment. The second layer can be considered a space usage probability and contains a walkable area map of the environment. The false alarm layer represents the place-dependent reliability of the detector.

6.3.1 Learning

This section shows how learning of the parameter of a single cell in the grid from a sequence $K_t = \{k_1, \ldots, k_t\}$ of t activity observations with $k_i \in \{0, 1\}$ is implemented. Bayesian inference for parameter learning is applied, since the Bayesian approach can provide information on cells via a prior distribution. The parameter λ is modeled using a Gamma distribution, as it is the conjugate prior of the Poisson distribution. Let λ be distributed according to the Gamma density, $\lambda \sim \text{Gamma}(\alpha, \beta)$, parametrized by the shape parameter α and the inverse scale parameter β called rate parameter, yields

$$\text{Gamma}(\lambda; \alpha, \beta) \;=\; \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\,\lambda} \qquad \text{for } \lambda > 0. \tag{6.10}$$

Then, learning the rate parameter λ consists in estimating the parameters α and β of the Gamma distribution. At discrete time index t, the posterior probability of λ_t according to Bayes' rule is computed as

$$P(\lambda_t \mid K_t) \;\sim\; P(k_t \mid \lambda_{t-1})\; P(\lambda_{t-1}) \tag{6.11}$$

with $P(\lambda_{t-1}) = \text{Gamma}(\alpha_{t-1}, \beta_{t-1})$ being the prior and $P(k_t \mid \lambda_{t-1}) = P(k_t)$ the likelihood from Eq. 6.8. Then by substitution, it can be shown that the update rules for the parameters are

$$\alpha_t = \alpha_{t-1} + k_t \quad \text{and} \quad \beta_t = \beta_{t-1} + 1. \tag{6.12}$$

The posterior mean of the rate parameter in a single cell is finally obtained as the expected value of the Gamma distribution,

$$\widehat{\lambda}_{\text{Bayesian}} \;=\; \mathbb{E}[\lambda] \;=\; \frac{\alpha}{\beta} \;=\; \frac{\#\text{positive events} + 1}{\#\text{observations} + 1}\,. \tag{6.13}$$

For $t = 0$ the quasi uniform Gamma prior for $\alpha = 1$, $\beta = 1$ is taken. The advantages of the Bayesian estimator are that it provides a variance estimate ($\mathbb{V}\text{ar}[\lambda] = {}^{\alpha}/_{\beta^2}$) which is a measure of confidence of the mean and that it allows to properly initialize never observed cells.

Given the learned rates the space distribution of the various events can be estimated. This distribution is obtained from the rate function of the spatial affordance map $\lambda(\vec{x}, t)$. While this estimation is hard in the general setting of a non-homogeneous spatial Poisson process, it becomes easy to compute if the separability property of Eq. 6.6 holds[34]. In this case, the pdf, $f(\vec{x})$, is obtained by

$$f(\vec{x}) \;=\; \frac{\lambda(\vec{x}, t)}{\lambda(t)} \tag{6.14}$$

where $\lambda(\vec{x}, t)$ is encoded in the spatial affordance map. The nominator, $\lambda(t)$, can be obtained from the map by substituting the expression for $f(\vec{x})$ into the constraint defined in Eq. 6.7. Hence,

$$\lambda(t) = \int_X \lambda(\vec{x}, t)\, d\vec{x}. \tag{6.15}$$

In the employed grid discretization, those quantities are computed as

$$f(\vec{x}) \;=\; \frac{\sum_{(i,j)\in X} \lambda_{ij}\, \mathbf{1}_{ij}(\vec{x})}{\sum_{(i,j)\in X} \lambda_{ij}}. \tag{6.16}$$

[34] Note that for a non-separable rate function, the Poisson process can model places whose importance changes over time.

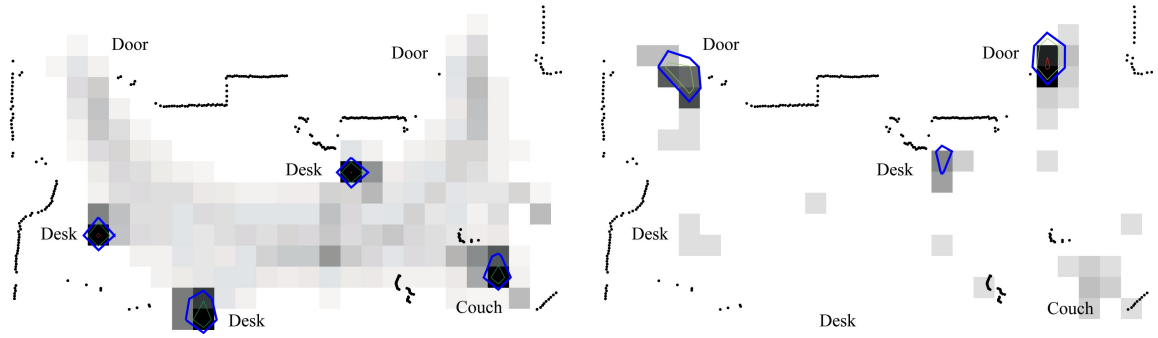


Figure 6.2: Spatial affordance map of a laboratory in experiment 1. The probability distribution of matched track events is shown on the left, the distribution of new track events is shown on the right. The marked locations in each distribution (extracted with a peak finder and visualized by contours of equal probability) have different meanings. While on the left they correspond to places that are often visited by people (three desks and a couch), the maxima of the new track distribution (right) denote locations where people appear in the sensor field of view (two doors, the couch and a desk).

Interesting places in the environment then correspond to the modes of the distribution $f(\vec{x})$ which can be extracted by any peak finding method. This method, of course, is simple in comparison to Bruce and Gordon [2004] where goals are found as end points of clustered people trajectories using a nonlinear optimization technique. Similarly, in Bennewitz et al. [2005], goals are found from sets of clustered trajectories using EM. The difference is that in both works, entire trajectories are considered that encode coherent motion information of people.

In case of several layers in the map, each layer contains the distribution $f(\vec{x})$ of the respective type of events. Note that learning in the spatial affordance map is simply realized by counting in a grid. This makes life-long learning particularly straightforward as new information can be added at any time by one or multiple robots.

Figure 6.2 shows two layers of the spatial affordance map of a laboratory, learned during the first experiment described in section 6.6. The picture on the left shows the space usage distribution of the environment. The modes in this distribution correspond to often used places and have the meaning of goal locations in that room (three desks and a couch). On the right, the distribution of new tracks is depicted whose peaks designate locations where people appear (doors). The reason for the small peaks at other locations than the doors is when subjects interact with objects (sit on a chair, lie on the couch), the tracker looses them. When they reenter space, they get detected as new tracks.

6.4 Data Association With Spatial Target Priors

Many tracking approaches rely on rather simple models for new track and false alarm events and ignore important information that is either directly available from environment and sensor-specific information (see Chapter 5) of can be learned from human observations as it is proposed in this chapter. For example, the Multi-Hypothesis Tracking (MHT) approach assumes a constant rate Poisson distribution for the occurrence of new tracks and false alarms over time and a uniform probability of these events over space within the sensor field of view V. While this is a valid assumption for a radar aimed upwards into the sky, it does not account for the place-dependent character of human behavior. The way how people move is often given by environmental constraints

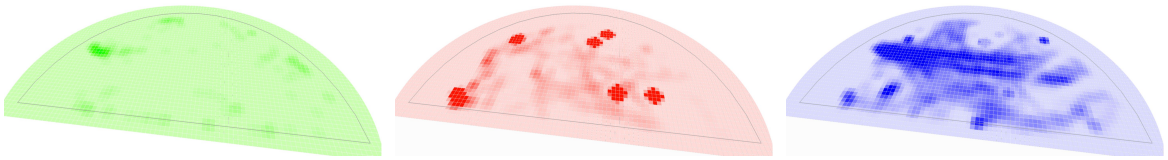


Figure 6.3: Learned spatial priors in the inner city of Freiburg. The probability distributions of new track, false alarms, and track continuation events are shown (from left to right). Local maxima of the new track distribution denote locations where people appear in the sensor field of view. High probability regions in the false alarm distribution denote clutter and areas in which detector failures are more likely.

that can be learned. Indoors, for instance, doors or convex corners are typical places where people appear. The same place-dependency applies for the behavior of a detector. Regions of clutter and complex background produce false alarms more likely than in open space, making a spatially uniform model a poor approximation.

The spatial affordance map exactly holds this kind of information as shown in Figure 6.3. Thus this work extends the original Multi-Hypothesis Tracker (MHT) by Reid [1979] with spatial priors and shows that the map allows for a seamless integration into the MHT framework. In particular, the temporal fixed-rate models for new tracks and false alarms are replaced by the learned Poisson rates for arrival events of people and false detections and the spatial uniform probability with the learned location statistics.

A detailed introduction into the MHT with spacial priors is presented in section 3.6 and briefly summarized in the following. The algorithm hypothesizes about the state of the world by considering all statistically feasible assignments between observations and tracks and all possible interpretations of observations as *false alarms* or *new tracks* and tracks as *matched*, *occluded* or *deleted*. Thereby, the MHT handles the entire life-cycle of tracks from creation and confirmation to occlusion and deletion. A hypothesis Ω_l^t is one possible set of assignments and interpretations at time t.

Formally, let Ω_l^t be the l^{th} hypothesis at time t and $\Omega_{p(l)}^{t-1}$ the parent hypothesis from which Ω_l^t was derived. Let $\mathcal{Z}(t) = \{\mathbf{z}_i(t)\}_{i=1}^{M_t}$ be the set of M_t observations which is in this case the set of detected people. Let further $\psi_l(t)$ denote a set of assignments which associates predicted tracks to observations in $\mathcal{Z}(t)$ and let $\mathcal{Z}^t = \{\mathcal{Z}(0), \ldots, \mathcal{Z}(t)\}$ be the set of all observations up to time t. Starting from a hypothesis of the previous time step $\Omega_{p(l)}^{t-1}$, and a new set of observations $\mathcal{Z}(t)$, there are many possible assignment sets $\psi_l(t)$, each giving birth to a child hypothesis that branches off the parent. This makes up an exponentially growing hypothesis tree as illustrated in Figure 1. For a real-time implementation, the growing tree needs to be pruned. To guide the pruning, each hypothesis receives a probability, recursively calculated as the product of a normalizer η, a measurement likelihood, an assignment set probability and the parent hypothesis probability,

$$p(\Omega_l^t \mid \mathcal{Z}^t) \;=\; \eta \;\; p(\mathcal{Z}(t) \mid \psi_l(t)) \;\; p(\psi_l(t) \mid \Omega_{p(l)}^{t-1}, \mathcal{Z}^{t-1}) \;\; p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \tag{6.17}$$

While the last term is known from the previous iteration, the two expressions that will be affected by the proposed extension are the measurement likelihood and the assignment set probability. More details on the measurement likelihood and the assignment set probability can be found in subsection 3.3.1 and subsection 3.3.2, respectively.

However, for the measurement likelihood, it is assumed that an observation $\mathbf{z}_i(t)$ associated to a track $\mathbf{x}_j(t-1)$ has a Gaussian pdf centered on the measurement prediction $\hat{\mathbf{z}}_j(t)$ with innovation covariance matrix $S_{ij}(t)$, $\mathcal{N}(\mathbf{z}_i(t)) := \mathcal{N}(\mathbf{z}_i(t); \hat{\mathbf{z}}_j(t), S_{i,j}(t))$. The regular MHT now assumes that

the pdf of an observation belonging to a new track or false alarm is uniform in V, the sensor field of view, with probability V^{-1}, thus

$$p(\mathcal{Z}(t) \mid \psi_l(t),\ \Omega_{p(l)}^{t-1},\ \mathcal{Z}^{t-1}) \ = \ V^{-N_{new}}\ V^{-N_{fal}}\ \prod_{i=1}^{M_t} \mathcal{N}(\mathbf{z}_i(t))^{\tau_i}, \tag{6.18}$$

with N_{new} and N_{fal} being the number of false alarms and new tracks, respectively, and τ_i is an indicator variable being 1 if observation i has been associated to a track, and 0 otherwise. Given the spatial affordance map, the term $V^{-N_{new}}$ changes as the probability of new tracks is now inferred from the map by

$$p_{new}(\mathbf{z}_i(t)) \ = \ \frac{\lambda_{new}(\mathbf{z}_i(t), t)}{\lambda_{new}(t)} \ = \ \frac{\lambda_{new}(\mathbf{z}_i(t), t)}{\int_V \lambda_{new}(\vec{x}, t)\, d\vec{x}} \tag{6.19}$$

where $\lambda_{new}(\mathbf{z}_i(t), t)$ is the learned Poisson rate of new tracks at positions $\mathbf{z}_i(t)$ transformed into global coordinates. Given the grid approximation, Eq. 6.19 becomes

$$p_{new}(\mathbf{z}_i(t)) \ = \ \frac{\lambda_{new}(\mathbf{z}_i(t), t)}{\sum_{(i,j)\in V} \lambda_{ij,new}}. \tag{6.20}$$

The probability of false alarms $p_{fal}(\mathbf{z}_i(t))$ is calculated in the same way using the learned Poisson rate of false alarms $\lambda_{fal}(\mathbf{z}_i(t), t)$ in the map. Although the theory presented so far is general, in this work, the appearing behavior of people and the false positive statistics of the detector are assumed to be time-invariant, and therefore, the Poisson processes to be only non-homogeneous over space. The rate partameters $\lambda_{new}(\vec{x}, t)$ and $\lambda_{fal}(\vec{x}, t)$ could be simplified to $\lambda_{new}(\vec{x})$ and $\lambda_{fal}(\vec{x})$, respectively.

The expression of the assignment set probability in the MHT can be shown (see Eq. 3.27) to be

$$p(\psi_l(t) \mid \Omega_{p(l)}^{t-1},\ \mathcal{Z}^{t-1}) \ = \ \eta\ p_{det}^{N_{det}}\ p_{occ}^{N_{occ}}\ p_{del}^{N_{del}}\ \lambda_{fal}^{N_{fal}}\ V^{N_{fal}}\ \lambda_{new}^{N_{new}}\ V^{N_{new}}, \tag{6.21}$$

where N_{det}, N_{occ}, and N_{del} are the number of matched, occluded and deleted tracks, respectively. The parameters p_{det}, p_{occ}, and p_{del} denote the probability of detection (matching), occlusion, and deletion that are subject to $p_{det} + p_{occ} + p_{del} = 1$. The regular MHT now assumes that the number of new tracks N_{new} and false alarms N_{fal} both follow a fixed rate Poisson distribution with expected number of occurrences $\lambda_{new}V$ and $\lambda_{fal}V$ in the observation volume V. Given the spatial affordance map, they can be replaced by rates from the learned spatial Poisson processes with rate functions $\lambda_{new}(t)$ and $\lambda_{fal}(t)$, respectively.

Substituting the modified terms back into Eq. 6.17 many terms cancel out – like in the original approach – leading to an easy-to-implement expression for a hypothesis probability

$$\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t) \ = \eta\ \prod_{i=1}^{M_t} \Big(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i}\ \lambda_{new}(\mathbf{z}_i(t), t)^{\nu_i}\ \lambda_{fal}(\mathbf{z}_i(t), t)^{\phi_i}\Big) \\ p_{det}^{N_{det}}\ p_{occ}^{N_{occ}}\ p_{del}^{N_{del}}\ p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \end{aligned} \tag{6.22}$$

with ν_i and ϕ_i indicating whether an observation is marked as new track or declared to be a false alarm, respectively.

The insight into this extension of the MHT is that fixed parameters are replaced by spatio-temporal priors on human behavior in the form of learned spatial rate functions. As the experiments will show, this domain knowledge leads to refined probability distributions over hypotheses and helps the tracker to better interpret observations and tracks. This extension comes at no additional runtime costs.

6.5 Place-Dependent Motion Model

People are highly dynamic targets to track. They can abruptly stop, turn back, left or right, make a sideway step or accelerate. However, human motion is not random but follows place-dependent patterns typically formed by the environment: people turn around convex corners, maneuver around obstacles, stop in front of doors and do not go through walls. The Brownian model, the constant velocity and even higher-order motion models are clearly unable to capture the complexity of these movements. In addition to this, people often undergo lengthy occlusion events during interaction with each other or with the environment. In this section a place-dependent motion model for short-term predictions of maneuvering targets is proposed. It relies on learned human motion priors in order to account for this complexity.

Formally, this means that the motion model $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{m})$ becomes conditioned on both, the previous track state $\mathbf{x}_{t-1}$ and the walkable area map $\mathbf{m}$ obtained by clipping the space usage probability defined in Eq. 6.16 and shown in Figure 6.3 (right) at a given probability. It describes a general density that follows the shape and topology of the environment, poorly described by a parametric distribution such as a Gaussian. Therefore a sampling approach is taken that represents the target distribution with a set of weighted samples

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{m}) \;\simeq\; \sum_i w_t^{(i)} \delta_{\mathbf{x}_t^{(i)}}(\mathbf{x}_t) \tag{6.23}$$

where $\delta_{\mathbf{x}_t^{(i)}}(\mathbf{x}_t)$ is the impulse function centered in $\mathbf{x}_t^{(i)}$.

Sampling directly from the distribution $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{m})$ is intractable in practice which is why a Monte Carlo approach is employed, in which samples are first drawn from a proposal distribution π and then evaluated according to the mismatch between the target distribution τ and the proposal distribution. In this case, the distribution is approximated by the following factorization

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{m}) \;\simeq\; p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \cdot p(\mathbf{x}_t \mid \mathbf{m}). \tag{6.24}$$

For the proposal distribution π a motion model $p(\mathbf{x}_t^{(i)} \mid \mathbf{x}_{t-1})$ is a natural choice, thus the samples are evaluated according to a weight

$$w_t^{(i)} \;=\; \frac{p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{m})}{p(\mathbf{x}_t^{(i)} \mid \mathbf{x}_{t-1})} \;=\; p(\mathbf{x}_t^{(i)} \mid \mathbf{m}). \tag{6.25}$$

In other words, samples are first spread out into the state space following the motion model $p(\mathbf{x}_t^{(i)} | \mathbf{x}_{t-1})$ and then weighted according to the map $\mathbf{m}$.

For $p(\mathbf{x}_t^{(i)} \mid \mathbf{x}_{t-1})$, the curvilinear model by Best and Norton [1997] is taken. This motion model is simple yet one of the most sophisticated target maneuver models in 2D as pointed out by Rong Li and Jilkov [2003]. It accounts for both, (cross-track) normal and (along-track) tangential target accelerations. As illustrated Figure 6.4, constant velocity and constant turn motion follow as special cases. Let $\mathbf{x}_t^{(i)} = (\; x_t \;\; y_t \;\; \dot{x}_t \;\; \dot{y}_t \;)^T$ be the state of particle i at discrete time t, $\vec{a}_t = (\; a_{tan} \;\; a_{nor} \;)^T$ the vector of tangential and normal accelerations, and A_t the transition matrix of the constant velocity model, then the particle states evolve according to

$$\mathbf{x}_t^{(i)} \;=\; A_t\, \mathbf{x}_{t-1}^{(i)} + G_t(\vec{a}_t^{(i)} + q_t) \tag{6.26}$$

with q_t being zero-mean Gaussian noise with covariance matrix Q_t accounting for unpredictable state changes and variation in $\vec{a}_t$ about their nominal values. The matrices A_t and Q_t are introduced in Eq. 4.12. Details on the 4×2 forcing matrix G_t can be found in Best and Norton.

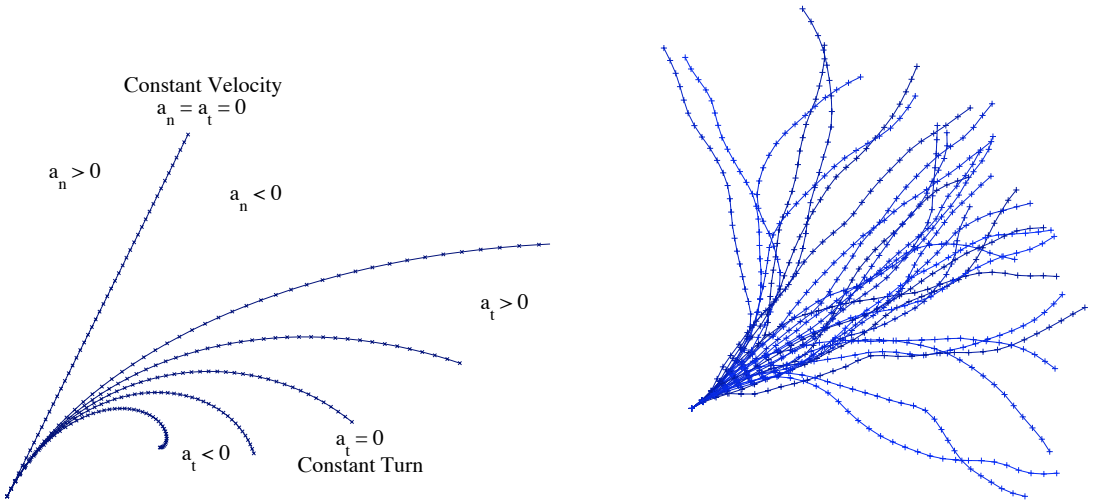


Figure 6.4: Curvilinear motion model according to Best and Norton [1997]. *Left:* The model accounts for both, (cross-track) normal acceleration a_n and (along-track) tangential acceleration a_t. *Right:* Example particles over 30 steps at $\Delta t = 0.1\ s$ subject to white zero-mean Gaussian noise with $\sigma_{a_t} = 0.1\ m/s^2$ and $\sigma_{a_n} = 1\ m/s^2$.

At each discrete time t and for each track, samples are drawn from the posterior state estimate $(\mathbf{x}_t,\ \Sigma_t)$ and sent into different directions by randomizing the accelerations $\vec{a}_t$ by a noise with covariance $Q_t = \mathrm{diag}[\sigma^2_{a_{tan}},\ \sigma^2_{a_{nor}}]$ (see Figure 6.4, right). When an occlusion event occurs, the particles will evolve through Eq. 6.26 and get weighted and resampled according to the strategy described hereafter.

Even a sophisticated motion model can strongly differ from the target distribution, especially at places where the walkable area is highly constrained by the environment. This makes that many samples fall into low probability regions leading to the known problem of particle depletion. For this reason, the auxiliary particle filter approach by Pitt and Shephard [1999] that has been developed for such mismatch situations is employed. In a nutshell, the auxiliary particle filter computes an improved proposal derived from an approximated observation likelihood. Here, this feature can be extended to a look-ahead ability into the future since the map $\mathbf{m}$ delivers the observation likelihood that can be probed at locations computed by forward-simulating the motion model.

Assuming a set of samples $\mathbf{x}^{(i)}_{t-1}$ representing the target distribution at discrete time $t-1$. The distribution at time t is then

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{m}) \ \simeq\ p(\mathbf{x}_t \mid \mathbf{m}) \ \sum_i \Big(p(\mathbf{x}_t \mid \mathbf{x}^{(i)}_{t-1})\ w^{(i)}_{t-1} \Big). \tag{6.27}$$

To avoid depletion, and following Pitt and Shephard, additional samples are drawn from the higher dimensional joint distribution $p(\mathbf{x}_t, k \mid \mathbf{x}_{t-1}, \mathbf{m})$, where the auxiliary variable k[35] denotes the index of the sample at time $t-1$ in the mixture defined above, thus

$$p(\mathbf{x}_t, k \mid \mathbf{x}_{t-1}, \mathbf{m}) \ \simeq\ p(\mathbf{x}_t \mid \mathbf{m})\ p(\mathbf{x}_t \mid \mathbf{x}^{(k)}_{t-1})\ w^{(i)}_{t-1}. \tag{6.28}$$

Drawing from this joint density and ignoring the sampled index, a sample from the original target density is obtained. Replacing $p(\mathbf{x}_t \mid \mathbf{m})$ Eq. 6.28 can be approximated by

$$g(\mathbf{x}_t, k \mid \mathbf{x}_{t-1}, \mathbf{m}) \ \simeq\ p(\vec{\mu}^{(k)}_t \mid \mathbf{m})\ p(\mathbf{x}_t \mid \mathbf{x}^{(k)}_{t-1})\ w^{(i)}_{t-1}, \tag{6.29}$$

[35] The auxiliary variable k is present simply to aid the task of forward simulating the motion.

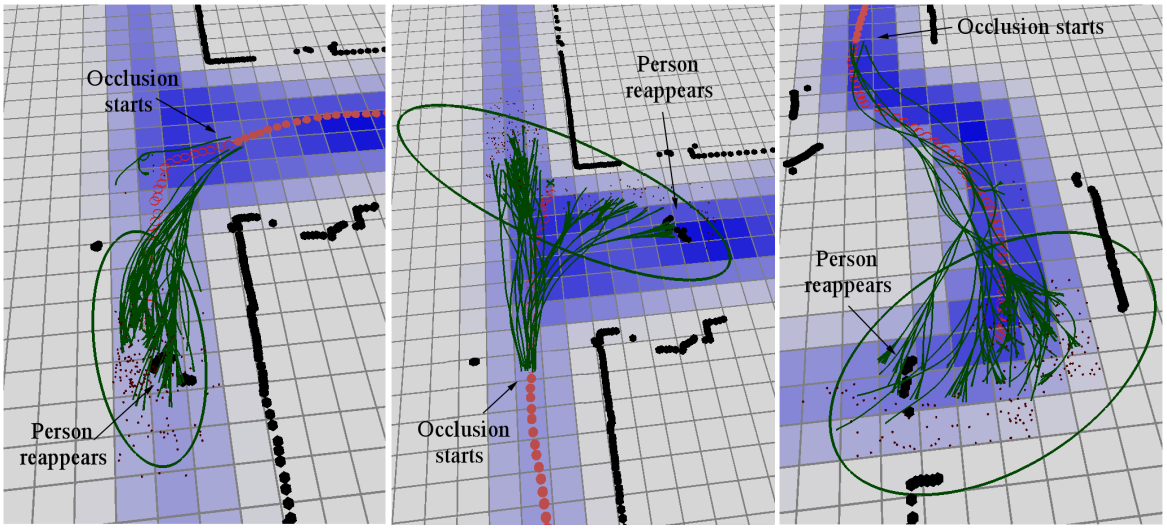


Figure 6.5: Place-dependent motion model in three example situations. The figures show a maneuvering target that reappears after a very long occlusion event. The background grid contains the learned space usage probabilities of the spatial affordance map, thick black dots are laser measurements, small dots are the look-ahead particles, and the green ellipse illustrates the weighted 99% sample covariance from the particles. The model is able to predict the targets "around the corner" and along the high-probability ridges in the map, yielding correct motion predictions in this type of situations.

where $\vec{\mu}_t^{(k)}$ is the mean, the mode, a draw, or some other likely value associated with the density of $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}^{(k)})$, used to evaluate the goodness of the parent sample $\mathbf{x}_{t-1}^{(k)}$.

The approximated joint density is designed such that one can sample from $g(\mathbf{x}_t, k \mid \mathbf{x}_{t-1}, \mathbf{m})$ by first sampling the index according to the pseudo-weight $\lambda_k \propto g(k \mid \mathbf{x}_{t-1}, \mathbf{m})$ and then sampling from the corresponding motion model $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}^{(k)})$, where

$$\begin{aligned} g(k \mid \mathbf{x}_{t-1}, \mathbf{m}) &= \int p(\vec{\mu}_t^{(k)} \mid \mathbf{m})\; p(\mathbf{x}_t \mid \mathbf{x}_{t-1}^{(k)})\; w_{t-1}^{(i)}\, d\,\mathbf{x}_t \\ &= p(\vec{\mu}_t^{(k)} \mid \mathbf{m})\; w_{t-1}^{(i)}. \end{aligned} \tag{6.30}$$

The weights of the new samples $\mathbf{x}_t^{(i)}$ at time t are finally computed by (replacing Eq. 6.25)

$$w_t^{(i)} = \frac{p(\mathbf{x}_t^{(i)} \mid \mathbf{m})\; p(\mathbf{x}_t \mid \mathbf{x}_{t-1}^{(k)})\; w_{t-1}^{(i)}}{p(\vec{\mu}_t^{(k)} \mid \mathbf{m})\; p(\mathbf{x}_t \mid \mathbf{x}_{t-1}^{(k)})\; w_{t-1}^{(i)}} = \frac{p(\mathbf{x}_t^{(i)} \mid \mathbf{m})}{p(\vec{\mu}_t^{(k)} \mid \mathbf{m})}. \tag{6.31}$$

In the concrete case, the above mentioned curvilinear motion model is used to compute a *look-ahead particle* as the future estimate $\vec{\mu}_t^{(k)}$ (see Figure 6.5). This is done by propagating the k-th sample at time step $t-1$ l steps into the future, that is, the sample is forward-simulated via the motion model over a time interval $l\,\Delta t$. The value that is finally taken for $p(\vec{\mu}_t^{(k)} \mid \mathbf{m})$ is then the value of $p(\mathbf{x}_t^{(k)} \mid \mathbf{m})$ evaluated at the position $(\; x_t \;\; y_t \;)^T$ of the look-ahead particle k.

Once the new motion model is obtained in the form of a set of weighted samples, it needs to be integrated into the MHT framework. Since the MHT relies on the Kalman filter for tracking, the

first two moments are computed as

$$\hat{\mu} = \sum_i w^{(i)} \mathbf{x}_t^{(i)} \quad \text{and} \quad \hat{\Sigma} = \frac{1}{1 - \sum_i (w^{(i)})^2} \sum_i w^{(i)} (\hat{\mu} - \mathbf{x}_t^{(i)})(\hat{\mu} - \mathbf{x}_t^{(i)})^T. \tag{6.32}$$

The target is then predicted using $\hat{\mu}$ as the state prediction with associated covariance $\hat{\Sigma}$. Obviously, the last step is not needed when using particle filters for tracking. Example situations that illustrate the place-dependent motion model are shown in Figure 6.5.

6.6 Experiments

For the experiments four data sets, two in indoor and two in outdoor environments, have been collected. The data sets are from a laboratory (Figure 6.6), an office building (Figure 6.9), the main station of Freiburg and a busy pedestrian zone in Freiburg downtown (Figure 6.8 and Figure 6.7). As sensor a SICK LMS 291 laser scanner with an angular resolution of 0.5 degrees mounted at $\sim$0.85 meter height and an acquisition rate of 12 Hz is used.

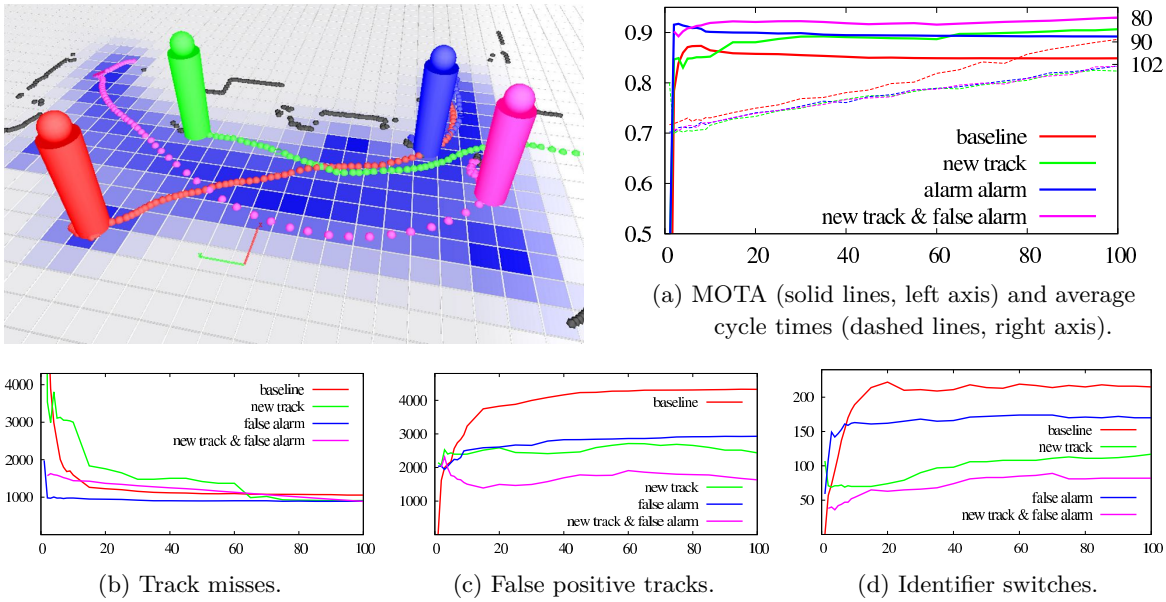
The spatial affordance maps have been trained with the baseline MHT tracker by Cox and Hingorani [1996] with a detection probability of $p_{det} = 0.999$, a termination likelihood $\lambda_{del} = 20$, and 300 hypothesis. The parameters of the tracker have been learned from training data with 95 labeled tracks over 28,242 frames. All detections and data associations including occlusions have been annotated by hand to determine ground truth information. This led to a fixed Poisson rate for new tracks $\lambda_{new} = 0.0002$ and a fixed Poisson rate of false alarms $\lambda_{fal} = 0.0041$, respectively. The rates have been estimated using the Bayesian approach in Eq. 6.13. Care has to be taken that the estimates of the expected number of events are normalized with the sensor field of view V. The grid cells of the map were chosen to be 30 cm in size. After the learning phase the map is assumed to be fixed. As pruning strategy the N-scan-back logic at a tree depth of 30 is employed. Additionally, the maximum number of hypotheses is limited to N_{Hyp} using the multi-parent variant of the algorithm proposed by Murty [1968] and discussed in subsection 3.8.2.

The parameters of the place-dependent motion model have been set to 300 samples, $\sigma^2_{a_{tan}} = 0.1$ and $\sigma^2_{a_{nor}} = 0.8$ for the noise in the tangential and normal accelerations, respectively, and $l = 5$ as look-ahead factor to compute the pseudo-weights λ_k.

6.6.1 MHT Data Association With Spatial Target Priors

In the first experiment, the original MHT approach is compared to the tracker using spatio-temporal prior information encoded in the spatial affordance map on the laboratory data set over 38,994 frames and with a total number of 134 people entering and leaving the sensor field of view. As mentioned, the data association ground truth of the 134 tracks has been determined manually.

To compare the impact of the presented models onto the tracking performance the individual models are test first against the baseline tracker. Afterwards, the combination of both models is evaluated. The accuracy of the resulting strategies is measured using the CLEAR MOT metrics proposed by Bernardin and Stiefelhagen [2008]. The metric counts three numbers with respect to the ground truth that are incremented at each frame: misses (missing tracks that should exist at a ground truth position), false positives (tracks that should not exist), and mismatches (track identifier switches). The latter value quantifies the ability to deal with occlusion events. From these numbers, two values are determined: MOTP (average metric distance between estimated targets and ground truth) and MOTA (the average number of times of a correct tracking output with respect to the ground truth). As it is based on a metric ground truth of target positions which is unavailable in the data MOTP is ignored. In order to show the evolution of the error as a function of N_{Hyp} which



(a) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

(b) Track misses.

(c) False positive tracks.

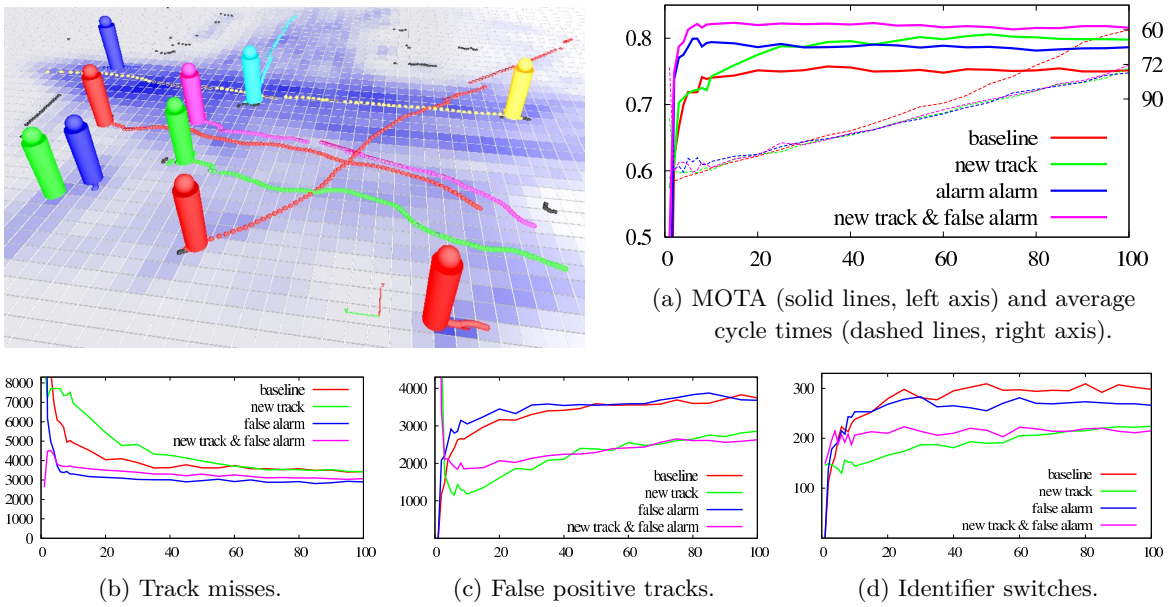
(d) Identifier switches.

Figure 6.6: Four of 134 example tracks from the laboratory data set (top left). Accuracy of the different tracking approaches (MOTA, top right), total number of mismatches, misses, and false positives as a function of N_{Hyp} (bottom, from left to right). The solid red lines show the results of the baseline MHT with fixed Poisson rates for new tracks and false alarms. The green and blue lines stand for the extended approach using the spatial priors for new tracks and false alarms, respectively. The results for the combined approach is denoted by the line in magenta. The tracker cycle times are underlaid with dotted lines in the top right diagram. The graphs show that when replacing the fixed Poisson rates by the learned, place-dependent ones, the tracker makes significantly fewer errors at slightly faster cycle times.

is proportional to the computational effort, N_{Hyp} is varied from 1 to 100. The results are presented in Figure 6.6 and Table 6.1.

The results show a significant improvement of the extended MHT with spatial priors over the regular approach especially for the number of mismatches. For $N_{Hyp} = 100$ the tracker makes 135 fewer id switches (217 vs. 82), the number of false positives decreases from 4328 to 1632, and the number of misses from 1055 to 887. The accuracy (MOTA) increases from 84.8% to 92.9%. The place-dependent new track and false alarm models applied in isolation (blue and green lines) already lead to a performance increase over the baseline MHT.

The insights into these improvements are as follows. As can be seen in Figure 6.2 right, few new track events have been observed in the center of the room. If, for instance, a track occlusion occurs at such a place (e.g. from another person), hypotheses that interpret this as an obsolete track followed by a new track receive a much smaller probability through the spatial affordance map than hypotheses that assume this to be an occlusion. The improvement in the false positive error is explained by fewer incorrect track creations in clutter. This is due to both, lower new track probabilities at such places and higher false alarm rates in regions of clutter. The combined approach benefits from both aspects and further reduces this type of error. Fewer misses are due to earlier track creations. Especially the modes in the new track distribution around doors allow the system to initialize tracks faster than with a fixed rate. Short sequences of observations are also tracked more accurately causing fewer



(a) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

(b) Track misses.

(c) False positive tracks.

(d) Identifier switches.

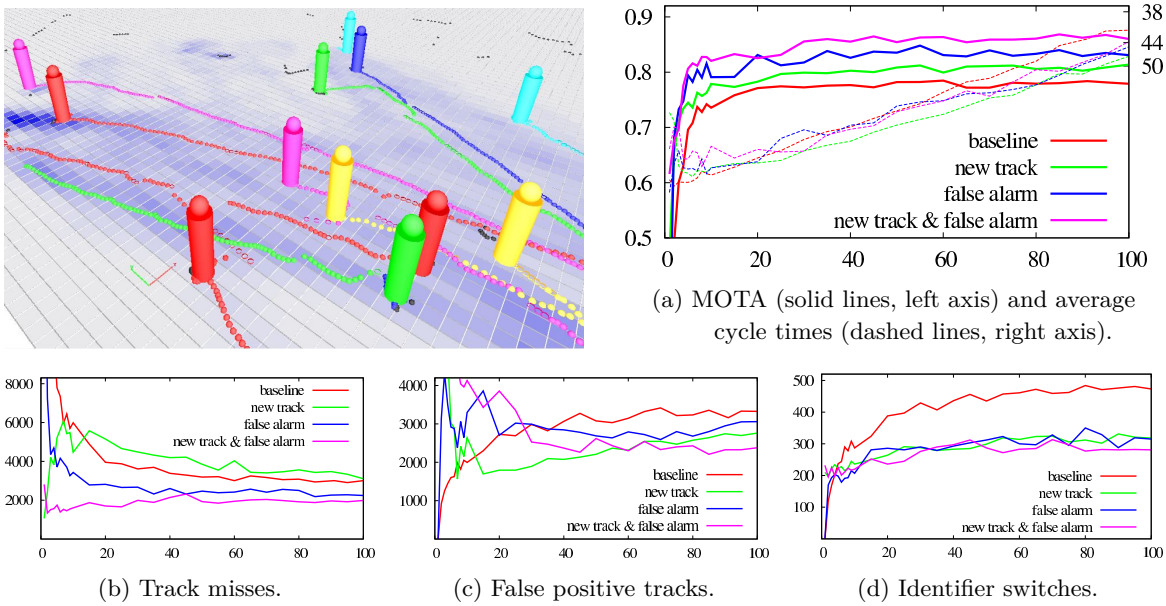
Figure 6.7: Ten of 168 example tracks from the experiment in the city center of Freiburg (top left). Accuracy of the different tracking approaches (MOTA, top right) and total number of mismatches, misses, and false positives as a function of N_{Hyp} (bottom, from left to right). The red, green, blue and magenta lines denote the baseline, the place-dependent false alarm and new track models and the combined approach, respectively. Again, the diagrams show that the place-dependent models significantly improve tracking performance.

errors of this type.

An additional data set with 15 people was collected to investigate whether the model is overfitted and generalizes poorly for unusual behavior of people. In this experiment, subjects entered the sensor field of view through entry points that have never been used (in between the couch and the desk at the bottom in Figure 6.2) or appeared in the center of the room by jumping off from tables. Manual inspection of the resulting trees (using the graphviz-lib for visualization) revealed that all 15 people are tracked correctly. The difference to the approach with fixed Poisson rates is that, after track creation, the best hypothesis is not the true one during the first few (less than five) iterations. However, the incorrect hypotheses that successively postulate the subjects being a false alarm become very unlikely, causing the algorithm to backtrack to the true hypothesis within milliseconds.

To demonstrate the scalability of the proposed extensions, they are evaluated in two unscripted large-scale outdoor scenarios. The first data set has been collected in a pedestrian zone in the city center of Freiburg and the second one in an underground hall in the Freiburg main station, both during a regular workday. The data sets consist of 55,475 frames during 25 minutes and 33,204 frames during 15 minutes, respectively. 10,000 frames with 168 persons and 6,000 frames with 160 persons have been labeled manually, again to determine the data association ground truth. These data sets are more difficult as occlusions are even more likely in scenarios with many people moving in a large space. The data sets contain up to 19 simultaneously visible targets with very frequent occlusions from other individuals or obstacles in the environment.

In Chapter 5 human-specific models using predefined areas of increased or decreased false alarms and new tracks likelihood are presented. As both approaches aim at the same their contribution to the tracking accuracy are compared in this section. The results of both approaches on the outdoor



(a) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

(b) Track misses.

(c) False positive tracks.

(d) Identifier switches.

Figure 6.8: 12 of 160 example tracks from the experiment in an underground hall of the Freiburg main station (top left). Accuracy of the different tracking approaches (MOTA, top right) and total number of mismatches, misses, and false positives as a function of N_{Hyp} (bottom, from left to right). The red line shows the baseline MHT with fixed Poisson rates, the green and blue lines stand for the system extended by the new track and false alarm models, the magenta line denotes the combined approach. Like in the indoor experiment, the diagrams show that the combined approach significantly reduces the number of mismatches, false positives, and misses.

data sets are presented in Table 6.1.

The results of the first outdoor experiment in the city center of Freiburg show that the extended MHT with spatial priors yields significant improvements comparable to the indoor data set (see Figure 6.7). As shown in Table 6.1 at $N_{Hyp} = 100$ the accuracy (MOTA) is increased by 6.5% (75.1% vs. 81.6%). Since the environment contains many regions of clutter, the number of false positives (FP) decreases substantially (3744 vs. 2627). The number of mismatches (ID) has also dropped from 298 to 211, and the number of misses (FN) decreased from 3440 to 3070. The "background learning ability" of the false alarm layer in the spatial affordance map is particularly appropriate in this data set as the environment contains a couple of person-shaped objects (trees, chairs, trash bins) that led to many false positives of the detector. The baseline approach with fixed rates of new track and false alarm events was not able to cope well with such systematic detection errors and incorrectly created tracks at these locations.

In comparison to the modeling approach proposed in Chapter 5 learning spatio-temporal target priors from human observations enables the system to adapt to human-specific behaviors which might not be predictable even by a human expert. Thus the improvements made by these manually designed models can further be improved. Especially, using the combined approach of learned new track and false alarm rates reduces the numbers of false positives and track identifier switches dramatically.

The results of the Freiburg main station show an even larger improvement (see Figure 6.8). At $N_{Hyp} = 100$ the accuracy (MOTA) increases from 79.4% to 86%. A detailed analysis of the CLEAR MOT metrics shows that the number of mismatches drops from 473 to 281 which is an improvement

Data Set	Model(s)		FN	FP	ID	MOTA	*Hz*
indoor		baseline	1055	4328	217	84.8%	87.3
	learned	new track	910 (-13.7%)	2436 (-43.7%)	117 (-46.1%)	90.6%	106
		false alarm	893 (-15.4%)	2929 (-32.3%)	170 (-21.7%)	89.2%	103
		new + false	887 (-15.9%)	1632 (-62.3%)	82 (-62.2%)	92.9%	103
Freiburg city center		baseline	3440	3744	298	75.1%	59.9
	modeled[36]	new track	3317 (-3.6%)	3550 (-5.2%)	239 (-19.8%)	77.9%	72.6
		false alarm	2940 (-14.5%)	3949 (+5.5%)	274 (-8.1%)	77.7%	68.5
		new + false	3328 (-3.3%)	3409 (-8.9%)	236 (-20.8%)	77.9%	76.1
	learned	new track	3424 (-0.5%)	2855 (-23.7%)	224 (-24.8%)	79.7%	77.1
		false alarm	2901 (-15.7%)	3682 (-1.7%)	266 (-10.7%)	78.6%	74.5
		new + false	3070 (-10.8%)	2627 (-29.8%)	211 (-29.2%)	81.6%	72.2
Freiburg main station		baseline	3006	3327	473	79.4%	41.4
	modeled	new track	3160 (+5.1%)	3189 (-4.1%)	351 (-25.8%)	79.5%	49.6
		false alarm	2343 (-22.1%)	3764 (+13.1%)	397 (-16.1%)	80.4%	45.3
		new + false	3219 (+7.1%)	3074 (-7.6%)	358 (-24.3%)	79.9%	48.5
	learned	new track	3105 (+3.3%)	2761 (-17.0%)	318 (-32.8%)	79.5%	47.3
		false alarm	2243 (-32.2%)	3059 (-8.1%)	315 (-33.4%)	83.1%	44.9
		new + false	1985 (-33.9%)	2378 (-28.5%)	281 (-40.6%)	86.0%	43.8

Table 6.1: Comparison of the CLEAR MOT results of the outdoor data sets using $N_{Hyp} = 100$ hypotheses. Employing place dependent models (introduced in Chapter 5) or learned spatial priors for the events of new tracks, false alarms, occlusions, and deletions, respectively, the number of identifier switches (ID) can be increased dramatically.

of over 40%. The number of false positives decreases from 3327 to 2378 (−28.5%) and the number of misses is reduced from 3006 to 1985 (−33.9%), respectively. The extensions lead to a slightly faster system. This behavior is due to a smaller number of tracks in the system that the regular approach had created for each incorrect track.

Compared to the modeling method (see Chapter 5) the results demonstrate that the proposed learning approach is more robust to unexpected human behavior. When the manually modeled information is inaccurate or contains errors (as shown in Figure 5.2, in subsection 5.6.1 an escalator was intentionally not annotated) the tracking accuracy can even get worse indicated by an increased number of misses (FN) by 7.1%. Using the learning technique human behavior is correctly modeled leading to an improvement in the (FN) by 33.9%. The faster run-time of the modeled approach (48.5 Hz vs. 43.8 Hz) is caused by the missed targets that are not tracked.

In the diagrams of all three experiments, the number of misses tend to decrease and the number of false positives tend to increase over N_{Hyp}. This behavior is explained by the fact that false alarms

[36] The results of the place dependent models using predefined areas of more or less likely new track and false alarm events are explained in more detail in section 5.6.

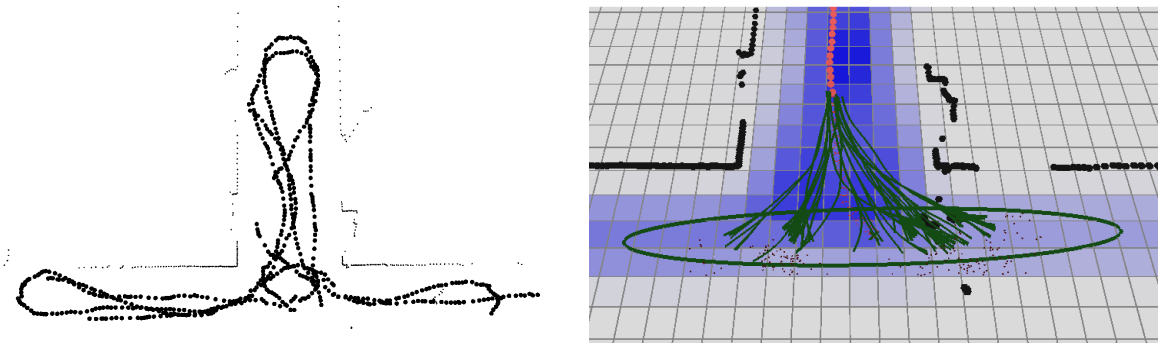


Figure 6.9: *Left*: Six (of 50) example tracks from experiment 2. *Right*: Motion prediction of the place-dependent model during a long occlusion event. The occlusion event starts one meter before the person is taking a left turn. The samples and the shape of the covariance predictions follow the walkable area map and predict that the person turns into one of the two possible corridors. The constant velocity motion model would predict the person directly into the wall behind the sensor (not visible in the figures).

are more likely than new tracks. Hypotheses that postulate observations as false alarms receive higher probabilities and can dominate the hypothesis ranking. This can lead to the rejection of lower probability hypotheses at small values for N_{Hyp}, that would have correctly interpreted observations as new tracks. With increasing N_{Hyp} more new track hypotheses survive the pruning process and the number of misses decreases.

The noise in the error plots such as the number of mismatches, for instance, means that more hypotheses do not always lead to a smaller error, which is counterintuitive. This is due to the pruning strategies in combination with numerical issues in the MHT. It follows from the combinatorics of the approach that several hypotheses can have the same probability value. If N_{Hyp} happens to prune within such a plateau in the distribution, the outcome of the tracker can partly become unpredictable since it depends on the order in which these hypotheses are stored in memory.

In addition to the improvement in tracking performance, the extended tracker is also more efficient than the baseline. As the new approach commits fewer track creation error, it has to maintain fewer tracks on average, especially in regions of clutter.

6.6.2 Place-Dependent Motion Model

In this section, the place-dependent motion model proposed in section 6.5 is evaluated. A data set has been collected in an office environment and divided into a training set and a test set. The training set contains 7,443 frames with 50 person tracks and has been used to learn the spatial affordance map (see Figure 6.9). To learn the walkable area map, the track confirmation events of the best hypothesis are counted. The test set with 6,971 frames and 28 people tracks was used to compare the model with a constant velocity motion model under different conditions. The data set was labeled by hand to determine both, the ground truth (x, y)-positions of subjects and the true data associations.

To analyze the robustness and accuracy of the new place-dependent motion prediction model, in a first experiment, areas are defined in which target observations are ignored as if the subjects had been occluded by an object or another person. These areas were placed at hallway corners and U-turns where people typically perform maneuvers. As the occlusions are only simulated, the ground truth position of the targets are still available. See Figure 6.5 for example frames.

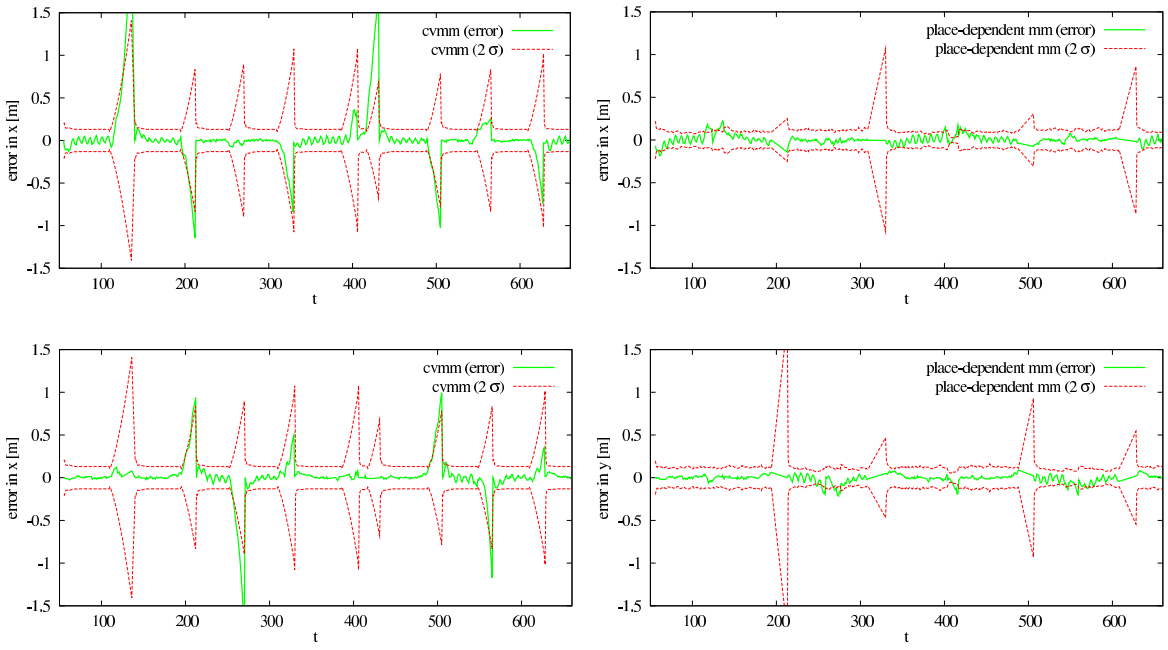


Figure 6.10: Estimation error in x of the constant velocity motion model (cvmm, left) and the place-dependent motion model (right). Peaks correspond to occluded target maneuvers. See also Figure 6.5, center, which shows the right turn of a person in this experiment. While both approaches are largely consistent from an estimation point of view, the place-dependent model results in an overall smaller estimation error and smaller uncertainties. For 28 manually inspected tracks, the constant velocity motion model lost a track 12 times while the new model had only one track loss.

For the 28 manually inspected tracks of the test set, the constant velocity motion model lost a track 12 times while the new model had only a single track loss. Clearly, as a naive countermeasure, one could enlarge the process noise covariance of the constant velocity motion model to avoid such losses. But in the multi-target case considered here, this is the wrong way to go that leads to enlarged validation gates and increased levels of data association ambiguity. Consequently, probability distribution over pruned hypothesis trees will be less accurate and lead to a less efficient tracker.

As a measure of metric accuracy, the resulting estimation errors in x are shown in Figure 6.10 (the errors in y are similar).

The diagram shows smaller estimation errors and 2σ bounds for the place-dependent motion model during most target maneuvers. The predicted covariances do not grow boundless during the occlusion events (peaks in the error plots) since the shape of the covariance predictions follows the walkable area map around the very position of the target. Example situations of this behavior are illustrated in Figure 6.5.

In a second experiment, the observation frequency is scaled down to 0.5 Hz where the tracker is allowed to initialize its targets for one second. The internal cycle time of the tracker was left unchanged at 12 Hz. This setting simulates a very slow data acquisition sensor or the realistic situation of an embedded CPU on which people detection runs concurrently with many other processes at a low rate.

The constant velocity motion model was not able to follow the maneuvering targets and lost all of them as soon as they passed the corner of the hallway. The place-dependent motion model was able

to predict the targets around corners as seen in Figure 6.5 and lost only six of the 28 tracks. The six track losses occurred when the targets were last observed briefly before the turn into the larger corridor. Explained by the zero mean noise in tangential acceleration most of the samples followed a straight motion.

6.7 Conclusions

This chapter presented the spatial affordance map for the purpose of extending a people tracker with spatial priors on human behavior. The problem of learning spatio-temporal models of human behavior is posed as a parameter estimation problem of a non-homogeneous spatio-temporal Poisson process. The model is learned using Bayesian inference from observations of track creation, confirmation, and false alarm events. It enables to overcome the usual fixed Poisson rate assumptions for new tracks and false alarms and to learn a place-dependent model for these events. Finally, it is shows that the Poisson process can be seamlessly integrated into the framework of an MHT tracker.

In large-scale experiments in different indoor and outdoor scenarios, it is demonstrated that the extended tracker is significantly more accurate in terms of the CLEAR MOT metrics. In particular, the number of track identifier switches could have been reduced by at least 36% up to several factors. This error is the most relevant metric for a people tracker as it quantifies the ability to keep correct identities over occlusion events and missed detections. The number of false positives dropped by at least 45% while track misses decreases at least by 17%.

The map further allowed us to derive a novel, place-dependent model for predicting maneuvering targets during lengthy occlusion events. The model is based on a walkable area map derived from the learned rate function of track confirmation events and uses an auxiliary particle filter that probes the map at locations of a look-ahead particle. In the experiments, the tracker could follow highly maneuvering people at an observation frequency as low as 0.5 Hz, clearly outperforming the constant velocity motion model in terms of track losses.

In the future, an extended representation of the spatial affordance map using non-stationary Poisson processes is planed to capture the requirements of mobile robots. Additionally, other place-dependent distributions (i.e. occluded and deleted track events) will be investigated. To support motion prediction the ability of the spatial affordance map to learn and predict short and long term goals of people should explored. This will allow to reason about the space usage in different period of time and of different activities. Moreover, long-term experiments have to be conducted to show that life-long learning of spatio-temporal dependent information improves motion prediction and data association.

7 Learning Social and Geometrical Relations for Group Tracking

Detecting and tracking people and groups of people is a key competence for social robots operating in populated environments. In many scenarios, people are encountered in groups formed by social and spatial relations between individuals. This chapter, addresses the problem of detecting, learning, and tracking such socio-spatial relations. Inferring from learned group affiliations improves tracking accuracy significantly, especially, when group members are occluded over extended periods of time. Assuming a mobile sensor that perceives the scene from a first-person perspective good tracking performance is achieved in real-time using only 2D range data.

Opposed to related work, the proposed method tracks individuals and reasons about multiple social grouping hypotheses in a recursive way using an extended multi-model multi-hypothesis tracking approach. Priors from the social science community guide the creation of social grouping hypotheses and inform on-line learning of spatial relations. Person-level tracking is improved in two ways: the social grouping information is fed back to jointly predict human motion over learned intra-group constraints using a particle-based approach. Furthermore, data association is supported by adapting track-specific occlusion probabilities for people sharing social relations. Both components lead to an improved occlusion handling, a reduced number of identifier switches, and a better trade-off between false negative and false positive tracks.

Experiments with a mobile robot equipped with two 2D laser range finders and on large-scale outdoor data sets demonstrate that the approach is able to model social grouping and to improve person tracking significantly. A reduction of track identifier switches by almost 50% and more than 28% less false negative tracks is achieved. Combining the approach with the previously presented physically grounded occlusion model yields even further improvements. The approach runs in real-time on a laptop PC embedded on the mobile robot.

This chapter is organized as follows: the introduction is provided in section 7.1 followed by a discussion of related work in section 7.2. The theory on social relation recognition is introduced in section 7.3. Section 7.4 describes how the geometric relations between people are learned in an on-line fashion. Section 7.5 presents the socially constraint motion model followed by section 7.6 that details the integration into the original multi-hypothesis tracking (MHT) framework and the extended multi-model MHT. Experiments are shown in section 7.7. Finally section 7.8 concludes the chapter.

7.1 Introduction

As robots enter domains in which they interact and cooperate closely with humans, people tracking becomes a key technology for many research and application areas in robotics, intelligent vehicles, and interactive systems. A particularly difficult problem is maintaining the identity of people in crowded scenarios. Such scenarios are highly important since an empirical experiments by Moussaïd et al. [2010] investigated that typically up to 70% of the pedestrians walk in groups. In the light of this, the problem of detecting, representing, and tracking groups of people, particularly from mobile platforms,

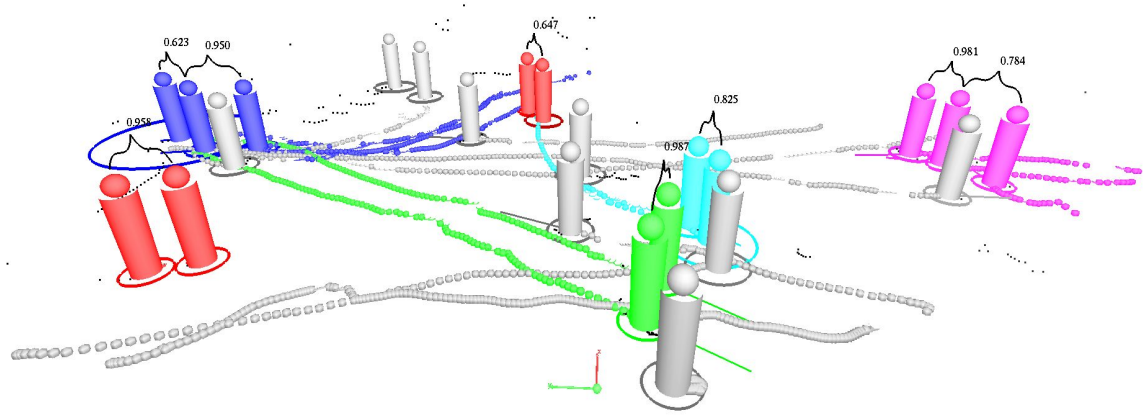


Figure 7.1: A situation recorded in the inner city of Freiburg showing 23 tracks organized in 6 groups (shown in the same color) and several individuals (shown in gray). Cylinders denote the positions of tracked pedestrians, the colored dots illustrate their past trajectories. The numbers on top of the cylinders show social relation probabilities between group members. Gray individuals that appear close to groups are correctly recognized to not belong to the groups as their motion directions, shown by the traces, are different.

is clearly underexplored. This section addresses these problems as they are relevant for a number of scenarios including multi-party human-robot interaction and collaboration, efficient and socially compliant robot navigation among people, analysis of social group activities, and understanding of social situations.

First, the problem is stated as an estimation problem of pairwise social relations between individuals from perceived track motion features using linear SVM classifiers and Bayesian smoothing. Since the spatial organization of groups is typically not random and remains largely stable over time, group-specific geometric relations[37] between individuals are learned in an on-line fashion using the current motion states and prior information provided in Moussaïd et al. and Yücel et al.. The grouping information is then fed back to improve person-level tracking by predicting occluded tracks jointly over the learned geometric intra-group constraints.

Opposed to previous works in which geometric relations and social groupings are only detected on a per-frame basis or found in an a posteriori fashion by batch methods, the proposed approach explicitly models and tracks group formations over time in a recursive multi-hypothesis model selection and data association framework. The recursiveness implies an anytime property where the tracker always provides a current best (but suboptimal) estimate that is refined with more incoming information. Using multi-hypothesis tracking, this happens by backtracking to branches in the hypothesis tree that become more probable with the new evidence. In contrast to batch methods, this property is crucial for mobile robots that need to take real-time decisions for interaction or navigation in unfolding social situations. The proposed approach, evaluated on 2D range data, results in substantially more accurate and fast social grouping estimates and outperforms several multi-hypothesis tracker variants. Combining the proposed approach with the physically grounded occlusion model introduced in subsection 5.4.1 leads to best results with more than 55% less track identifier switches and more than 35% less track misses compared to a state-of-the-art baseline approach.

[37] The geometric relations provide information about the spatial arrangement of people. However, to prevent confusion with the social relations they are not called spatial relation.

7.2 Related Work

The organization of pedestrian social groups, their impact on the complex dynamics of crowd behavior, and the identification of models are research subjects in the social science community. Empirical studies of McPhail and Wohlstein [1982], Moussaïd et al. [2010], and Yücel et al. [2012] provide relevant background knowledge in this context. They developed individual-based models to describe how people interact with group members and other pedestrians. Furthermore, it was found that the spatial organization of pedestrians into groups can be well described by only three motion features namely relative motion *direction*, relative *distance*, and relative *velocity*. These properties apply to pairs of people as well as to groups with many members.

Social grouping from sensory data has recently gained increasing attention by researchers from the computer vision and social computing communities. One group of works is concerned with the understanding of social situations. Using interpersonal distance and relative body orientation Groh et al. [2010] study social situation recognition of standing people from static cameras. Similarly, Cristani et al. [2011] address the problem of social relation recognition in standing conversation scenarios. Using interpersonal distance only, they estimate pairwise stable spatial arrangements called F-formations. Ge et al. [2009] focus on the analysis of social behavior by identifying small groups of people traveling together in crowds employing a pairwise distance that combines proximity and velocity cues informed by the sociological models of collective behavior found by McPhail and Wohlstein [1982]. An alternative approach is taken by Choi et al. [2011] who recognize collective human activities by modeling crowd context with spatio-temporal bins. A Random Forest structure is used to classify and localize collective activities in the scene.

A second group addresses social relation recognition in still images and video. In the work of Wang et al. [2010] social relations are extracted from photographs. They use the knowledge that social relations between people in photographs typically influence their appearance and relative image position. Learning a model on weakly labeled data representing social relations and appearances using features and an EM algorithm improves identification of unknown faces and allows to recognize relationships in previously unseen images. From the learned models, they are able to predict relationships in previously unseen images. Social relations between film actors in video are estimated by Ding and Yilmaz [2011]. A graphical social network representation with temporal smoothing is learned using actor occurrence patterns. The approach also allows for changes in social relations over time. Based on tracklet observations Choi and Savarese [2012] recognize atomic activities of individuals, interaction activities of pairs, and collective activities of groups, jointly, using an energy maximization framework. Corresponding tracklets are associated following the central idea that the atomic activities of individuals are highly correlated with the overall collective activity, through the interactions between people.

A third group, most related to this context, is concerned with detecting and tracking groups from image or range data. Yu et al. [2009] address the problem of discovery and analysis of social networks from individuals tracked in surveillance videos. A social network graph is built over time from observations of interacting individuals using face recognition and track matching. Expanding plain social relation recognition Fathi et al. [2012] present a method to detect and characterize social interactions in a day-long first-person video recorded with a wearable camera (egocentric video) at a social event. First-person motion is used to transform face detections into 3D space and serve to recognize specific roles of people in social interactions based on patterns of attention shift and turn-taking over time. Social relations between persons in overhead video data are recognized by Pellegrini et al. [2010]. They use approximate inference on a third-order graphical model to jointly reason about correct person trajectories and group memberships. Based on learned statistical models on people's behavior in groups, they also perform group-constraint prediction of motion. This results in improved

tracking performance of individuals. Leal-Taixé et al. [2011] model social and grouping behavior from tracked individuals in video data using a minimum-cost network flow formulation. The weighting of the edges in the network is informed by learned group models and a social force-enabled motion prediction. Qin and Shelton [2012] improve tracking of individuals by considering social grouping in a tracklet linking approach. Using large numbers of hypothetical partitionings of people into groups, solutions are evaluated based on the geometrical similarity of trajectories of individuals with the hypothesized group.

Groups of people are tracked in Lau et al. [2009, 2010] using data from a mobile robot equipped with a 2D laser range finder. As maintaining the state of individual people is intractable in densely crowded situations, a multi-model hypothesis tracking approach is developed to estimate the formation of tracks into groups that split and merge. Group formation behavior is modeled and tracked using constant prior probabilities of group continuation and split events and trajectory-dependent probabilities of group merge events. Groups are collapsed into single states loosing the individual person tracks. A very similar multi-model hypothesis approach has been developed independently by Chang et al. [2010] to track and group neural signals whose locations are inferred from clusters of observations. Based on the work of Wolf and Burdick [2009] tracking is combined with multiple cluster hypotheses (cluster models) in a single sound MHT framework.

In contrast, this chapter extends the state of the art as follows:

Opposed to Groh et al., Cristani et al., Yu et al., Pellegrini et al., Leal-Taixé et al., or Qin and Shelton which rely on static overhead cameras to perceive the scene, the problem of social grouping and tracking is addressed from a mobile sensor and a first-person perspective. Overhead cameras are sufficient for surveillance but fall short of scenarios where a robot, an intelligent vehicle, or an interactive system coexists and acts in the same space with people. Occlusions and misdetections occur much more often from an in-scene view than in an overhead setup. Thus, the goal is to make tracking particularly robust with respect to lengthly occlusion events and the mobility of the sensor.

Additionally, the problem of using 2D range data which add the difficulty that targets have the same appearance and cannot be distinguished to guide data association is addressed. The use of 2D range data is relevant for robots and intelligent vehicles where such sensors are used due to their large field of view and robustness with respect to illumination and vibration. The proposed approach can obviously be applied to other sensory data, too. Furthermore, unlike Yu et al., Pellegrini et al., Leal-Taixé et al., Qin and Shelton, or Choi and Savarese which employ (partly very slow) batch methods, the proposed approach runs recursively and in real-time on a laptop PC – again highly relevant for interactive robots or vehicles that need to respond quickly to group formation changes and unpredictable events.

Unlike Groh et al. or Cristani et al. that detect and analyze social relations on a per-frame basis, such relations and the inferred social groupings are tracked over time. To this end, the multi-model hypothesis tracking approach developed by Lau et al. and Chang et al. is adopted. The method suits the problem as it allows to simultaneously hypothesize about the clustering of tracks (models) and the assignment of measurement to tracks (data association) in a consistent probabilistic framework. However, opposed to Lau et al. who represent groups of people in a single collapsed state without spatial extension information, the proposed approach keeps track of both the state of individual group members and the group affiliation. This allows for a much more detailed group analysis such as estimating the group's spatial extension or understanding group activities and social situations.

Further related works from the target tracking community consider aerial tracking of convoys of ground vehicles. In Mallick et al. [2011] a new filtering algorithm that integrates inter-vehicle constraints is presented to predict the motion of a group of vehicles moving along a one dimensional lane. Using a velocity-dependent safe driving distance motion is predicted from the front to the rear. This approach is extended to incorporate geometrical constraints of multiple group members in 2D.

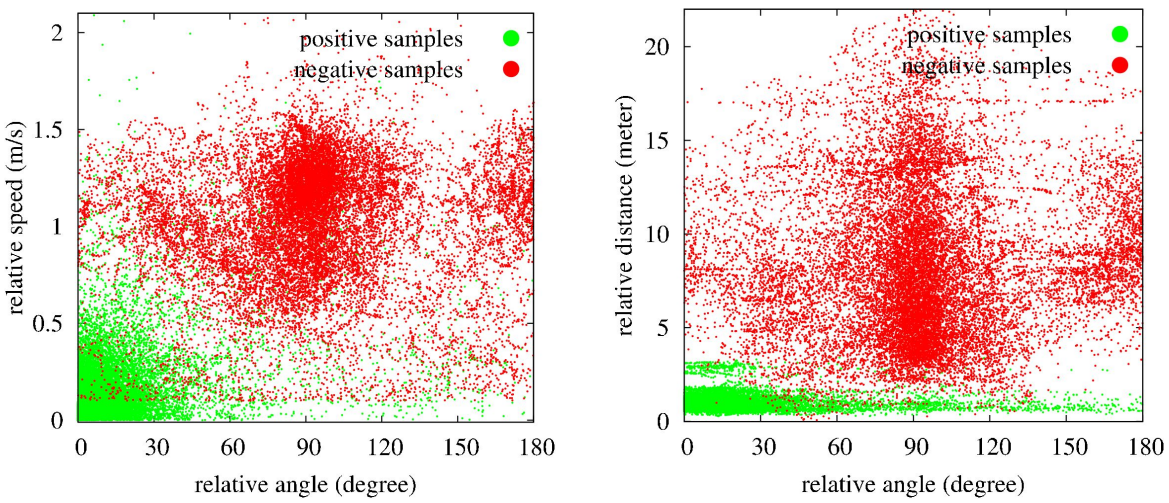


Figure 7.2: Distributions of feature values of positive and negative samples. *Left:* Relative angle vs. relative speed of tracks in the same group (green dots) and tracks not in the same group (red dots), respectively. *Right:* Relative angle vs. relative distance. The distributions are consistent with the empirical observations in McPhail and Wohlstein [1982], Moussaïd et al. [2010], and Yücel et al. [2012] and show a good separability of the data.

7.3 Social Relations Learning and Group Detection

Empirical social science studies by McPhail and Wohlstein [1982] and Moussaïd et al. [2010] have found three dominant coherent motion indicators of people that walk in groups: relative distance, relative orientation, and similar velocity to their direct neighbors. Using this insight, social relation candidates are detected by classifying pairs of tracks according to their relative motion properties. The relations are used to build up a weighted social network graph (social relation graph), hereafter called $\mathcal{G} = \{U, E\}$, in which each person corresponds to a node $u_i \in U$ and edges $\varepsilon_{i,j} \in E$ are weighted with the probabilities of the pairwise relation between the pair of persons i and j. Once the graph is constructed a graph-cut algorithm is applied to extract the groups denoted with $\mathcal{G}_k$[38].

7.3.1 Detection of Pairwise Social Relations

A pairwise relation candidate is obtained by computing the three coherent motion indicators proposed by McPhail and Wohlstein [1982] for two tracks of people that are: *relative distance*, *relative angle*, and *relative velocity*, respectively, and by classifying the sample.

Assuming tracks of people to be represented by $\mathbf{x}_t = (\ x_t \ \ y_t \ \ \dot{x}_t \ \ \dot{y}_t\ )^T$ encoding position and velocity in 2D space at time t with orientation $\phi_t = \text{atan2}(\dot{y}_t, \dot{x}_t)$ and velocity $v_t = \sqrt{\dot{x}_t^2 + \dot{y}_t^2}$. While constructing the social network graph $\mathcal{G}$ all pairs of people i, j are inspected and classified to share a social relation or not based on the calculated features. With $\mathbf{x}_t^i$ being the state of persons i and $\mathbf{x}_t^j$ the state of person j, respectively, the feature values are defined as:

(1) relative distance: $\Delta x_t^{i,j} = \sqrt{(x_t^i - x_t^j)^2 + (y_t^i - y_t^j)^2}$,

(2) relative orientation: $\Delta \phi_t^{i,j} = |\phi_t^i - \phi_t^j|$, and

[38] A group $\mathcal{G}_k$ found in the social network graph $\mathcal{G} = \{U, E\}$ corresponds to the subgraph $\mathcal{G}_k = \{U_k, E_k\}$ with $U_k \subset U$,$E_k \subset E$, and $\forall i\ \exists j\ \varepsilon_{i,j} > \theta$. Where θ is the probability threshold indicating that persons i and j share a social relation.

(3) relative velocity: $\Delta v_t^{i,j} = |v_t^i - v_t^j|$.

Combining these values the feature vector $\mathcal{F}_{i,j}$ of the pair i, j yields

$$\mathcal{F}_t^{i,j} \;=\; \left\{\Delta x_t^{i,j},\ \Delta\phi_{i,j}^t,\ \Delta v_t^{i,j}\right\}. \tag{7.1}$$

From training data (see section 7.7) a set of positive ($\mathcal{F}^+$) [39] and negative ($\mathcal{F}^-$) samples visualized in Figure 7.2 can be extracted and used to train a classifier. After training the probability of a social relation $\mathcal{S}_t^{i,j}$ between persons i and j at time t is obtained from the classifier output $p(\mathcal{F}_t^{i,j})$. For classification a linear SVM proposed by Platt [2000] is employed that provides both, the class label $l(\mathcal{F}_t^{i,j}) \in \{\mathcal{F}^+, \mathcal{F}^-\}$ and a probability estimate $p(\mathcal{F}_t^{i,j}) \in [0, 1]$. The set of social relations of person i to all his or her neighbors $j_1, \ldots, j_{N_i}$ at time t is denoted as

$$\mathcal{S}_t^i \;=\; \left\{\mathcal{S}_t^{i,j_1}, \ldots, \mathcal{S}_t^{i,j_{N_i}}\right\}. \tag{7.2}$$

In the following the double indexations i, j_k denoting features and social relations between pairs of people will be omitted for the sake of simplicity. The set of social relations shared between all existing people at time t is denoted as $\mathcal{S}(t)$.

7.3.2 Bayes Filtered Pairwise Social Relations

The detection results from the previous subsection rely only on information from a single frame/scan and are still noisy. Thus, the classification probabilities are integrated over time using a simple Bayes filter to achieve smoother and more stable probability estimates. Let $\mathcal{F}(t) = \{\mathcal{F}_0, \ldots, \mathcal{F}_t\}$ be the sequence of all observed feature values of a pair of people from time zero to the current time t. The probability $p(\mathcal{S}_t \mid \mathcal{F}(t))$ of a social relation $\mathcal{S}_t$ is then be calculated from the parent estimate $p(\mathcal{S}_{t-1} \mid \mathcal{F}(t-1))$ and the current feature values $\mathcal{F}_t$ in a recursive fashion using Bayes' rule

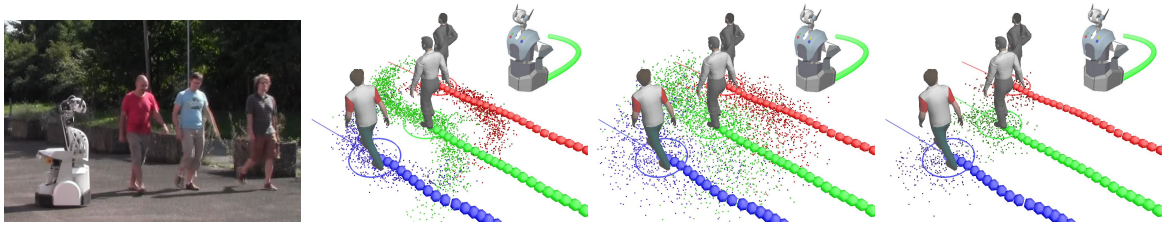
$$p(\mathcal{S}_t \mid \mathcal{F}(t)) \;= \eta\ \ p(\mathcal{F}_t \mid \mathcal{S}_t)\ \ p(\mathcal{S}_t \mid \mathcal{S}_{t-1})\ \ p(\mathcal{S}_{t-1} \mid \mathcal{F}(t-1)), \tag{7.3}$$

with $p(\mathcal{S}_t \mid \mathcal{S}_{t-1})$ encoding the event probabilities of social relations to arise, be confirmed, or to break, respectively. Assuming uniformly distributed prior probabilities for the states of the social relation $p(\mathcal{F}_t \mid \mathcal{S}_t)$ is transformed to obtain the likelihood $p(\mathcal{S}_t \mid \mathcal{F}_t)$ that is the classification result of the one-shot detection procedure explained in the previous subsection 7.3.1.

7.3.3 Detection of Groups

After the construction of the social network graph $\mathcal{G}$ in which the edges $\varepsilon_{i,j} \in E$ are weighted with the Bayes filtered social relation probabilities $p(\varepsilon_{i,j}) = p(\mathcal{S}_t^{i,j} \mid \mathcal{F}^{i,j}(t))$ defined in Eq. 7.3 the currently existing groups $\mathcal{G}_i(t)$ are detected using a simple graph-cut algorithm. Therefore, all edges $\varepsilon_{i,j} \in E$ with a probabilities lower than a given threshold θ are removed (cut). In a second step all nodes connected by the remaining edges – now encoding that people share a social relation with a high likelihood – are collected using depth first search and combined into subgraphs. After detecting all groups the corresponding persons of the nodes in the subgraphs are marked as members of the specific group. The information of the groups affiliation is later used to adapt the occlusion probability of people walking in groups. Further details are explained in subsection 7.6.1, an example tracking situation with six groups of two and three group members is shown in Figure 7.1.

[39] The feature values of the positive samples agree with statistics in McPhail and Wohlstein [1982], Moussaïd et al. [2010] and Yücel et al. [2012].



(a) Tracking three persons using the mobile robot Daryl.

(b) Spatial prior distributions modeled in Moussaïd et al..

(c) Predictied spatial distributions using Brownian motion.

(d) Resulting geometrical relations learned from multiple observations.

Figure 7.3: Learning geometrical relations while tracking a group of three persons using the mobile robot Daryl (a). The persons trajectories and current position estimates are shown with colored dots and circles, respectively. The spatial distributions are initialized with prior knowledge taken from Moussaïd et al. [2010] shown as colored particles (b). Uncertainty is modeled using Brownian motion (c). The spatial distributions are updated with new tracking information (d). The person in the middle (shown with green trajectory) shares social relations to both neighbors. Thus, two geometrical relations are learned.

7.4 On-line Learning of Geometric Relations

In addition to the social relation probabilities between people, geometric (or spatial) intra-group relations are learned in this section. Empirical analyses presented in Moussaïd et al. [2010] and Yücel et al. [2012] investigated the spatial arrangement of pedestrians and found that people in groups form specific stable patterns of geometrical group organizations. Thus, such patterns can be learned for person tracks in groups in an on-line fashion which amounts to estimating a track-specific spatial probability distribution of a person in the local reference frame of another person. Let $\Psi_t^{i,j}$ be the geometric relation of person track i in the local reference frame of track j and $p(\Psi_t^{i,j})$ its time-dependent distribution.

As learning is performed on a per pair basis a group of two persons i and j sharing a social relation $\mathcal{S}_t^{i,j}$ can be assumed without loss of generality. A Monte Carlo approach is taken to represent the geometric group relations[40] since their distributions have arbitrary shapes especially when used for human motion prediction (see Figure 7.3 or section 7.5). Thus, spatial relations are learned by recursively estimating a particle-based distribution

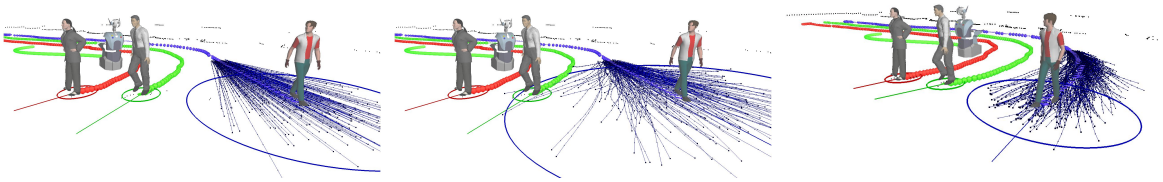
$$p(\Psi_t^{i,j} \mid \mathbf{z}_i^j(0), \ldots, \mathbf{z}_i^j(t)) \;=\; p(\mathbf{z}_i^j \,|\, \widehat{\Psi}_t^{i,j})\; p(\widehat{\Psi}_t^{i,j} \mid \Psi_{t-1}^{i,j})\; p(\Psi_{t-1}^{i,j} \mid \mathbf{z}_i^j(0), \ldots, \mathbf{z}_i^j(t-1)), \tag{7.4}$$

from sequences of relative track positions $\mathbf{z}_i^j(t)$. To update $\Psi_t^{i,j}$ over time the consecutive Kalman filtered observations $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$ of persons i and j, respectively, are integrate into the spatial distribution $p(\Psi_t^{i,j})$ as follows. Using Gaussian state estimates $\mathbf{x}_i(t) \sim \mathcal{N}_i(\mu_i, \Sigma_i)$ and $\mathbf{x}_j(t) \sim \mathcal{N}_j(\mu_j, \Sigma_j)$ the observations $\mathbf{z}_i^j \sim \mathcal{N}(\mu_i^j, \Sigma_i^j)$ are obtained by transforming the Gaussian state estimates from the tracker into the local frame of person j. Skipping time indices t, the mean is then computed as

$$\mu_i^j \;=\; \ominus H\mathbf{x}_i \oplus H\mathbf{x}_j, \tag{7.5}$$

a so called tail-to-tail relationship with $H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$. The covariance Σ_i^j is obtained by first-order

[40] In a group of two persons i and j two geometric relations are learned: $\Psi_t^{i,j}$ represents the geometric relation from person i in the local reference frame of j while $\Psi_t^{j,i}$ represents the position of person j in the local reference frame of i. They are not identical. For example, if i is walking in front of j, j is walking behind i.



(a) Particle prediction using constant velocity motion model.

(b) Particle prediction using curvilinear motion model of Best and Norton [1997].

(c) Particle prediction using curvilinear model constrained on learned geometric relations.

Figure 7.4: Comparison of different motion models. 3D models are drown at the person mean position estimates. Colored circles, dots, and lines show position uncertainties, motion particles, and trajectories, respectively. Using constant motion (a) the person is predicted straight. The curvilinear motion model (b) accounts for turns and velocity changes. Restricting the model on the learned spatial relations (c) the learned spatial organization of the group is maintained.

error propagation of the two state covariances Σ_i and Σ_j through the frame transform using the tail-to-tail Jacobians as derived in Smith et al. [1990]. The proposal distribution $p(\widehat{\Psi}_t^{i,j} \mid \Psi_{t-1}^{i,j})$ is chosen to be a Brownian motion model shown in Figure 7.3c as this model makes the least commitment for predicting the evolution of the relation. Finally, for the filter initialization priors, the values for $\Psi_0^{i,j}$ are taken from Moussaïd et al., learned from large-scale observations. Prior information is provided by two 1-dimensional Gaussian distributions of relative angle and relative distance, respectively. Although they can not be easily combined into a 2-dimensional distribution in Cartesian coordinates the particle based approach allows their integration into the proposed method. A visualization of the prior knowledge is presented in Figure 7.3b.

The three terms, the proposal $p(\widehat{\Psi}_t^{i,j} \mid \Psi_{t-1}^{i,j})$, the (sampled) Gaussian observation likelihood $p(\mathbf{z}_i^j \mid \widehat{\Psi}_t^{i,j})$, and the priors $p(\Psi_0^{i,j})$ are then used in a particle filter with importance re-sampling to estimate spatial relations. A complete initialization, prediction, and update cycle in a group of three persons is illustrated in Figure 7.3.

In the general case of arbitrary group sizes with people sharing multiple social relations the set of geometric relations of person i to all his or her neighbors $j_1, \ldots, j_{N_i}$ is defined as

$$\Psi_t^i = \left\{ \Psi_t^{i,j_1}, \ldots, \Psi_t^{i,j_{N_i}} \right\}. \tag{7.6}$$

Details on the integration of all geometric relations into the motion estimate of a person are presented in the next section.

7.5 Motion Prediction using Geometric Relations

In this section, a model for short-term motion predictions of people moving in groups is proposed. Based on prior knowledge of group behavior analyzed in empirical studies by McPhail and Wohlstein, Moussaïd et al., and Yücel et al. and showing that people in groups largely maintain their spatial organization, the on-line learned geometric relations allow to predict occluded group members over their visible neighbors. Therefore, the approach of Mallick et al. [2011] is extended to multiple interaction partners and motion prediction in 2D.

Formally, with $\Psi_{t-1} = \{\Psi_{t-1}^{\mathbf{x},j}\}_{j=1}^{N}$ being the set of known geometric relations of track $\mathbf{x}$ to its N neighboring group member tracks, the motion model $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \Psi_{t-1})$ becomes conditioned on

both, the previous track state $\mathbf{x}_{t-1}$ and the set Ψ_{t-1} learned from previous observations up to time $t-1$. The model describes a general density that follows the shape and topology of the spatial prior presented in Moussaïd et al. and Yücel et al. and the learned geometric relations introduced in section 7.4, poorly described by a parametric distribution such as a Gaussian. According to the particle-based representation of geometric relations, the target distribution is therefore represented with a set of weighted samples

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \Psi_{t-1}) \;\simeq\; \sum_i w_t^{(i)} \delta_{\mathbf{x}_t^{(i)}}(\mathbf{x}_t) \tag{7.7}$$

where $\delta_{\mathbf{x}_t^{(i)}}(\mathbf{x}_t)$ is the impulse function centered in $\mathbf{x}_t^{(i)}$.

Sampling directly from the distribution $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \Psi_{t-1})$ is intractable in practice which is why a Monte Carlo approach is taken. Samples are first drawn from a proposal distribution π and then evaluated according to the mismatch between the target distribution τ and the proposal distribution. In the concrete case, the distribution is approximated by the following factorization

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \Psi_{t-1}) \;\simeq\; p(\mathbf{x}_t \mid \mathbf{x}_{t-1})\; p(\mathbf{x}_t \mid \Psi_{t-1}). \tag{7.8}$$

The natural choice to use a motion model $p(\mathbf{x}_t^{(i)} \mid \mathbf{x}_{t-1})$ as proposal distribution is adopted and the samples are evaluated according to

$$w_t^{(i)} \;=\; \frac{p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \Psi_t)}{p(\mathbf{x}_t^{(i)} \mid \mathbf{x}_{t-1})} \;=\; p(\mathbf{x}_t^{(i)} \mid \Psi_{t-1}). \tag{7.9}$$

In other words, samples are first spread out into the state space following the motion model $p(\mathbf{x}_t^{(i)} \mid \mathbf{x}_{t-1})$ and then weighted according to the set of geometric relations Ψ_{t-1}.

For $p(\mathbf{x}_t^{(i)} \mid \mathbf{x}_{t-1})$, the curvilinear model by Best and Norton [1997] is taken to account for accelerations needed to maintain the geometric relation during turns. This motion model is simple yet one of the most sophisticated target maneuver models in 2D. In contrast to the constant velocity motion model it accounts for both, (cross-track) normal and (along-track) tangential target accelerations needed to properly maintain the spatial intra-group relations during direction changes of the entire group (see Figure 7.4 and Figure 7.5). A comparison of both models is presented in Figure 7.4a and Figure 7.4b. Let $\mathbf{x}_t^{(i)} = (\; x_t \;\; y_t \;\; \dot{x}_t \;\; \dot{y}_t \;)^T$ be the state of particle i, $\vec{a}_t = (\; a_{tan} \;\; a_{nor} \;)^T$ the vector of tangential and normal accelerations, and A_t the transition matrix of the constant velocity model, then the particle states evolve according to

$$\mathbf{x}_t^{(i)} \;=\; A_t\, \mathbf{x}_{t-1}^{(i)} + G_t(\vec{a}_t^{(i)} + q_t) \tag{7.10}$$

with q_t being zero-mean Gaussian noise with covariance matrix Q_t. The matrices A_t and Q_t are introduced in Eq. 4.12. The details on the 4×2 forcing matrix G_t can be found in Best and Norton [1997]. An illustration of the model is provided in Figure 6.4.

To evaluate the likelihood $p(\mathbf{x}_t^{(i)} \mid \Psi_{t-1})$ of particle i the N geometric relations of Ψ_{t-1} are considered as equally important components of a mixture model having equal mixture weights, thus

$$p(\mathbf{x}_t^{(i)} \mid \Psi_{t-1}) \;=\; \frac{1}{N} \sum_{j=1}^{N} p(\mathbf{x}_t \mid \Psi_{t-1}^{\mathbf{x},j}), \tag{7.11}$$

with $p(\mathbf{x}_t \mid \Psi_{t-1}^{\mathbf{x},j})$ being the pairwise geometric relation explained in section 7.4. A visualization of the constrained curvilinear motion model is shown in Figure 7.4.

During tracking, visible group tracks are predicted using a constant velocity motion model and, after the Kalman update, are used to predict tracks of the same group that have been declared

as occluded by the MHT. The occluded tracks are predicted by spreading their particles according to Eq. 7.10. To evaluate their weights, the geometric relations need to be transformed into the reference frame of the tracker. The transformations are based on the positions predicted in the first step forcing the particles to follow the motions of the adjacent group members. If all group members were occluded, motion prediction would fall back onto the constant velocity model.

7.6 Integration into the Multi-Hypothesis Tracker

This section shows, how the proposed methods of social relation learning and geometrically constraint motion prediction are integrated into the Multi-Hypothesis Tracking (MHT) framework to describe dynamic group formation processes. For the sake of brevity, only the relevant parts of the MHT theory are presented. Refer to Chapter 3 and the work of Arras et al. [2008] for more details.

Summarizing, the MHT algorithm hypothesizes about the state of the world by considering all statistically feasible assignments between observations and tracks and all possible interpretations of observations as false alarms or new track and tracks as matched, occluded or deleted. A hypothesis Ω_l^t is one possible set of assignments and interpretations at time t. Let $\mathcal{Z}(t) = \{\mathbf{z}_i(t)\}_{i=1}^{M_t}$ be the set of M_t observations which in this case is the set of detected people[41] and let $\psi_l(t)$ be the set of assignments which associates the tracks $\mathbf{x}_j(t-1)$ from the previous time step to the observations in $\mathcal{Z}(t)$. Further, let $\mathcal{Z}^t = \{\mathcal{Z}(0), \ldots, \mathcal{Z}(t)\}$ be the set of all observations up to time t. Starting from a hypothesis of the previous time step, called a parent hypothesis $\Omega_{p(l)}^{t-1}$, and a new set $\mathcal{Z}(t)$, there are many possible assignment sets $\psi_l(t)$, each giving birth to a child hypothesis that branches off the parent. This makes up an exponentially growing hypothesis tree. For a real-time implementation, the growing tree needs to be pruned. An example of a pruned tree is given in Figure 1. To guide the pruning, each hypothesis receives a probability, recursively calculated as the product of a normalizer η, a measurement likelihood, an assignment set probability and the parent hypothesis probability, thus

$$p(\Omega_l^t \mid \mathcal{Z}^t) \;=\; \eta \;\; p(\mathcal{Z}(t) \mid \psi_l(t)) \;\; p(\psi_l(t) \mid \Omega_{p(l)}^{t-1}, \mathcal{Z}^{t-1}) \;\; p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}), \tag{7.12}$$

where the last term is known from the previous iteration. More details on the measurement likelihood and the assignment set probability can be found in subsection 3.3.1 and subsection 3.3.2, respectively.

There exist two alternatives to integrate social relations into the MHT framework that are presented in the next subsections and compared in the experiments. The first approach – described in subsection 7.6.1 – follows the idea of Arras et al. [2008] who track pairs of human legs that frequently occluding each other using explicitly modeled adaptive occlusion probabilities. The second approach – introduced in subsection 7.6.2 – models group affiliations explicitly by an *intermediate tree level at each time step*, on which models spring off from parent hypotheses. As shown in Lau et al. [2009] each model branch has its own data association tree, conditioned on that model. Finally, the integration of the on-line learned geometric relations is introduced in subsection 7.6.3.

7.6.1 Bayes Filtered Social Relations and Adaptive Occlusion Probabilities

When perceiving the scene from a first-person perspective, occlusions occur particularly often for people in groups. Thus, person tracks for which the system predicts a high social relation probability will have a higher occlusion probability. Formally, this can be implemented using a simple extension

[41] The set of observations $\mathcal{Z}(t)$ also contains misdetections from clutter that need to be declared as false alarms by the tracking framework.

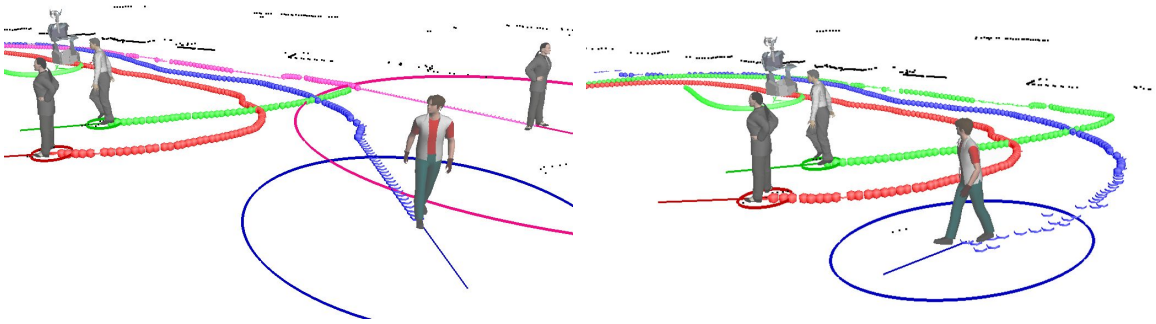


Figure 7.5: Tracking a group of three people using the mobile robot. 3D models are drawn at the person mean position estimates. Colored circles and dots show position uncertainties and trajectories, respectively. *Left:* The fixed parameter MHT using constant velocity motion model predicts the persons according their last motion causing track losses. *Right:* Using socio-spatial relations the inter-group organization is maintained. No track losses occur.

of the MHT initially developed for leg tracks[42] in Arras et al. [2008]. As described in section 3.5 the extension allows the MHT to not only reason about the interpretation of tracks to be detected or deleted but also to be occluded. This implies a generalization to an arbitrary number of track interpretation labels and the modeling of their numbers in an assignment set by a multinomial distribution. With occlusion being a label on its own, Eq. 7.12 is extended to adapt the occlusion probability of individual tracks dynamically. In the following the assignment set dependent numbers of observation events are defined as N_{fal} (number of false alarms) and N_{new} (new tracks), respectively. The numbers of track events are denoted as N_{det} (detected tracks), N_{occ} (occluded tracks), and N_{del} (deleted tracks), respectively. With $N_{det} + N_{fal} + N_{new} = M_t$ and $N_{det} + N_{occ} + N_{del} = N_{t-1}$ Eq. 7.12 yields

$$p(\Omega_l^t \mid \mathcal{Z}^t) = \eta \; \lambda_{fal}^{N_{fal}} \; \lambda_{new}^{N_{new}} \; \prod_{i=1}^{M_t} \left(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i}\right) \; p_{det}^{N_{det}} \; p_{occ}^{N_{occ}} \; p_{del}^{Ndel} \; p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}), \tag{7.13}$$

where τ_i indicates matched observations and $\mathcal{N}(\mathbf{z}_i(t)) := \mathcal{N}(\mathbf{z}_i(t);\, \hat{\mathbf{z}}_j(t),\, S_{i,j}(t))$ denotes the measurement likelihood of a single matched observation $\mathbf{z}_i(t)$ associated to an existing track $\mathbf{x}_j(t)$ assumed to be a Gaussian pdf centered around the measurement prediction $\hat{\mathbf{z}}_j(t)$ with innovation covariance matrix $S_{i,j}(t)$. Furthermore, the numbers of new tracks N_{new} and false alarms N_{fal} are assumed to follow Poisson distributions with constant expected numbers of events λ_{new} and λ_{fal} in the observation volume V, respectively. The occurrence of N_{det}, N_{occ}, and N_{del} track detection, occlusion, and deletion events is modeled jointly using a multinomial distribution with constant parameters p_{det}, p_{occ}, and p_{del}, respectively.

Employing the knowledge of Bayes filtered group affiliations an additional indicator is introduced marking people in groups[43] with „G“ (group) and those who walk solely with „F“ (free). Applying the indicator to all track events yields N_{det}^F detected tracks that walk alone and N_{det}^G detected tracks in groups. Furthermore, $N_{occ}^{F/G}$ and $N_{del}^{F/G}$ are the number of occluded and deleted tracks that

[42] Tracking groups of people can be compared to the problem of tracking pairs of human legs. Both – two legs and the people in a group – frequently occlude each other.

[43] The groups are detected using the graph cut algorithm explained in subsection 7.3.3. Groups of size 1 are filtered out and their members are marked with F.

are free or marked as group members, respectively. While the non-adaptive MHT has parameters for track detection, occlusion and deletion events, the MHT with adaptive occlusion probabilities requires to learn those probabilities for tracks in groups separately, thus the constant parameters p_{det}, p_{occ}, and p_{del} are replaced by $p_{det|F}$, $p_{det|G}$, $p_{occ|F}$, $p_{occ|G}$, $p_{del|F}$, and $p_{del|G}$, denoting the probability of matching, occlusion, and deletion given a free track (F) or a group member (G), respectively. Both sets, best learned from large-scale datasets, are subject to the multinomial constraint $p_{det|F} + p_{occ|F} + p_{del|F} = 1$ and $p_{det|G} + p_{occ|G} + p_{del|G} = 1$.

With the numbers of detected, occluded, and deleted tracks of people walking alone or in groups summing up to the total number of tracks, e.g. $N^F_{det} + N^G_{det} = N_{det}$, it can be shown that Eq. 7.12 yields

$$p(\Omega^t_l \mid \mathcal{Z}^t) = \eta\, \lambda_{fal}^{N_{fal}}\ \lambda_{new}^{N_{new}}\ \prod_{i=1}^{M_t} \left(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i}\right) \tag{7.14}$$
$$p_{det|F}^{N^F_{det}}\ p_{occ|F}^{N^F_{occ}}\ p_{del|F}^{N^F_{del}}\ p_{det|G}^{N^G_{det}}\ p_{occ|G}^{N^G_{occ}}\ p_{del|G}^{N^G_{del}}\ p(\Omega^{t-1}_{p(l)} \mid \mathcal{Z}^{t-1}).$$

With two different occlusion probabilities $p_{occ|F}^{N^F_{occ}}$ and $p_{occ|G}^{N^G_{occ}}$, this expression is a spezial case of Eq. 3.29 in which general track-specific occlusion probabilities are introduced.

Integrating the knowledge of the Bayes filtered social relations using Eq. 7.14 is rather simple. The group detector explained in subsection 7.3.3 serves to mark the tracks in groups[44]. Hence, the detection, occlusion, and deletion probabilities of those tracks are adapted accordingly. Social relation learning and adapted probabilities influence each other since latter lead to more robust tracking with less track losses – especially during lengthy occlusion events – and longer tracks aid to find stable social relations and groups.

7.6.2 Tracking Social Relations using Explicit Group Models

An important novelty of the proposed approach is the ability to hypothesize about the most likely partitions of the social network graph and to track them over time in a multi-model multi-hypothesis framework. Instead of smoothing the social relations probabilities using a Bayes filter this approach integrates social information into the MHT explicitly. The approach is inspired by Lau et al. [2009, 2010] who track groups of people using a recursive multi-hypothesis model selection framework. Introducing explicit group models hypotheses describe both, the partitioning of tracks into groups (group models) and the associations of observations to tracks (assignments). In contrast to their work, where the states of multiple tracks in a groups are merged to a common group state the approach presented in this section maintains the individual states and identities of the tracked people.

Group Modeling

A specific partitioning of the social network graph $\mathcal{G}$ into subgraphs $\mathcal{G}_i$ encoding the groups of people is called a group model $\mathcal{M}$, that is formally defined as

$$\mathcal{M} \;=\; \{\mathcal{G}_i\}_{i=1}^{N_\mathcal{G}}\,, \tag{7.15}$$

where $N_\mathcal{G}$ is the number of groups in the model and $\mathcal{G}_i = \{U_i, E_i\}$ are the subgraphs with $U_i \subset U$ and $E_i \subset E\ \forall i$ each describing a group of people. Further, each tracker person belongs to exactly one group, people without relations to others form groups of size 1.

[44] The group detection is performed on each hypothesis independently. Thus, the tracks of the same person can be marked as free in one hypothesis and at the same time as group member in another one.

Model Generation

Groups are initialized when the tracker signals a new track event, e.g. when a person enters the sensor field of view. Then, a new group of size 1 is created. Social relation computation with this group is delayed until the track state have reached steady state in the filter, typically after four or five steps. Once the social grouping detection is stable, each group can, at any point in time, be continued, split up into an unknown number of new groups, merged with an unknown number of other groups. Since the possible number of model transitions is large this space is bounded by the assumption that split and merge events are binary operations as defined in Lau et al. [2010]. In each step, a group can only split up into two child groups and at most two groups can merge into a larger group. It is also assumed that a group can not be involved into a split and merge operation at the same time. In detail, let $\mathcal{M}(t-1)$ be a group model at time $t-1$, the possible transitions for each group $\mathcal{G}_i(t-1), \mathcal{G}_j(t-1) \in \mathcal{M}(t-1)$ in one frame are:

(1) group $\mathcal{G}_i(t-1)$ continues without changes[45] into group $\mathcal{G}_k(t)$, or

(2) group $\mathcal{G}_i(t-1)$ splits into two groups $\mathcal{G}_k(t)$ and $\mathcal{G}_l(t)$, or vice versa,

(3) two groups $\mathcal{G}_i(t-1)$ and $\mathcal{G}_j(t-1)$ merge into one group $\mathcal{G}_k(t)$.

A group is terminated if all its members are declared as obsolete by the tracker. The binary operation assumption is not a sensible limitation because, for example, an instantaneous breakup of a group into three subgroups would be correctly reflected by the tracker after only two cycles.

As the number of possible group model $\mathcal{M}(t)$ grows exponentially over time their generation is limited by a set of simple rules and guided by the probabilities of the existing social relations $\mathcal{S}(t)$, thereby implementing a data-driven aspect into the model generation step. The rules include, that an existing group $\mathcal{G}_i(t-1)$ can only split if none of the tracks in the two child groups $\mathcal{G}_k(t)$ and $\mathcal{G}_l(t)$ share a social relation above probability threshold θ. Thus, a split event occurs with a probability that scales with the strongest remaining social relation. Formally, let U_k be the nodes (people) in group $\mathcal{G}_k(t)$ and U_l be the nodes in group $\mathcal{G}_l(t)$, respectively. It must be satisfied that each edge (social relation) $\varepsilon_{k,l} \in E_i$ in the social relation graph of group $\mathcal{G}_i(t-1)$ connecting a node in U_k to a node in U_l has a probability lower than the threshold θ defined in subsection 7.3.3. The edge with the highest probability (but still lower than θ) is defined as ε_i^-. Similarly, two groups $\mathcal{G}_i(t-1)$ and $\mathcal{G}_j(t-1)$ are only allowed to merge into $\mathcal{G}_k(t)$ if there is at least one social relation between the members of those groups greater than θ. Thus, $\exists\ i,j \quad p(\varepsilon_{i,j}) > \theta$ and merge events depend on the highest probability of an across-group relation defined as $\varepsilon_{i,j}^+$.

Model Probability

The probability of a group model $\mathcal{M}(t)$ at time t, follows from the probabilities of the split (*spl*), merge (*mer*), and continuation (*con*) events of its groups assumed to have constant prior probabilities p_{con}, p_{spl}, and p_{mer}, respectively. Furthermore, the probabilities of the social relations ε_i^- and $\varepsilon_{i,j}^+$ forcing split or merge events are integrated. Given the recursiveness of the problem, it conditionally depends on the model of the previous time step $\mathcal{M}(t-1)$ and the current social relations $\mathcal{S}(t)$ both encoded in parent hypothesis Ω^{t-1} at time $t-1$, thus

$$p(\mathcal{M}(t) \mid \Omega^{t-1}) \;=\; p_{con}^{N_{con}} \prod_{\mathcal{G}_i} \left(p_{spl}\,(1 - p(\varepsilon_i^-))\right)^{\sigma_i} \prod_{\mathcal{G}_i, \mathcal{G}_j} \left(p_{mer}\, p(\varepsilon_{i,j}^+)\right)^{\mu_{i,j}}, \tag{7.16}$$

[45] The only *change* allowed in the continued groups is that obsolete tracks marked for deletion are removed. This case is not interpreted as a split event.

where N_{con} is the number of continued groups, $\mathcal{G}_i, \mathcal{G}_j \in \mathcal{M}(t-1)$, and σ_i and $\mu_{i,j}$ are indicator variables set to 1 if $\mathcal{G}_i$ is split or $\mathcal{G}_i, \mathcal{G}_j$ merge, respectively, and 0 otherwise. The model probability can be conditioned on a parent hypothesis Ω^{t-1} since group affiliations are encoded in each hypotheses explicitly and the social relation graph $\mathcal{S}(t)$ is calculated based in the states of the tracks in Ω^{t-1}, hence $p(\mathcal{M}(t) \mid \mathcal{M}(t-1), \mathcal{S}(t)) = p(\mathcal{M}(t) \mid \Omega^{t-1})$.

Multi-Model Multi-Hypotheses Tracking

To integrate explicit group models and to hypothesizes over data associations and group models an *intermediate tree level at each time step*, on which models spring off from parent hypotheses is introduced. Formally, this adds a model probability term to Eq. 7.12 and introduces the group model as a conditioning variable (see derivation in Lau et al. [2010]). The hypothesis probability $p(\Omega_l^t \mid \mathcal{Z}^t)$ formerly calculated based on all sensor readings $\mathcal{Z}^t$, its parent hypothesis $\Omega_{p(l)}^{t-1}$, and a track to observation assignment set $\psi_l(t)$, is now also conditioned on a group model, thus

$$p(\Omega_l^t \mid \mathcal{Z}^t) \;=\; p(\psi_l(t),\; \mathcal{M}(t),\; \Omega_{p(l)}^{t-1} \mid \mathcal{Z}(t),\; \mathcal{Z}^{t-1}) \tag{7.17}$$

Using Bayes' rule, Eq. 7.17 yields

$$\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t) \;=\; & \eta\; p(\mathcal{Z}(t) \mid \psi_l(t),\; \mathcal{M}(t),\; \Omega_{p(l)}^{t-1},\; \mathcal{Z}^{t-1}) \\ & p(\psi_l(t) \mid \mathcal{M}(t),\; \Omega_{p(l)}^{t-1})\;\; p(\mathcal{M}(t) \mid \Omega_{p(l)}^{t-1})\;\; p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}), \end{aligned} \tag{7.18}$$

where η is a normalizer, followed by the measurement likelihood, the assignment set probability and the model probability defined in Eq. 7.16. The rightmost term represents the probability of the parent hypothesis known from the previous iteration. The measurement likelihood and assignment set probability are discussed in more detail in subsection 3.3.1 and subsection 3.3.2, respectively.

Based on explicitly modeled group affiliations the indicator variable introduced in subsection 7.6.1 marks tracks in groups of at least two persons with „G“ and adapts their detection, occlusion, and deletion probability to $p_{det|G}$, $p_{occ|G}$, and $p_{del|G}$, respectively, where $p_{det|G} + p_{occ|G} + p_{del|G} = 1$. The people walking solely are marked with „F“ and tracked with the usual parameters $p_{det|F}$, $p_{occ|F}$, and $p_{del|F}$, conditioned on $p_{det|F} + p_{occ|F} + p_{del|F} = 1$. With N_{con} being the number of continued groups, $N_{det}^{F/G}$, $N_{occ}^{F/G}$, and $N_{del}^{F/G}$ the number of free and group tracks that are detected, occluded, or deleted, respectively, Eq. 7.18 yields

$$\begin{aligned} p(\Omega_l^t \mid \mathcal{Z}^t) \;=\; \eta\;\; & \lambda_{fal}^{N_{fal}}\;\; \lambda_{new}^{N_{new}}\;\; \prod_{i=1}^{M_t} \left(\mathcal{N}(\mathbf{z}_i(t))^{\tau_i}\right) \\ & p_{con}^{N_{con}}\;\; \prod_{\mathcal{G}_i} \left(p_{spl}\, p(\varepsilon_i^-)\right)^{\sigma_i}\;\; \prod_{\mathcal{G}_i, \mathcal{G}_j} \left(p_{mer}\, p(\varepsilon_{i,j}^+)\right)^{\mu_{i,j}} \\ & p_{det|F}^{N_{det}^F}\;\; p_{occ|F}^{N_{occ}^F}\;\; p_{del|F}^{N_{del}^F}\;\; p_{det|G}^{N_{det}^G}\;\; p_{occ|G}^{N_{occ}^G}\;\; p_{del|G}^{N_{del}^G}\;\; p(\Omega_{p(l)}^{t-1} \mid \mathcal{Z}^{t-1}). \end{aligned} \tag{7.19}$$

The number of hypotheses grows boundlessly thus pruning of the hypotheses tree is essential. The k-best pruning strategy proposed by Murty [1968] and described in section 3.8 is employed to limit the maximum number of hypotheses generated at every step to k. The hypotheses probability $p(\Omega_l^t \mid \mathcal{Z}^t)$ is used to guide the pruning process. As multiple group models $\mathcal{M}_1(t), \ldots, \mathcal{M}_N(t)$ emerging from the same parent hypothesis may achieve a similar (high) probability value the hypotheses tree quickly looses diversity since the tree concentrates on few probable branches. To avoid this effect a second pruning strategy is implemented to limit the number of model arising from a common parent to $l \leq k$. Similar to k-best pruning this strategy is guided by the probability of the groups models $p(\mathcal{M}_i(t))$.

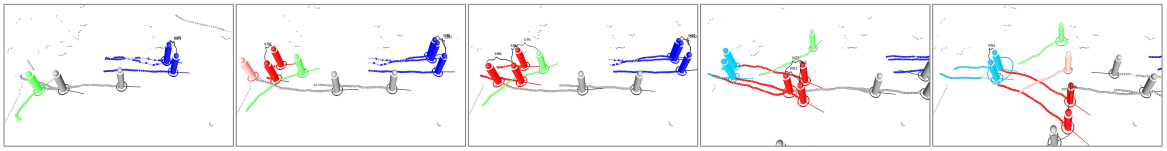


Figure 7.6: Tracking sequence in the city center data set. The cylinders and dots show positions and trajectories. People in the same groups have identical color, those in gray and green share no social relations. The green-colored track (correctly) does not merge with the red group (max. social relation probability 0.396 in frame 3). The person in light red merges with the red group in frame 2 and splits up in frame 5.

7.6.3 Integration of Geometric Relations

The Geometric relations are now used for better prediction of human motion in groups. Motion prediction with geometric relations, as introduced in section 7.5, is applied analog to Mallick et al. [2011] where a group of cars is predicted from the front to the rear constraining the motion on a safe driving distance. Given a group of people $\mathcal{G}(t-1)$ known from the previous time step. First, the positions $\widehat{\mathbf{x}}_j(t)$ of all visible group members $\mathbf{x}_j(t-1) \in \mathcal{G}(t-1)$ are predicted using constant motion assumption

$$\widehat{\mathbf{x}}_j(t) = A_t\,\mathbf{x}_j(t-1), \qquad \widehat{\Sigma}_j(t) = A_t \Sigma_j(t-1) A_t^T + Q_t, \tag{7.20}$$

with A_t being the state transition matrix and Q_t the process noise Matrix encoding the acceleration capabilities of walking pedestrians. Both matrices are introduced in Eq. 4.12. Subsequently, each occluded track $\mathbf{x}_i(t-1)$ is predicted by spreading its particles according to Eq. 7.10. The evaluation of the particle weights using Eq. 7.11 requires to transform the position invariant spacial relations $\Psi_t^{\mathbf{x},j}$ into the global reference frame. The transformations are based on the positions predicted in the first step forcing the particles to follow the motions of the adjacent group members. Note, the geometrical relations with group members that are occluded are not taken into account. If all group members are occluded the motion prediction falls back to the constant velocity motion model.

7.7 Experiments

In this section the proposed tracking system is evaluated and the contribution of all extensions to the tracking performance is analyzed. The experiments are carried out on two exemplary data sets collected indoors and outdoors with the mobile robot Daryl and two large, unscripted outdoor data sets collected in a city center of Freiburg and a main station environment during a regular work day. The sensor is always a SICK LMS 291 laser range finder mounted at a height of $\sim$0.85 meter and with an angular resolution of 0.5 degree. The large outdoor data sets of 55,475 and 33,204 frames (25 and 15 minutes, respectively) recorded at fairly busy city locations, contain data on individuals, couples, groups of people, bicycles, cars, wheelchairs, skaters and person-shaped static obstacles that all undergo countless occlusions (see Figure 7.1 for an example frame). The data have been manually annotated to determine the detection, data association, and social grouping ground truth. Criteria for the social grouping annotations were people's trajectories, behaviors, and appearances (camera data were available). In detail, the city center data set consists of 10,000 frames with 190 person tracks including 31 groups. The main station set contains 6,000 frames with 168 person tracks and 25 groups.

As people detector the place-dependent cascade of specialized boosted features classifiers presented in Chapter 2 and based on the approach of Arras et al. [2007] is applied. Shortly, a set of geometrical

Approach	TP	FP	TN	FN	PR	RE	ACC
training set	6342	879	12703	520	0.88	0.92	0.93
single frame detection	4598	733	12849	2264	0.86	0.67	0.85
Bayes filtered detection	5701	1256	12326	1161	0.81	0.83	0.88

Table 7.1: Results of the social relation detection between pairs of people on the training set (1^{st} row), the test set using only single frames without Bayesian smoothing (2^{nd} row), and with the Bayes filter (3^{rd} row).

and statistical features is computed for each potential object and classified by a cascade of strong classifiers trained for specific range intervals. Social relations are detected using a linear SVM trained on the relative motion features defined in McPhail and Wohlstein [1982] and shown in Figure 7.2. Both classifiers have been trained on a separate training set. An experimental evaluation of the accuracy of the people detector is presented in section 2.5.

The MHT parameters for detections, occlusions, and deletions and the fixed rates for false alarms and new tracks have been learned from a training data set with 95 tracks over 28,242 frames. In detail, $p_{det|F} = 0.7$, $p_{occ|F} = 0.27$, $p_{del|F} = 0.03$, $\lambda_{new} = 0.0003$, and $\lambda_{fal} = 0.005$, respectively. The adapted parameter for people in groups are $p_{det|G} = 0.6$, $p_{occ|G} = 0.39$, $p_{del|G} = 0.01$, respectively. The baseline tracker is set up with the adapted parameters to allow longer occlusion events. Geometric relations are learned and predicted using 200 particles per track. The maximum number of MHT hypothesis is $k = 100$ and the maximum number of model branches per hypothesis is $l = 10$. The system is fully integrated on the mobile robot Daryl.

7.7.1 Detecting Social Relations and Groups

As the performance of the tracking system strongly depends on the ability to detect groups the accuracy of the social relation detection and the impact of the Bayesian filtering is evaluated first. The results are presented in Table 7.1 and discussed hereafter.

On the training set, the detection accuracy (ACC) of the SVM classifier is 93% with only 879 false positives (FP) and 520 false negatives (FN), respectively. While tracking on the test set, this decreases to 85% accuracy mostly due to missed social relations (2264 FN) during the track initialization phase when the orientation and velocity state estimates are not yet in steady state. Bayes filtering the social relation probabilities over time as depicted in subsection 7.3.2 improves the number of misses by 50% but comes at the expense of delayed responses, e.g. when people leave a group. This causes the number of false positives to increase by almost a factor of two. However, the overall detection accuracy increases to 88%.

This accuracy is sufficient for the current purposes, however, inferring relations only from motion features is clearly limited and future work will focus on recognizing more attributes of people as cues for social relations. See Table 7.1 for all numbers including true positives (TP), true negatives (TN), precision (PR), recall (RE), and accuracy (ACC).

7.7.2 Tracking Social Relations

Evaluating the impact of tracking the social grouping hypotheses, it shows that the proposed approach is able to resolve the trade off between lower numbers of false negatives and delayed response times. This is demonstrated in the indoor experiment in Figure 7.7 where three people meet, form a group during interaction and split up again. The multi-model hypothesis tracker is able to reflect

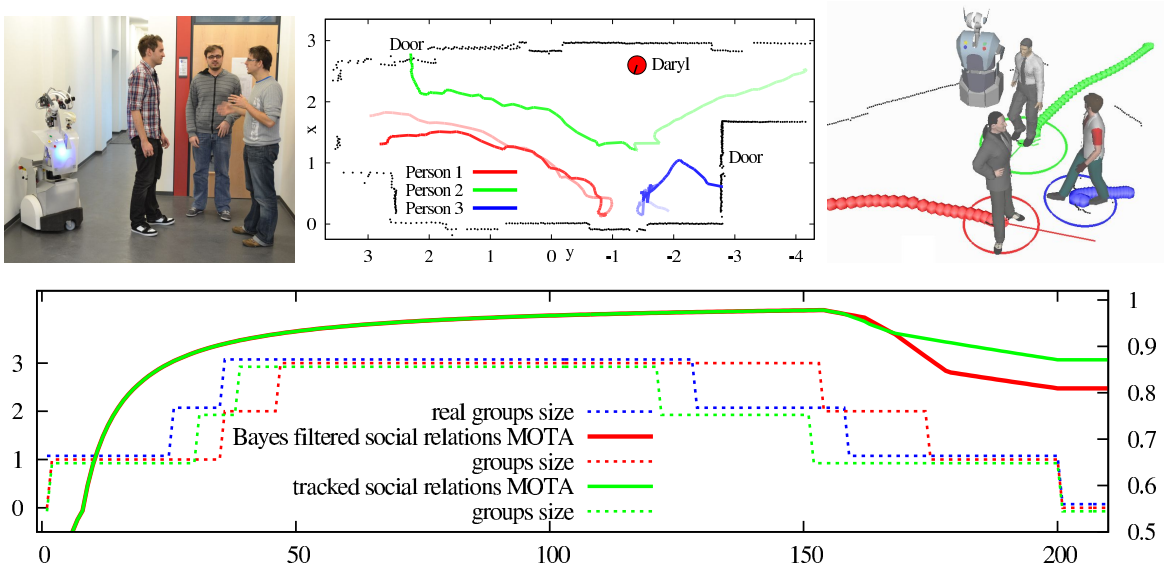
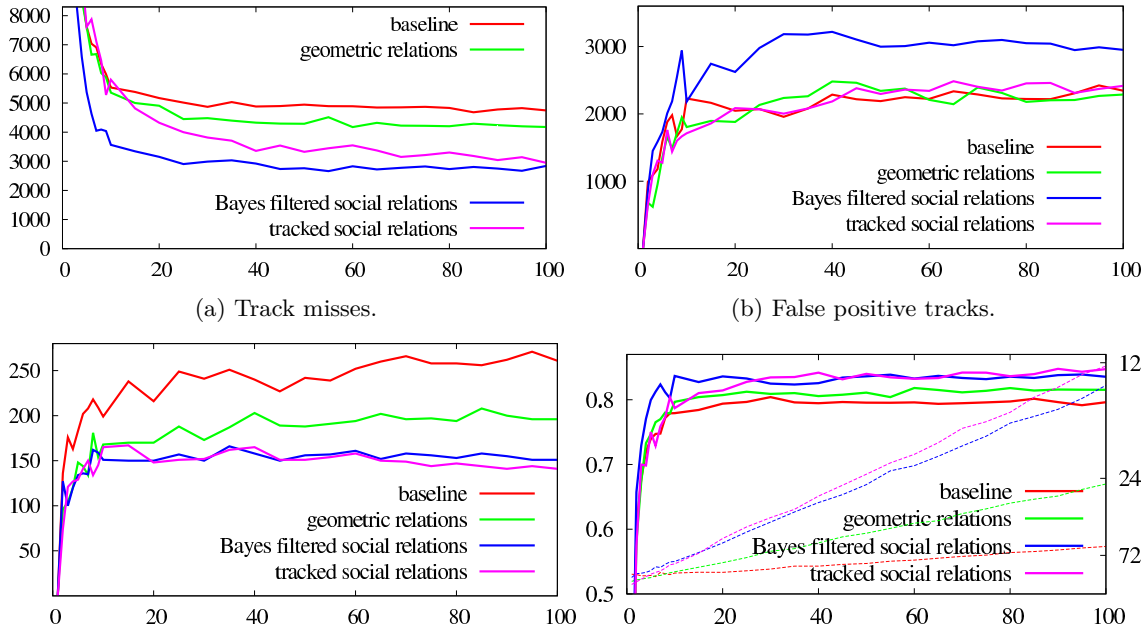


Figure 7.7: Comparison of the filtered per-frame approach vs. the tracking approach. Three people meet, interact, and split up. The top middle image shows their trajectories. The bottom graph shows the group formation process and person-level tracking accuracy. The Bayes approach (red) overly delays group merge and split events and cause the tracking accuracy (solid) to drop. Tracked social relations (green) are clearly closer to ground truth (blue).

those group formation changes much faster than the Bayesian smoothing approach (see dotted lines in the bottom diagram). This is due to the ability of the multi-hypothesis approach to consider multiple model explanations at a time and backtrack to branches that have become more probable with more incoming information. The delayed responses of the Bayesian approach make that the group merge phase lags behind and that the persons are kept in one group overly long after the split up. The solid lines in the same diagram show the person-level tracking accuracy (MOTA) which is consistently high for the multi-hypothesis tracking approach versus a drop from 87% to 81% for the Bayesian filtering approach. The tracking accuracy is calculated as $1 - \frac{\#err}{\#evt}$, with $\#err$ being the number of tracking errors and $\#evt$ the number of tracking events. A second interesting insight is, that the correct partitioning of the people into one groups is also achieved earlier. This does not lead to a tracking improvement in this simple scenario but shows that the multiple models enable a better representation of the social relations between people. While in this experiment the improvement seems not dramatic, the faster response times are crucial for robots that reactively navigate among people. See below for more results on the large-scale data sets.

7.7.3 People Tracking using the Mobile Robot Daryl

This experiment evaluates the adaptive occlusion probabilities, the on-line estimated geometric relations, and their ability to predict group tracks over lengthy occlusion events. The baseline is the tracker without social and geometric relation information. During a sequence of six minutes three persons were instructed to walk and stand in the vicinity of the mobile robot while changing their spatial arrangement. The robot was also moving during the experiment. An example situation is shown in Figure 7.5 during a turn of the group. The baseline approach is unable to maintain the spatial organization of the group and looses track of one person. The total number of track losses during the experiment is 12. The proposed system with social and geometric relations maintains



(a) Track misses.

(b) False positive tracks.

(c) Identifier switches.

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

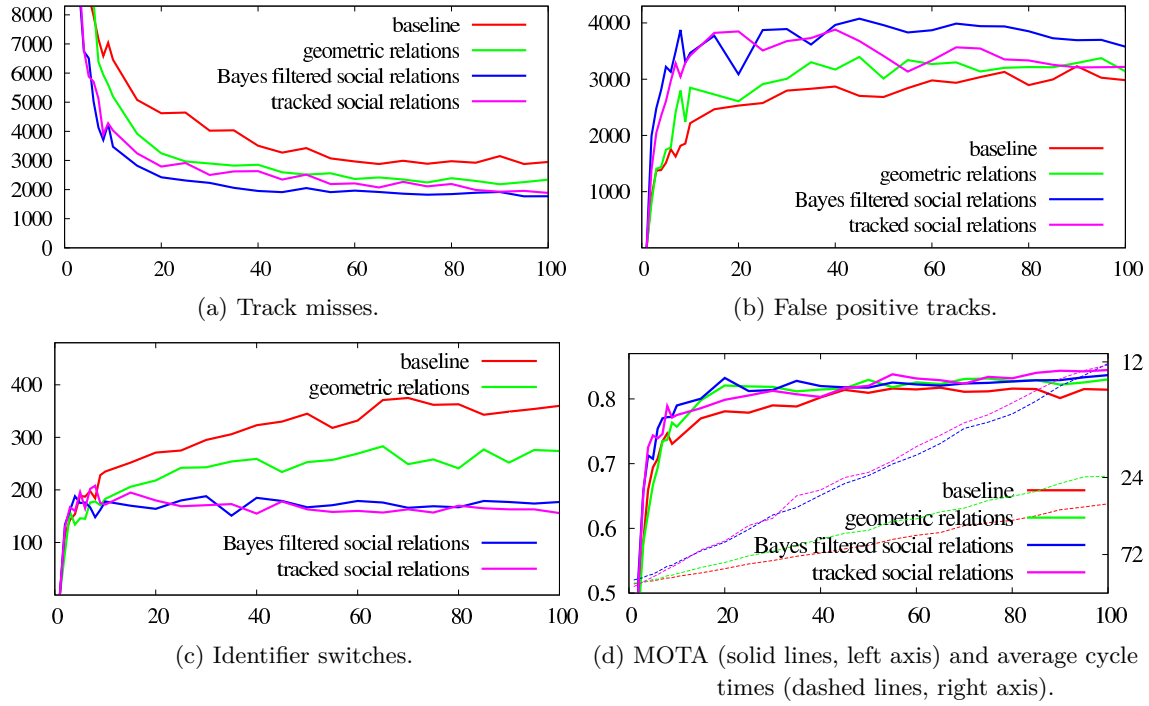
Figure 7.8: CLEAR MOT result of the Freiburg city center data set using socio-spatial relations. The baseline tracker (shown in red) uses no social and geometrical relations. Employing geometrical relations only (green) leads to improvements during occlusions and group maneuvers. The Bayesian filtering approach (blue) improves the numbers of track misses (a) and identifier switches (c) but causes a significant increase of false positives (b). The model approach (purple) resolves that problem due to faster response times to group formation changes and a proper incorporation of domain knowledge into occlusion handling.

the spatial organization and has no track losses. The particle filter is also able to quickly adapt the on-line learned geometric relations to changes of the spatial arrangement of the group.

7.7.4 People Tracking using Social and Geometric Information

Finally, the last experiments evaluate the multi-model hypothesis MHT on the two large-scale outdoor data sets (city center and main station) using the CLEAR MOT metrics introduced by Bernardin and Stiefelhagen [2008]. The metrics count three numbers with respect to ground truth: misses (missing person tracks that should exist at a ground truth position, FN), false positives (person tracks that should not exist, FP), and mismatches (track identifier switches, ID). From these numbers the tracking accuracy MOTA defined above is calculated. Note that due to the normalization by the total number of events, even large reductions of the errors may result in only small changes of the MOTA score.

The proposed approaches are compared to a regular MHT without the group model hypothesis extension. Furthermore, the contributions of the social and geometric relation information are studied separately. Additionally, the physically grounded occlusion model proposed in subsection 5.4.1 is integrated into the system to improve tracking in case people are occluded by static obstacles over a longer period of time. The tracking behavior with different numbers of hypotheses varying from $k = 1$ to 100 are discussed in more detail hereafter. The results using $k = 100$ hypotheses are



(a) Track misses.

(b) False positive tracks.

(c) Identifier switches.

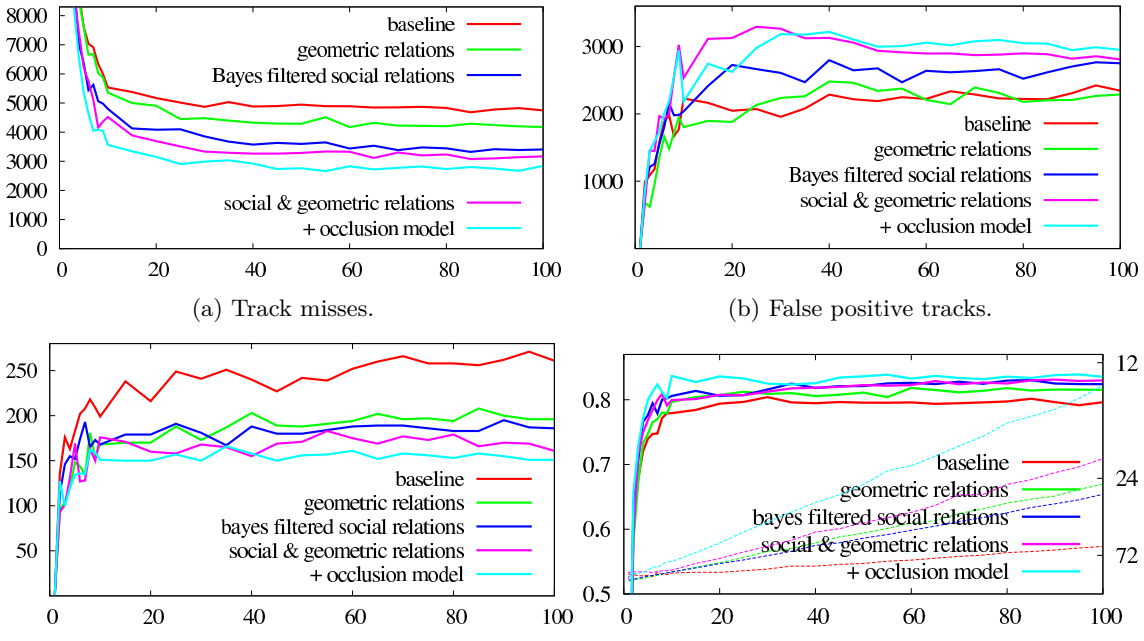
(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

Figure 7.9: CLEAR MOT analysis of the Freiburg main station data set using socio-spatial relations. Using the combination of all models (shown as purple line) and 100 hypotheses the number of identifier switches decrease by 56.7%. The approach can be applied in real time but due to a sample based occlusion model the average run-time is 12.1 Hz.

presented in Table 7.2.

Geometric Relations

Using on-line learned geometric intra-group relations for the joint motion prediction of people in groups has a strong positive impact on the number of track identifier switches (ID) decreasing by 24.9% and 23.9%, respectively. As shown in Figure 7.5, the geometric relations allow the tracker to properly track persons during occlusions and group maneuvers maintaining their correct identities. This property is particularly important when targets have identical appearance and no target-specific models can be learned to re-identify individuals. It is noteworthy that the approach finds a good trade-off between occlusion handling and an increase of false positive tracks. The effect on latter is neutral shown by changed numbers of false positives FP by $-2,4\%$ and $+5,2\%$. Naive methods handle occlusions simply by delaying the deletion of tracks but cause the number of wrong tracks (e.g. from false positive people detections) to persist longer in the system as well resulting in a stronger increase of FP. The impoved occlusion handling has also a positive effect on the number of false negatives FN increasing by $11,9\%$ and $20,6\%$. After occlusions refined motion predictions lead to higher matching likelihoods thus the number of track misses are reduced. The tracking accuracy (MOTA) increases by 1.9% and 1.6%, respectively. Furthermore, detailed evaluations on the Freiburg city center (Figure 7.8) and Freiburg main station (Figure 7.9) data sets revealed that a maximum number of 15 to 20 hypotheses is sufficient to obtain the improvements. Ergo, the results suggest that the approach of learning spatial group arrangements is a well suited method to deal with occlusions in this context.



(a) Track misses.

(b) False positive tracks.

(c) Identifier switches.

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

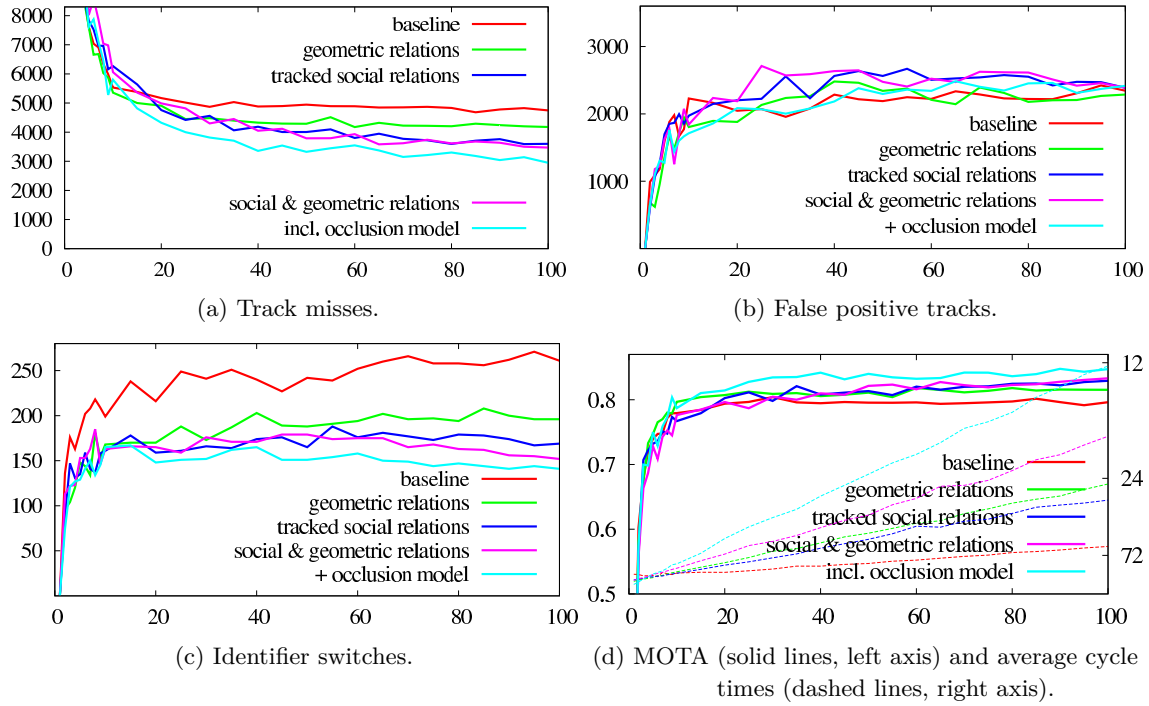
Figure 7.10: Detailed analysis of the results on the city center data set using Bayes filtered social relations. In combination with geometric relations (purple) and the occlusion model (cyan) the number of false positives (b) increases dramatically for low numbers of hypotheses.

Smoothed vs. Tracked Social Relations

The Bayesian filtering approach appears to have this very problem. It improves the number of false negatives by almost 30% on both data sets and the number of identifier switches by 28.7% (city center) and 34.7% (main station), respectively, but causes a significant increase in the number of false positive measures by 17.4% and 16.9%. A detailed analysis on the Freiburg city center data presented in Figure 7.10 shows that this effect already occurs with a low number of hypotheses ($k = 20$) and can not be diminished by increasing k. The method is too simple to find a good track management trade-off, one reason being the slow response to group formation changes as already shown in Figure 7.7. In contrast, the multi-model multi-hypothesis approach finds the so far best FP versus FN/ID trade-off. The numbers of false positives are only slightly increased by 1.8% and 5.9%, respectively. This is mainly due to the faster response times to group formation changes and a proper incorporation of domain knowledge into occlusion handling. As presented in Figure 7.11 this improvement is already achieved with a low number of hypotheses ($k = 30$) but can further be advanced by increasing k.

Combining Social and Geometric Information

Adding the constraint-based motion prediction model improves both approaches (Bayes filter and multi-model multi-hypothesis MHT) in the measures of FN and ID. The number of false positives (FP) remain almost stable. In this settings, the multi-model multi-hypothesis MHT yields the overall best results, expressed also by the highest MOTA scores. Notice that a key improvement over the baseline of -38% (Bayes filter) and up to -49% (group models) fewer track ID switches achieved. The number of identifier switches is the most important performance measure in scenarios that



(a) Track misses.

(b) False positive tracks.

(c) Identifier switches.

(d) MOTA (solid lines, left axis) and average cycle times (dashed lines, right axis).

Figure 7.11: Detailed analysis of the results on the city center data set using modeled social relations. Tracking social relations an employing a physically grounded occlusion (cyan) the model the most accurate tracking results are achieved. These improvements come at the expense of a lower frame-rate.

involve interaction with and motion prediction of people.

Adding Physically Grounded Occlusion Information

Adding the occlusion model proposed in subsection 5.4.1 allows to better estimate the occlusion probability of people walking in groups or solely based on their position and the geometric constraints of the environment (see Katz et al. [2008]). This leads to further improvements of the tracking accuracy (MOTA) increasing to 84.7% (+5.1%) on the Freiburg city center and to 84.5% (+3.1%) on the main station data set. Especially, when lengthly occlusion events are not caused by group members but by static obstacles of the environment track losses are avoided leading to an reduction of the identifier switches (ID) and track misses (FN). This improvement comes at the expense of an increased number of false positives as systematic misdetections and targets disappearing from the sensor field of view are maintained longer. As shown in Figure 7.10 this effect is particularly high using Bayes filtered social relations with low number of hypotheses ($k < 50$).

7.8 Conclusions

In this chapter the problems of detecting and learning socio-spatial relations between people as well as inferring and tracking their social groupings are addressed. The proposed approach, that relies on an extension of a multi-hypothesis tracking approach, also improves person-level tracking in two ways: the social grouping information is used to predict human motion over learned intra-group constraints and to support data association by adapting track-specific occlusion probabilities. Both

Data Set	Approach	FN	FP	ID	MOTA	*Hz*
Freiburg city center	baseline	4746	2344	261	79.6%	52.5
	geometric rel.	4179 (-11.9%)	2287 (-2.4%)	196 (-24.9%)	81.5%	25.2
	filtered social rel.	3407 (-28.2%)	2752 (+17.4%)	186 (-28.7%)	82.4%	27.8
	+ geometric rel.	3169 (-33.2%)	2808 (+19.8%)	161 (-38.3%)	83.0%	20.5
	+ occlusion model	2841 (-40.1%)	2950 (+25.8%)	151 (-42.1%)	83.6%	13.3
	tracked social rel.	3600 (-24.1%)	2387 (+1.8%)	169 (-35.2%)	82.9%	29.6
	+ geometric rel.	3472 (-26.8%)	2390 (+1.9%)	152 (-41.8%)	83.3%	17.6
	+ occlusion model	2949 (-37.8%)	2416 (+3.1%)	141 (-45.9%)	84.7%	12.2
Freiburg main station	baseline	2949	2982	360	81.4%	31.1
	geometric rel.	2342 (-20.6%)	3138 (+5.2%)	274 (-23.9%)	83.0%	23.7
	filtered social rel.	2067 (-29.9%)	3488 (+16.9%)	235 (-34.7%)	82.9%	25.3
	+ geometric rel.	1878 (-36.3%)	3459 (+15.9%)	221 (-38.6%)	83.6%	18.8
	+ occlusion model	1773 (-39.9%)	3579 (+20.0%)	177 (-50.8%)	83.7%	12.1
	tracked social rel.	2122 (-28.0%)	3158 (+5.9%)	192 (-46.7%)	83.8%	23.7
	+ geometric rel.	2108 (-28.5%)	3150 (+5.6%)	183 (-49.2%)	83.9%	18.4
	+ occlusion model	1887 (-36.0%)	3218 (+7.9%)	156 (-56.7%)	84.5%	12.1

Table 7.2: CLEAR MOT results of both data sets using $N_{Hyp} = 100$ hypotheses. Learning socio-spatial relations and inferring group affiliations from models found in the social science community the numbers of track misses (FN) and identifier switches (ID) can be increased by a factor of two. Combined with the occlusion model proposed in subsection 5.4.1 the most accurate tracking result is achieved. These improvements come at the expense of an increased number of false positives (FP) and a lower frame-rate.

measures lead to an improved occlusion handling and a better trade-off between false negative and false positive tracks.

Opposed to most related works that use static overhead cameras and batch approaches, a mobile platform is used, geometric relations are learned in an on-line fashion, and group affiliations are tracked with a recursive multi-model multi-hypothesis tracker in real-time. With up to 50% fewer track identity switches and 28% fewer false negative tracks, the results suggest that tracking people in 2D range data can strongly benefit from estimates on social and geometrical relations, mostly due to their ability to explain lengthy occlusion events and maneuvers of groups. Combining the approach with the previously presented physically grounded occlusion model yields even further improvements.

In future work the system is planed to be deployed on RGB-D data. Furthermore, additional attribute information on people such as age and gender will be incorporated to improve the social relation estimation.

Part IV

Learning Appearances and Appearance Dynamics

8 Unsupervised Learning Of Dynamic Objects

For robots operating in real-world environments, the ability to deal with dynamic entities such as humans, animals, vehicles, or other robots is of fundamental importance. The variability of dynamic objects, however, is large in general, which makes it hard to manually design suitable models for their appearance and dynamics. In this chapter, an unsupervised learning approach to this model-building problem is presented. An exemplar-based model is employed to describe and represent the time-varying appearance of dynamic objects in planar laser range data. Additionally, a clustering procedure is presented that builds a set of object classes from given observation sequences. Incoming information is normalized to achieve translational and rotational invariance which enables to employ (self-) similarity measures to classify tracks of relevant object classes in a Bayesian filtering framework. Extensive experiments with 500 tracks in real environments demonstrate that the proposed system is able to classify objects of six classes and to autonomously learn useful models for, e.g., pedestrians, skaters, or cyclists without being provided with external class information. The classification accuracy reaches more than 98%.

This chapter is structured as follows. Introduction and related work are presented in section 8.1 and section 8.2. The theory on learning exemplar-based models of dynamic objects for range-bearing observations is introduced in section 8.3 followed by section 8.4 explaining the classification of dynamic objects using a Bayesian filtering framework. Subsequently, the approach on unsupervised learning of a set of object classes is presented in section 8.5. The segmentation and tracking system is presented in section 8.6 followed by the experiments in section 8.7. Finally, section 8.8 concludes the chapter.

8.1 Introduction

The problem of tracking dynamic objects and modeling their time-varying appearance has been studied extensively in robotics, engineering, computer vision, and other areas. On one hand, the problem is hard as the appearance of objects is ambiguous, partly occluded, may vary quickly over time, and is perceived via a high-dimensional measurement space. On the other hand, the problem is highly relevant in practice – especially in future applications for mobile robots and intelligent cars. Consider, for example, a service robot deployed in a populated environment such as a pedestrian precinct. Tasks like collision-free navigation or interaction require the ability to recognize, distinguish, and track moving objects including reliable estimates of object classes, e.g., “adult”, “infant”, “car”, “dog”, etc.

In this chapter, the problem of detecting, tracking, and classifying moving objects in sequences of planar range scans acquired by a 2D laser range finder is considered. It introduces an exemplar-based model for representing the time-varying appearance of moving objects as well as a clustering procedure that builds a set of object classes from given observation sequences in conjunction with a Bayes filtering scheme for classification. The proposed system, which has been implemented and tested on a real robot, does not require labeled object trajectories, but rather uses an unsupervised

Figure 8.1: Six examples of relevant object classes considered in this work. The proposed system learns probabilistic models of their appearance in planar range scans and the corresponding dynamics. The classes are denominated Pedestrian (PED), Buggy (BUG), Skater (SKA), Suitcase (SUI), Cyclist (CYC), and Kangaroo-shoes (KAN).

clustering scheme to automatically build appropriate class assignments. By pre-processing the sensor stream using state-of-the-art feature detection and tracking algorithms, a system is obtained that is able to learn and re-use object models on-the-fly and without human intervention. The resulting set of object models can then be used to recognize previously seen object classes and to improve data segmentation and association in ambiguous multi-target tracking situations. Furthermore, the object models can be used in various applications to associate semantics with recognized objects depending on their classes.

8.2 Related Work

Exemplar-based models are frequently applied in computer vision systems for dealing with the high dimensionality of visual input. In Toyama and Blake [2002], for instance, probabilistic exemplar models are used for representing and tracking human motion. Their approach is similar in that they also learn probabilistic transition models. As the major differences, the range-bearing observations used in this work are substantially more sparse than visual input. Additionally, the presented approach also addresses the problem of learning different object classes in an unsupervised fashion. The work of Plagemann et al. [2005] used exemplars to represent the visual appearance of 3D objects in the context of an object localization framework. In Kruger et al. [2006] exemplar models are learned to realize a face recognition system for video streams and Wolf et al. [2005] have been used image retrieval methods for localizing a mobile robot in an environment. Latter can also be regarded as an exemplar-based technique for dealing with the high-dimensional and continuously changing appearance of places. Exemplar-based approaches have also been used in other areas such as action recognition as addressed in Drumwright et al. [2004] or word sense disambiguation as proposed by Ng and Lee [1996]. The work of Wren et al. [1997] introduces a people modeling and tracking system for color images. It uses a multi-class model of shape and color and has an explicit background model to perform image segmentation.

There exists a large body of work on laser-based object and people tracking in the robotics literature. Most relevant are the methods introduced by Schulz et al. [2001], Kluge et al. [2001], Montemerlo and Thrun [2002], Fod et al. [2002], Topp and Christensen [2005], and Arras et al. [2007]. People tracking typically requires carefully engineered or learned features for track identification and data association and often a-priori information about motion models. MacLachlan and Mertz [2006] showed, that this is also the case for geometrically simpler and rigid object such as vehicles in traffic scenarios. Assuming the relevant motion models are known in advance, in the work of Cui et al. [2006b] a system for tracking single persons within a larger set of people is described.

The work most closely related to the proposed approach has been presented by Schulz [2006], who

combined vision- and laser-based exemplar models to realize a people tracking system. In contrast to his work, the main contribution here is the unsupervised learning of multiple object classes that can be used for tracking as well as for classifying dynamic objects. Ilg et al. [2004] also follow a prototype-based approach. In contrast to this work, they explicitly align time series using Dynamic Time Warping to perform a clustering into prototypes.

Periodicity and self-similarity have been studied by Cutler and Davis [2000], who developed a classification system based on the autocorrelation of appearances, which is able to distinguish, for example, walking humans from dogs.

A central component of the presented approach described in the following section is an unsupervised clustering algorithm to produce a suitable set of exemplars. Most approaches to cluster analysis (see Hartigan [1975]) assume that all data is available from the beginning and that the number of clusters is given. Other works in this area like Tasoulis et al. [2006] and Chis and Grosan [2006] also deal with sequential data and incremental model updates. Fei-Fei et al. [2003] learn visual object classes in an unsupervised manner proposing an one-shot approach to efficiently learn new models using information from previously seen classes of unrelated categories. Kulic et al. [2008] have proposed an hierarchical on-line clustering algorithm for unsupervised learning of HMM models from motion trajectories based on work by Kohlmorgen-Lemm segmentation. The work of Ghahramani [2004] gives an easily accessible overview of the state-of-the-art in unsupervised learning.

As an alternative to the exemplar-based approach, researchers have applied generic dimensionality reduction techniques to deal with high-dimensional and/or dynamic appearance distributions. PCA and ICA have, for example, been used by Wang and Han [2005] and Fortuna and Capson [2004] to recognize people from iris images or their faces, respectively. Recent advances in this area include latent variable models, such as Gaussian process latent variable models (GPLVMs) proposed by Lawrence [2005]. However, this approach is not feasible in an unsupervised approach where the latent space can not be learned from training data

The approach of Wang et al. [2006], termed Gaussian process dynamical models (GPDMs), builds on the idea that the high-dimensional data which is observed over time actually lies on a low-dimensional manifold. They build on GPLVMs to learn and represent the low-dimensional embedding in a nonparametric way. The feasibility of this approach has been shown for the different problem of body pose tracking from visual input. The general approach nevertheless constitutes an alternative to the model presented in this chapter.

The work of Jenkins and Matarić [2004] extends the Isomap proposed by Tenenbaum et al. [2000], which is another popular method for nonlinear dimensionality reduction, by a spatio-temporal component which allows to model high-dimensional data that changes over time. One of their example instantiations of the model shows that it develops into a HMM-like structure for clustered data. More details on dimensionality reduction are provided in the standard text book by Bishop [2006].

8.3 Modeling Object Appearance and Dynamics Using Exemplars

Exemplar models are representations for both, time-varying *appearance* and appearance *dynamics*. They are a choice consistent with the motivation for an unsupervised learning approach avoiding manual feature selection, parameterized physical models (e.g., human gait models), and hand-tuned classifier creation.

This section describes, how the exemplar-based models of dynamic objects are learned. Based on a segmentation and the tracking system presented in section 8.6, it is assumed to have a discrete track for each dynamic object in the current scene. Over time, these tracks describe trajectories that

are analyzed regarding temporal-dependent appearance and dynamics of the corresponding objects.

8.3.1 Problem Description

Let $\mathcal{T} = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ be a *track*, in example, a time-indexed observation sequence of appearances $\{\mathbf{z}_t\}_{t=1}^m$, of an object belonging to an *object class* $\mathcal{C}$. The two problems addressed in this chapter can be formalized as follows:

1. **Unsupervised learning:** Given a set of observed tracks $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots\}$, learn the classes $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ of objects in an unsupervised manner. This amounts to setting an appropriate number n of classes and to learn for each class $\mathcal{C}_j$ a probabilistic model $p(\,\mathcal{T} \mid \mathcal{C}_j\,)$ that characterizes the time-varying appearance of tracks $\mathcal{T}$ associated with that class.

2. **Classification:** Given a newly observed track $\mathcal{T}$ and a set of known object classes $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$, estimate the class probabilities $p(\,\mathcal{C}_j \mid \mathcal{T}\,)$ for all classes.

Note that "unsupervised" in this context does not mean that *all* model parameters are learned from scratch, but rather that the important class information (e.g. "pedestrian", "cyclist") is not supplied to the system. The underlying segmentation, tracking, and feature extraction subsystems employed in this chapter are designed to capture a wide variety of possible object appearances and the unsupervised learning task is to build a compact representation of object appearance that generalizes across instances.

8.3.2 The Exemplar Model

The exemplar models introduced by Toyama and Blake [2002] aim at approximating the typically high-dimensional and dynamic appearance distribution of objects using a sparse set $\mathcal{E} = \{\mathbb{E}_1, \ldots, \mathbb{E}_r\}$ of significant observations, termed *exemplars* $\mathbb{E}_i$. Similarities between concrete observations and exemplars as well as between two exemplars are specified by a distance function $\rho(\mathbb{E}_i, \mathbb{E}_j)$ in exemplar space. Furthermore, each exemplar is given a prior probability $\pi_i = p(\mathbb{E}_i)$, which reflects the prior probability of a new observation being associated with this exemplar. Changes in appearance over time are dealt with by introducing transition probabilities $p(\mathbb{E}_i \mid \mathbb{E}_j)$ between exemplars w.r.t. a predefined iteration frequency. Formally, this renders the exemplar model a first-order Markov chain, specified by the four elements $\mathcal{M} = (\mathcal{E}, \mathcal{B}, \pi, \rho)$, which are the exemplar set $\mathcal{E}$, the transition probability matrix $\mathcal{B}$ with elements $b_{i,j} = p(\mathbb{E}_i \mid \mathbb{E}_j)$, the priors π, and the distance function ρ. All these components can be learned from data, which is one of the central topics of this chapter.

8.3.3 Exemplars for Range-Bearing Observations

In a laser-based object tracking scenario, the raw laser measurements associated with each track constitute the appearance $\mathbf{z} = \{(\phi_j, \rho_j)\}_{j=1}^l$ of the objects, where ϕ_i is the bearing, ρ_i is the range measurement, and l is the number of laser end points in the respective laser segment.

To cluster the laser segments into exemplars, the individual laser segments need to be normalized with respect to rotation and translation. This is achieved using the state information estimated by the underlying tracker[46]. Here, the state of a track $\mathbf{x}_t = (\; x_t \;\; y_t \;\; \dot{x}_t \;\; \dot{y}_t \;)^T = (\; \boldsymbol{x}_t \;\; \boldsymbol{v}_t \;)^T$ is composed of the target position $\boldsymbol{x}_t = (\, x_t, y_t \,)$ and velocities $\boldsymbol{v}_t = (\, \dot{x}_t, \dot{y}_t \,)$. The velocity vector $\boldsymbol{v}_t$ can then be used to calculate the heading of the object. As shown in Figure 8.2 translational invariance is achieved by shifting the center of gravity of the segment to $(0, 0)$, while rotational invariance is gained from

[46] In this work, the state information is estimated using Kalman filters. Other techniques, like particle filters, can be employed as well.

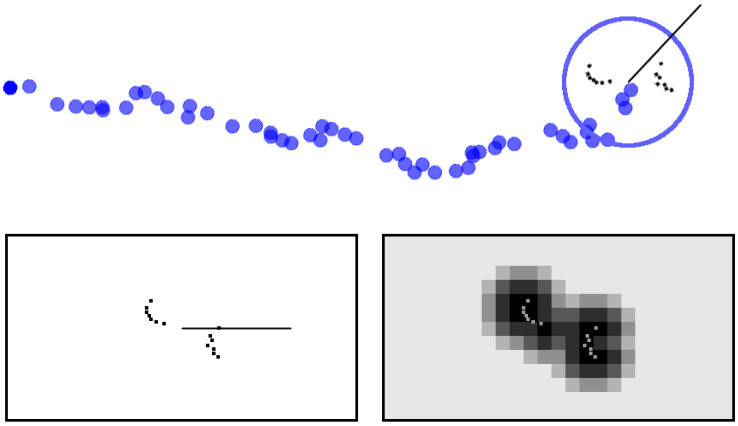


Figure 8.2: Pre-processing steps illustrated with a pedestrian observed via a 2D laser range finder. The top figure shows the trajectory of the subject moving from left to right. The segmentation and tracking system yields estimates of target location (shown as a trace of large dots), orientation (shown by a line), and velocity. The bottom left figure shows the raw range readings (small dots) that are normalized such that the estimated motion direction is zeroed. The resulting grid-based representation G generated from the set of normalized laser end points is depicted in the bottom right figure.

zeroing the orientation in the same way. After normalization, all segments have a fixed position and orientation simplifying further processing steps.

Rather than using the raw laser end points (ϕ, ρ) of the normalized segments as observations (see Schulz Schulz [2006]), the so called *likelihood field* introduced by Thrun [2001] is calculated for each of them on a regular grid. In this model, the likelihood of a range measurement is a function of the Euclidean distance d_{euc} of the respective endpoint of the beam to the closest obstacle in the environment. The likelihood of each cell (x, y) is then calculated using a Gaussian distribution $\mathcal{N}(d_{euc}(x, y); 0, \sigma^2)$ with zero mean and standard deviation σ which reflects the sensor noise. In the past, likelihood fields have been used successfully for tasks like localization or scan matching. The main advantage of this approach is that the distance function for observations can be defined independently of the number of laser end points in the segment and that likelihood estimation for new observations can be performed efficiently. Henceforth, the grid representation of an appearance $\mathbf{z}_i$ will be denote as $\mathbf{G}_i$[47]. Figure 8.2 shows an example of a track, a laser segment, the normalized segment, and the corresponding grid for a walking pedestrian.

8.3.4 Validation of the Exemplar Approach

The choice of the exemplar representation has a strong impact on both the creation of the exemplar set $\mathcal{E}$ from a sequence of appearances $\mathbf{z}$ and the unsupervised creation of new object classes $\mathcal{C}$. This motivates a careful analysis of the choices made. To illustrate the practicability of the exemplar model for the purpose of representing time-varying appearances of objects in 2D range data with few representative exemplars, the self-similarity of the (grid) observations $\mathbf{G}_i$ for tracks of objects from relevant object classes are analyzed first. The similarity $\mathcal{S}(\mathbf{G}_{t_1}, \mathbf{G}_{t_2})$ of two observations obtained at

[47] In the remainder of the chapter, the grid representation $\mathbf{G}_i$ is treated as input to the learning and classification methods and therefore also called observation.

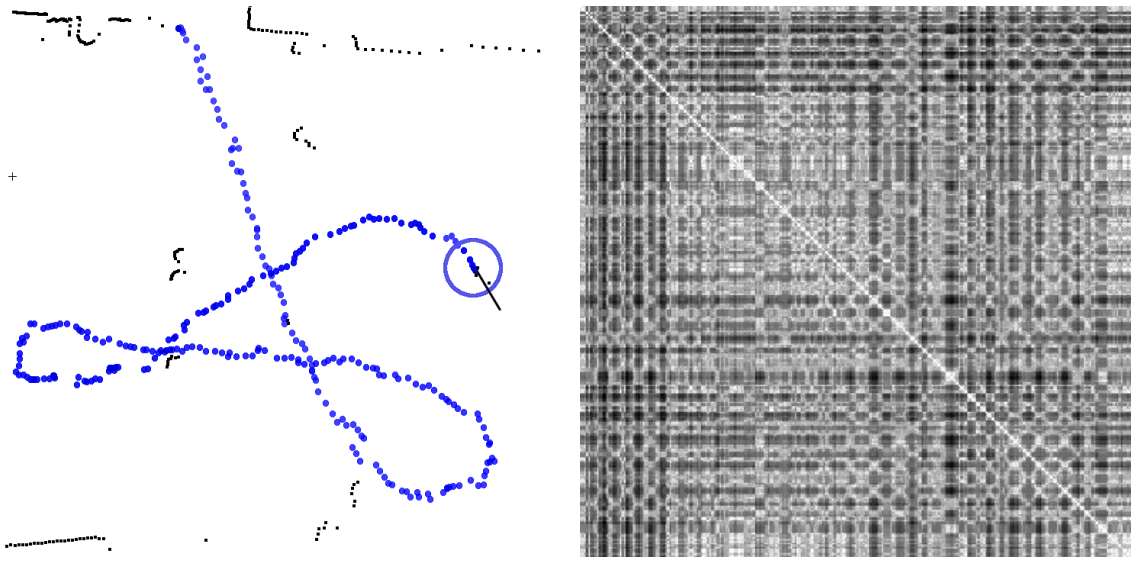


Figure 8.3: Trajectory (left) and self-similarity matrix (right) of a pedestrian walking in a large hallway. The track consists of 387 observations. The walking cycle of the pedestrian causes a periodic structure of the self-similarity matrix. The relative stability of this structure demonstrates how the exemplar representation is able to uncover salient appearance properties with invariance to the subject's heading and to self-occlusion by the legs.

times t_1 and t_2 is defined by the absolute correlation function, thus

$$\mathcal{S}(\mathbf{G}_{t_1}, \mathbf{G}_{t_2}) := \sum_{(x,y)\in\mathbf{B}} | \mathbf{G}_{t_1}(x,y) - \mathbf{G}_{t_2}(x,y) |, \tag{8.1}$$

where $\mathbf{B}$ is the bounding box of the grid-based representations $\mathbf{G}_{t_1}$ and $\mathbf{G}_{t_2}$ of the observations $\mathbf{z}_{t_1}$ and $\mathbf{z}_{t_2}$, respectively.

Figure 8.3 visualizes the self-similarity matrix for 387 observations of a pedestrian. Both axes of this matrix (Figure 8.3, right) correspond to the time with t_1 along the horizontal and t_2 along the vertical axis. The gray values that encode self-similarity range from bright to dark. Whereas light gray stands for maximal correlation, black represents minimal self-similarity. The diagonal is maximal by definition as the distance of an observation to itself is zero.

A periodic structure of the matrix can be recognized, which is caused by the strong self-similarity of the appearance of the pedestrian over a walking cycle. This is not self-evident as the appearance of the walking person in laser range data changes with the heading of the person relative to the sensor. Poor normalization (e.g., because of inaccurate heading estimates of the underlying tracker) or a poor exemplar representation (e.g., which is too sensitive to measurement noise) would have removed the periodicity in the pre-processed data. This illustrates that the normalization and the grid-based representation of appearance has sufficiently good invariance properties, so that a small amount of salient appearance patterns, i.e. *exemplars*, and the transitions between them are well suited for the goal to learn and classify dynamic objects.

8.3.5 Learning the Exemplar Model

This section describes how exemplar models are learned from observation sequences $\mathbf{z}$. This involves the definition of an appropriate distance function ρ and learning the exemplar set $\mathcal{E}$, the prior

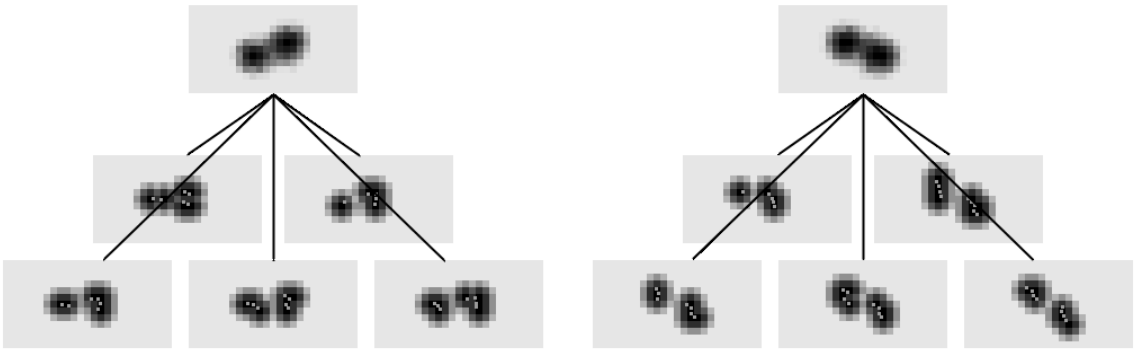


Figure 8.4: Example clusters of pedestrian observations $\mathbf{G}_i$. The diagram shows the centroids of two clusters (exemplars $\mathbb{E}_1, \mathbb{E}_2$) each created from a set of five observations $\mathbf{G}_i$. Goal of the clustering process is to obtain exemplars with high intra-cluster similarity $\mathcal{S}(\mathbf{G}_i, \mathbf{G}_j)$ of the observation $\mathbf{G}_i$ and $\mathbf{G}_j$. Vice versa, the distance $\rho(\mathbf{G}_i, \mathbf{G}_j)$ between those observations is small. Exemplars can be described as as representative yet temporally smoothed appearances of dynamic objects.

probabilities π_i, and the transition matrix $\mathcal{B}$ encoding the transition probabilities $b_{i,j} = p(\mathbb{E}_i \mid \mathbb{E}_j)$.

Distance Function For Exemplar Learning

The similarity between two observations $\mathbf{z}_i$ and $\mathbf{z}_j$ is assess based on a distance function applied to the corresponding grid-based representations $\mathbf{G}_i$ and $\mathbf{G}_j$. Interpreting the grids as two dimensional histograms various distance functions compared in Gonzalez-Barbosa and Lacroix [2002] can be employed for this purpose. Some of these functions have been compared, such as

- Euclidean distance

$$\rho_{euc}(\mathbf{G}_i, \mathbf{G}_j) \;=\; \sqrt{\sum_{(x,y)\in\mathcal{I}} \left(\mathbf{G}_i(x,y) - \mathbf{G}_j(x,y)\right)^2}, \tag{8.2}$$

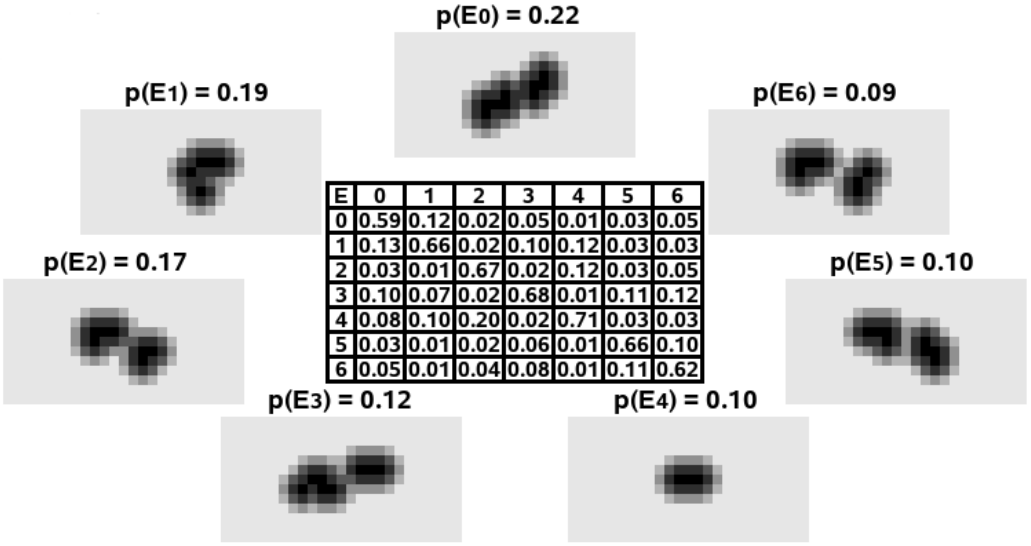
- the Haussler distance

$$\rho_h(\mathbf{G}_i, \mathbf{G}_j) \;=\; \sum_{(x,y)\in\mathcal{I}} \frac{|\mathbf{G}_i(x,y) - \mathbf{G}_j(x,y)|}{1 + \mathbf{G}_i(x,y) + \mathbf{G}_j(x,y)}, \tag{8.3}$$

- and the χ^2 statistics

$$\rho_{\chi^2}(\mathbf{G}_i, \mathbf{G}_j)^2 \;=\; \sum_{(x,y)\in\mathcal{I}} \frac{\left(\mathbf{G}_i(x,y) - \mathbf{G}_j(x,y)\right)^2}{\mathbf{G}_i(x,y) + \mathbf{G}_j(x,y)}\ . \tag{8.4}$$

The Euclidean distance achieved the best results, thus it is used in the rest of the chapter and $\rho(\mathbf{G}_i, \mathbf{G}_j) := \rho_{euc}(\mathbf{G}_i, \mathbf{G}_j)$. Note, the function is used for both, the clustering and the Gaussian observation model in Eq. 8.9 described hereafter.



E	0	1	2	3	4	5	6
0	0.59	0.12	0.02	0.05	0.01	0.03	0.05
1	0.13	0.66	0.02	0.10	0.12	0.03	0.03
2	0.03	0.01	0.67	0.02	0.12	0.03	0.05
3	0.10	0.07	0.02	0.68	0.01	0.11	0.12
4	0.08	0.10	0.20	0.02	0.71	0.03	0.03
5	0.03	0.01	0.02	0.06	0.01	0.66	0.10
6	0.05	0.01	0.04	0.08	0.01	0.11	0.62

Figure 8.5: Laser-based exemplar model of a pedestrian. The transition matrix is shown in the center with the exemplars sorted counterclockwise according to their prior probability.

Exemplar Set

Exemplars are representations that generalize distinctive object appearance. To this aim, similar appearances are associated and merged into clusters. In the proposed system this is achieved by applying k-means clustering[48] (see Hartigan [1975]) on the set of grid based observations $\{\mathbf{G}_1, \ldots, \mathbf{G}_m\}$ to partition the full data set into r clusters $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_r$. At the beginning of the clustering procedure, r observations are taken randomly as single-element clusters and the remaining observations are assigned to the cluster with minimal distance, calculated using Eq. 8.2. Until convergence, the centroids are recomputed and the observations moved to that correct clusters. Both, centroids and resulting exemplars are calculated simply by averaging the likelihoods of each grid cell. Figure 8.4 shows two of the resulting clusters and some assigned observations.

Strong outliers in the training set – which cannot be merged with other observations – are retained by the clustering process as additional, non-representative exemplars. Such observations may occur for several reasons, e.g., when a tracked object performs atypical movements, when the underlying segmentation method fails to produce a proper foreground segment, or due to sensor noise. To achieve robustness with respect to such outliers, an exemplar is accepted only if it was created from a minimum number of observations[49]. This assures that the resulting set of exemplars $\mathcal{E}$ characterizes only states of the appearance dynamics that occur often and are representative.

Transition Probabilities and Priors

Once the clustered exemplar set $\mathcal{E}$ has been generated from the observation sequence $\mathbf{z}$, the transition probabilities between exemplars can be learned. As defined in subsection 8.3.2, the dynamics of the appearance of an object are modeled using hidden Markov models (HMM). The transition probabil-

[48] Alternatively, agglomerative hierarchical clustering (AHC) can be employed. Experiments with both approaches yield comparable results.

[49] To achieve the requested number of clusters the observations must be partitioned into more than r examples to allow outliers to be filtered out.

ities are obtained by pair-wise counting. A transition between two exemplars $\mathbb{E}_i$ and $\mathbb{E}_j$ is counted each time an observation that has minimal distance to $\mathbb{E}_i$ is followed by an observation with minimal distance to $\mathbb{E}_j$. As there is a non-zero probability that some transitions are never observed although they exist, the transition probabilities $b_{i,j}$ are initialized with a small value to moderately smooth the resulting model. Accordingly, the exemplar priors π_i are determined by counting the number of contributing observations $\mathbf{G}$ in each cluster. See Figure 8.5 for the learned exemplar model of a pedestrian.

8.4 Classification

Having learned the exemplar set and the transition probabilities as described in the previous section, both can be used to classify tracks of different objects in a Bayesian filtering framework. More formally, given the grid representations $\{\mathbf{G}_1, \ldots, \mathbf{G}_m\}$ of the observations of a track $\mathcal{T}$ and a set of learned classes $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ including their corresponding representative exemplar models $\{\mathcal{M}^1, \ldots, \mathcal{M}^n\}$, an algorithm to estimate the class probabilities $p_t(\mathcal{C}_k \mid \mathcal{T})_{k=1}^{n}$[50] for every time step t is wanted. The estimates for the last time step m then reflect the consistency of the entire track with the different exemplar models. Finally, these quantities can be used to make classification decisions.

8.4.1 Estimating Class Probabilities over Time

Each exemplar model $\mathcal{M}^i$ represents the distribution of track appearances for its corresponding object class $\mathcal{C}_i$. Thus, a combination of *all* known exemplar models can be defined as

$$\mathcal{M}^{comb} \;=\; \{\mathcal{M}^1, \ldots, \mathcal{M}^n\}. \tag{8.5}$$

$\mathcal{M}^{comb}$ covers the entire space of possible appearances – or, more precisely, of all appearances that have been observed so far[51]. The combined set of exemplar $\mathcal{E}^{comb}$ of $\mathcal{M}^{comb}$ is constructed by simply building the union set of the individual exemplar sets $\mathcal{E}^k$ of all models $\mathcal{M}^k$, hence

$$\mathcal{E}^{comb} \;=\; \bigcup_k \mathcal{E}^k. \tag{8.6}$$

The transition probability matrix $\mathcal{B}^{comb}$ as well as the exemplar priors π^{comb} can be obtained from the $\mathcal{B}^k$ matrices and the π^k prior probabilities in a straightforward way

$$\mathcal{B}^{comb} \;=\; \eta_{\mathcal{B}} \begin{pmatrix} \mathcal{B}^1 & & \varepsilon \\ & \ddots & \\ \varepsilon & & \mathcal{B}^n \end{pmatrix} \tag{8.7}$$

$$\pi^{comb} \;=\; \eta_{\pi} \left(\pi^1, \ldots, \pi^n\right)^T,$$

where $\eta_{\mathcal{B}}$ and η_{π} are normalizers. If it is assumed that the corresponding exemplar sets do not intersect all cross-model transition probabilities $\varepsilon_{i,j}$ in $\mathcal{B}^{comb}$ can be set to zero. This assumption means that objects do not change their class during the time of observation, that is to say, for example, that no skater takes off his shoes and becomes a pedestrian.

[50] In this work only one exemplar model is learned for each object class $\mathcal{C}_k$. Therefore the class probability is equivalent to the probability of the corresponding model, hence $p_t(\mathcal{C}_k \mid \mathcal{T}) = p_t(\mathcal{M}_k \mid \mathcal{T})$.

[51] Excluding outliers that have been filtered out.

Algorithm 4: Bayes filtering algorithm for object classification.

Input : Sequence of grid observations $\{\mathbf{G}_1, \ldots, \mathbf{G}_m\}$ of a track $\mathcal{T}$.
Set of exemplar models $\{\mathcal{M}^1, \ldots, \mathcal{M}^n\}$ describing time-varying appearances of n different object classes $\{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ and the prior model probabilities $p(\mathcal{M}^k)$.
Cross-model transition probability ε.

Output : Class probabilities $p_m(\,\mathcal{M}^k \mid \mathcal{T}\,)$ and most likely class assignment $\mathcal{M}^{best}(\mathcal{T})$.

Variables: Belief function $Bel_t(\mathbb{E}_i^k)$ encoding the probability of the i^{th} exemplar in the k^{th} model $\mathcal{M}^k$.

```
/* construct the combined exemplar model M^comb */
E^comb ← ⋃_{k=1}^n E^k;
B^comb ← η_B ( diag(B^1, ..., B^n) + (1 − I)ε );
π^comb ← η_π (π^1, ..., π^n)^T;
/* initialize the belief function Bel_0(E_i^k) ∀i,k */
for k ← 1 to n and ∀i do
    Bel_0(E_i^k) ← p(M^k) π_i^k;
end
/* recursive update of the belief function (main loop) */
for t ← 1 to m do
    // update probabilities
    for k ← 1 to n and ∀i do
        Bel̂_t(E_i^k) ← p( G_t | E_i^k ) Σ_l Σ_j ( p( E_i^k | E_j^l ) Bel_{t−1}(E_j^l) );
    end
    // normalize
    η_t ← Σ_{i,k} Bel̂_t(E_i^k);
    for k ← 1 to n and ∀i do
        Bel_t(E_i^k) ← Bel̂_t(E_i^k);
    end
end
/* calculate individual class probabilities p_m( M^k | T ) */
for k ← 1 to n do
    p_t( M^k | T ) ← Σ_i Bel_t(E_i^k);
end
/* return classification result */
M^best(T) ← argmax_k p_m( M^k | T );
return M^best(T);
```

Figure 8.6: Bayes filtering algorithm for object classification.

Given this combined exemplar model $\mathcal{M}^{comb}$, a belief function Bel_t for the class probabilities $p_t(\,\mathcal{C}_k \mid \mathcal{T}\,)_{k=1}^n$ can be updated recursively over time using the well-known Bayes filtering scheme. For better readability, the notation $\mathbb{E}_i^k$ is introduced to refer to the i^{th} exemplar of model $\mathcal{M}^k$. According to the Bayes filter, the belief about object classes is initialized as

$$Bel_0(\mathbb{E}_i^k) \;=\; p(\mathcal{M}^k)\;\pi_i^k, \tag{8.8}$$

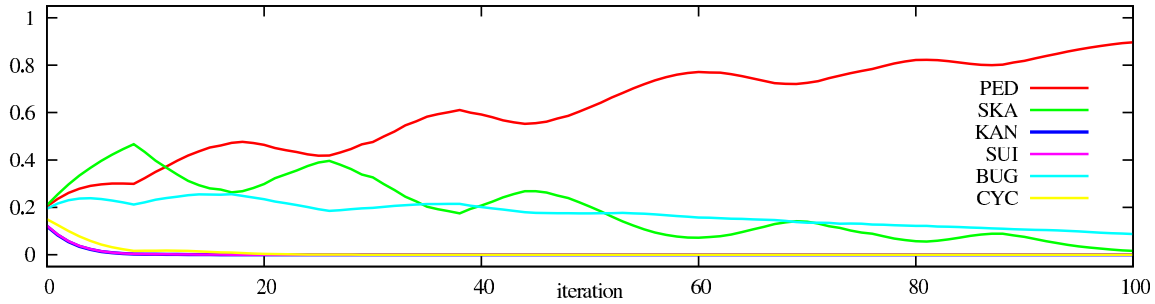


Figure 8.7: The graphs show the evolution of the probabilities of different classes over time during an experiment in which a pedestrian is being observed. The x-axis refers to the time t. The classes that compete most are pedestrian (solid line) and skater (dashed line). The periodicity in the graphs corresponds to the walking or skating cycle respectively. Both cycles have very similar appearance in the laser data. Integrated over time, however, the pedestrian class obtains maximum posterior probability, which corresponds to the ground truth.

where π_i^k denotes the prior probability of $\mathbb{E}_i^k$ and $p(\mathcal{M}^k)$ stands for the model prior, which is assumed to be uniform (or can be estimated from a training set). Starting with $\mathbf{G}_1$, the recursive update of the belief function is performed for every $\mathbf{G}_t$ following

$$Bel_t(\mathbb{E}_i^k) \;=\; \eta_t \; p(\,\mathbf{G}_t \mid \mathbb{E}_i^k\,) \sum_l \sum_j \Big(p(\,\mathbb{E}_i^k \mid \mathbb{E}_j^l\,) \; Bel_{t-1}(\mathbb{E}_j^l) \Big). \tag{8.9}$$

In this equation, η_t is a normalizing factor ensuring that $Bel_t(\mathbb{E}_i^k)$ sums up to one over all i and k, and $p(\,\mathbf{G}_t \mid \mathbb{E}_i^k\,)$ is the Gaussian observation likelihood using the distance function in Eq. 8.2.

The estimates of the exemplar probabilities $Bel_t(\mathbb{E}_i^k)$ at time t can finally be summed up to yield the individual class probabilities

$$p_t(\,\mathcal{M}^k \mid \mathcal{T}\,) \;=\; \sum_i Bel_t(\mathbb{E}_i^k). \tag{8.10}$$

At time $t = m$, that is, when the entire observation sequence has been processed, $p_m(\,\mathcal{M}^k \mid \mathcal{T}\,)$ constitute the resulting estimates of the model probabilities. In particular, the most likely model assignment for track T can be defined as

$$\mathcal{M}^{best}(\mathcal{T}) \;:=\; \mathrm{argmax}_k \; p_m(\,\mathcal{M}^k \mid \mathcal{T}\,). \tag{8.11}$$

Finally, the most likely class associated with the best model is defined as $\mathcal{C}^{best}(\mathcal{T})$. The filtering process described above is visualized by an example run for a pedestrian track $\mathcal{T}$ (see Figure 8.3). The plot in Figure 8.7 shows the class probabilities for six alternative object classes over time. In algorithm 4 the pseudo code of the Bayes filtered classification process is given.

8.5 Unsupervised Learning

As the variety of dynamic objects in the world is hard to predict a-priori, an unsupervised approach to learn such objects without external class information is proposed. This section explains, how a set of object classes can be learned from scratch in an unsupervised manner.

Objects of a previously unknown type $\mathcal{C}^{new}$ will always be assigned to some class $\mathcal{C}^{best}$ by the Bayes filter. The class with the highest resulting probability estimate provides the current best, yet

	$\mathbf{K} \geq 1$	$\mathbf{K} \geq 2$	$\mathbf{K} \geq 4$	$\mathbf{K} \geq 8$	$\mathbf{K} \geq 20$
PED / PED	41%	2%	0%	0%	0%
SKA / SKA	58%	7%	0%	0%	0%
CYC / CYC	79%	32%	14%	10%	8%
BUG / BUG	78%	47%	21%	9%	1%
KAN / KAN	60%	40%	21%	11%	3%
PED / KAN	46%	3%	0%	0%	0%
PED / SKA	100%	83%	40%	10%	0%
CYC / BUG	100%	100%	100%	99%	50%
BUG / KAN	100%	100%	100%	100%	82%
CYC / KAN	100%	100%	100%	98%	92%

Table 8.1: Percentages of incorrectly separated (top five rows) and correctly separated (bottom five rows) track pairs. A Bayes factor is sought that trades off separation of tracks from different classes and association of tracks from the same class.

suboptimal description of the object at that time. A better fit would always be achieved by creating a new, specifically trained model for this particular object instance. Thus, the classic model selection problem is faced. The model selection problem is defined as choosing between a more compact vs. a more precise model for explaining the observed data. As a selection criterion, the *Bayes factor* by Kass and Raftery [1995] is employed. The Bayes factor considers the amount of evidence in favor of a model relative to an alternative one.

More formally, given a set of known classes $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ and their respective models $\{\mathcal{M}^1, \ldots, \mathcal{M}^n\}$, further let $\mathcal{T}$ be the track of an object to be classified. First, the current best matching model $\mathcal{M}^{best}(\mathcal{T})$ according to Eq. 8.11 is determined. Thereafter, a new, fitted model $\mathcal{M}^{new}(\mathcal{T})$ that describes the new data is learned as described in subsection 8.3.5. To decide whether $\mathcal{T}$ should be added to $\mathcal{M}^{best}(\mathcal{T})$ (and considered to belong to the corresponding known object class $\mathcal{C}^{best}$) or rather to $\mathcal{M}^{new}(\mathcal{T})$ by adding a new object class $\mathcal{C}^{new}$ to the existing set of classes, the model probabilities $p(\mathcal{M}^{best}(\mathcal{T}) \mid \mathcal{T})$ and $p(\mathcal{M}^{new}(\mathcal{T}) \mid \mathcal{T})$ are calculated using the Bayes filter. The ratio of these probabilities yields the factor

$$\mathbf{K} = \frac{p(\mathcal{M}^{new}(\mathcal{T}) \mid \mathcal{T})}{p(\mathcal{M}^{best}(\mathcal{T}) \mid \mathcal{T})}, \tag{8.12}$$

that quantifies how much better the new model describes this object instance relative to the current best matching model. While large values for a threshold on $\mathbf{K}$ favor more compact models (fewer classes and lower data-fit), lower values lead to more precise models (more classes, in the extreme case overfitting the data). As alternative model selection criteria, one could use the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC), for example. However, during the experimental evaluation, the Bayes factor yielded accurate results and, thus, the comparison to alternative choices is left to future work.

Identify a Threshold on K

This subsection describes how to identify a threshold on $\mathbf{K}$, so that the system achieves a human-like class granularity, that is, a balance between model precision and compactness which is similar to how humans classify dynamic objects. To this aim, a training set that consists of instances of the classes *pedestrian*, *skater*, *cyclist*, *buggy*, and *kangaroo* was collected (see Figure 8.1 for example instances). First, the current best models and the fitted models of objects of the same class are compared by calculating the factors $\mathbf{K}$ according to Eq. 8.12. Subsequently, the same comparison is carried out with objects of different classes with randomly selected tracks. Table 8.1 gives the relative number of pairs for which different values of $\mathbf{K}$ – ranging from 1 to 20 – were exceeded. It can be seen that, e.g., for $\mathbf{K} \geq 4$, all pedestrians are merged to the same class (PED / PED), but also that there is a poor separation (40%) between pedestrians and skaters (PED / SKA). Given this set of tested thresholds $\mathbf{K}$, the best trade-off between precision and recall is achieved between $\mathbf{K} \geq 2$ and $\mathbf{K} \geq 4$. Therefore $\mathbf{K} \geq 3$ is chosen.

Interestingly, this threshold on $\mathbf{K}$ coincides with the interpretation of "substantial evidence against the alternative model" of Kass and Raftery. Note that fitting the threshold $\mathbf{K}$ to a labeled data-set does not render the approach a supervised one, since no specific class labels – which is the crucial information in this task – are supplied to the system. This step can rather be compared to learning regularization parameters in alternative models to balance data-fit against model complexity.

8.6 Segmentation and Tracking

The segmentation and tracking system takes the raw laser range data as input and produces the tracks $\mathcal{T}$ with associated laser segments $\mathbf{z}$ for the exemplar generation step. To this end, a Kalman filter-based multi-target tracker with a constant velocity motion model is employed. The constant velocity model is used since it makes only mild assumptions about the motion of targets of unknown type[52]. Practical experiments with other models – as the constant acceleration motion model – have been made without sensible changes in performance.

The observation step in the filter amounts to the problem of partitioning the laser range image into segments that consist in measurements on the same dynamic objects and to estimate their center. As the occurring objects are not known in advance no specific classifier can be trained. Instead, successive laser scans subtracted to extract beams that belong to dynamic objects. If the beam-wise difference is above the sensor noise level, the measurement is marked and grouped into a segment with other moving points in a pre-defined radius threshold θ_r.

Four different techniques to calculate the segment center have been compared: mean, median, average of extrema, and the center of a circle fitted through the segments points (for the latter the closed-form solutions from Arras et al. [2007] were taken). The last approach leads to very accurate results when tracking pedestrians, skaters, and people on kangaroo shoes but fails to produce good estimates with person pushing a buggy and cyclists. The mean turned out to be the smoothest estimator of the segment center.

Data association is realized with a modified nearest neighbor filter. It was adapted so as to associate multiple observations to a single track. This is necessary to correctly associate the two legs of pedestrians, skaters, and kangaroo shoes that appear as nearby blobs in the laser range image. Although more advanced data association strategies, motion models, or segmentation techniques have been described in the related literature, the proposed system is effective for its purposes.

[52] The constant velocity motion model well captures the differenced in dynamics of various object classes as pedestrians, skaters, cyclists, etc.

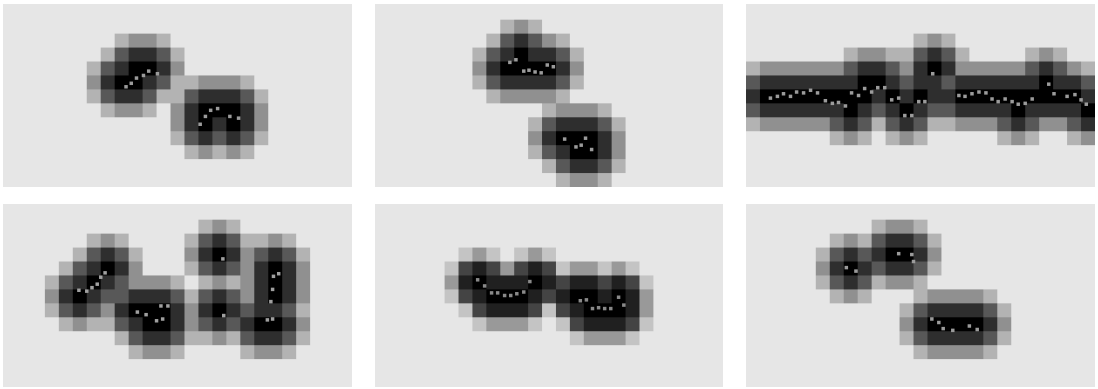


Figure 8.8: Top left to bottom right: Typical exemplars of the object classes *pedestrian*, *skater*, *cyclist*, *buggy*, *suitcase*, and *kangaroo*. The grid representations are invariant to rotation and translation thus the direction of motion is from left to right. Pedestrians and skaters have very similar appearance but differ in their dynamics. Pedestrians and subjects on kangaroo-shoes have similar dynamics but different appearances (mainly due to metal springs attached at the backside of the shoes). In the proposed approach both information are used to classify these objects.

8.7 Experiments

The proposed approach is experimentally evaluated with six object classes, namely *pedestrian* (PED), *skater* (SKA), *cyclist* (CYC), *person pushing a buggy* (BUG), *person pulling a suitcase* (SUI), and *people on kangaroo-shoes* (KAN) (see Figure 8.1 for examples). A set of 500 object tracks in total has been recorded. The sensor employed was a SICK LMS291 laser range finder with an angular resolution of 0.5 degree mounted at a height of $\sim$0.15 meter above ground. The tracks include walking and running pedestrians, skaters with small, wide, or no pace (just rolling), cyclists at slow and medium speeds, people pushing a buggy, pedestrians pulling a suitcase, and subjects on kangaroo shoes that walk slowly and fast. Note that pedestrians, skaters, and partly also kangaroo shoes have very similar appearance in the laser range data but differ in their dynamics. See Figure 8.8 for typical exemplars of each class. The implementation of the system runs in real-time on a typical desktop computer. The cycle time when using a 2 GHz single-core CPU for single tracks is around 43 Hz when sensor data are immediately available. Most time is spent in the k-means clustering algorithm (about 65%).

8.7.1 Supervised Learning Experiments

In the first group of experiments, the classification performance in the supervised case is tested. Each training set was composed of a single, typical track for each class including their labels *PED*, *SKA*, *CYC*, *BUG*, *SUI*, or *KAN*. The exemplar models were then learned from these single tracks. Based on the resulting prototype models, the remaining 494 tracks are classified. This experiment was repeated ten times with different training and testing tracks, the averaged results are shown in Table 8.2.

Pedestrians are classified correctly in 96.2% of the cases whereas 3.3% are incorrectly associated to the skater class. A manual analysis of these 3.3% revealed that the misclassification occurred typically with running pedestrians whose appearance *and* dynamics resemble those of skaters. A percentage of 0.5% were classified to be a person on kangaroo-shoes. All these tracks belonged to

Classes	PED	SKA	CYC	BUG	SUI	KAN
Pedestrian	96.2%	**3.3%**	0%	0%	0%	0.5%
Skater	2.4%	97.5%	0%	0%	0%	0.1%
Cyclist	0%	1.6%	98.4%	0%	0%	0%
Buggy	0%	0%	0%	97.2%	0%	2.8%
Suitcase	**6.4%**	0%	0%	0%	85.4%	**8.0%**
Kangaroo	**14.7%**	0%	0%	0%	0%	85.3%

Table 8.2: Classification rates in percent in the supervised experiment. Whereas the rows correspond to the ground truth object classes, the columns contain the obtained classification results. Mentionable misclassification rates are marked in bold. Most false classifications occurred for people walking slowly on kangaroo shoes having a pedestrian like gait.

running pedestrians, too. A rate of 97.5% for skaters with one track (0.1%) falsely classified as kangaroo-shoes and 2.4% classified as pedestrians is obtained. The latter group was found to skate slower than usual with a small pace, thereby resembling pedestrians. Cyclists are classified correctly in 98.4% of the cases. None of them was falsely recognized as pedestrians, buggies, suitcases, or person on kangaroo-shoes. But it appeared that the bicycle wheels produced measurements that resemble skaters taking big steps. This lead to a rate of 1.6% of cyclists falsely classified as skaters. A percentage of 97.2% of the buggy tracks were classified correctly. Only 2.8% were found to be a subject on kangaroo-shoes. In this particular case, the track contained measurements in which the front of the buggy was partially outside the field of view of the sensor with two legs of the person still visible. The pedestrians pulling a suitcase were correctly classified in 85.4%. Unfortunately, 6.4% were classified as pedestrians and 8.0% were considered to walk on kangaroo shoes. Typically, the people in these tracks walked with a lower pace, so that both legs and the suitcase appeared as the legs of a pedestrian or the kangaroo shoes. Subjects on kangaroo shoes were correctly recognized at a rate of 85.3% with 14.7% of the tracks falsely classified as pedestrians. The manual analysis revealed that the latter group consisted mainly of kangaroo shoe novices taking small steps and thus appearing like pedestrians.

In conclusion, it was found that, given the limited information provided by the laser range data and the high level of self-occlusion naturally occurring in this setting, the results indicate that the proposed exemplar models are expressive enough to discriminate between relevant object classes accurately. Misclassification typically occur at the boundaries where objects of different classes appear similarly or have similar dynamics.

8.7.2 Unsupervised Learning Experiments

In the second experiment object classes were learned in an unsupervised manner. Therefore, the entire set of 500 tracks from all six classes was presented to the system in a random order[53].

Each track was either assigned to the best existing class $\mathcal{C}^{best}$ so far or was taken as basis for a new class $\mathcal{C}^{new}$ according to the learning procedure described above. As can be seen in Table 8.3, eight object classes have been generated for the data set presented: one class for *pedestrians* (PED), one for *skaters* (SKA), two for *cyclists* (CYC), two for *buggies* (BUG), two for *suitcases* (SUI), and

[53] In detail, the raw laser range data is presented to the system. Since, the parameters of segmentation and tracking are kept fixed the same object tracks are obtained and examined.

Classes	PED	SKA	CYC	BUG	SUI	KAN	
class 1 (209)	**187**	5	0	0	3	17	"PED"
class 2 (114)	7	**107**	0	0	0	0	"SKA"
class 3 (41)	0	0	**41**	0	0	0	"CYC"
class 4 (23)	0	0	**23**	0	0	0	"CYC"
class 5 (26)	0	0	1	**25**	0	0	"BUG"
class 6 (23)	0	0	0	**23**	0	0	"BUG"
class 7 (38)	0	0	0	0	**38**	0	"SUI"
class 8 (23)	0	0	0	0	**23**	0	"SUI"
total (500)	194	112	65	48	64	17	

Table 8.3: Results of the unsupervised learning of object classes. The last column shows the manually added labels and the last row contains the total number of tracks of each class.

none for people walking on *kangaroo shoes* (KAN).

Class one (labeled PED[54]) contains 187 pedestrian tracks (out of 194), 5 skater tracks, 4 suitcase tracks, and 17 kangaroo tracks resulting in a true positive rate of 89.5%. Class two (labeled SKA) holds 107 skater tracks (out of 112) and 7 pedestrian tracks yielding a true positive rate of 93.9%. Given the resemblance of pedestrians and skaters, the total number of tracks and the extent of intra-class variety, this is an encouraging result that shows the ability of the system to discriminate objects that vary predominantly in their dynamics. Classes three and four (labeled CYC) contain 41 and 23 cyclist tracks respectively. No misclassification occurred. The classes five and six (labeled BUG), hold 25 and 23 buggy tracks with a bicycle track as the single false negative in class five. The last two classes, seven and eight (labeled SUI), consists of 38 and 23 tracks of pedestrians pulling a suitcase. Again no misclassification occurred. The representation of cyclists, buggies, and suitcases by two classes is due to the larger variability in their appearance and more complex dynamics. The discrimination from the other three classes is exact – no pedestrians, skaters, or subjects on kangaroo shoes were classified to be a cyclist or a buggy.

The system did not produce a specific class for subjects on kangaroo shoes as all instances of the latter class were included in the pedestrian class. The best known model for all 17 kangaroo tracks was always class one which has previously been created from a pedestrian track. This results in a false negative rate of 8.1% from the view point of the pedestrian class. This result confirms the outcome of the supervised experiment where the highest misclassification rate (14.7%) was found to be between pedestrians and subjects on kangaroo shoes (see Table 8.2).

8.7.3 Analysis of Track Velocities

The data set of object trajectories that was used in the experiments contains a high level of intra-class variation, like for example skaters moving significantly slower than average pedestrians or even pedestrians running at double their typical velocity. To visualize this diversity and to show that a simple velocity-based classification approach would yield unsatisfactory results, a velocity histogram

[54] The class labels have been assigned manually by inspecting the associated object classes occurring most often.

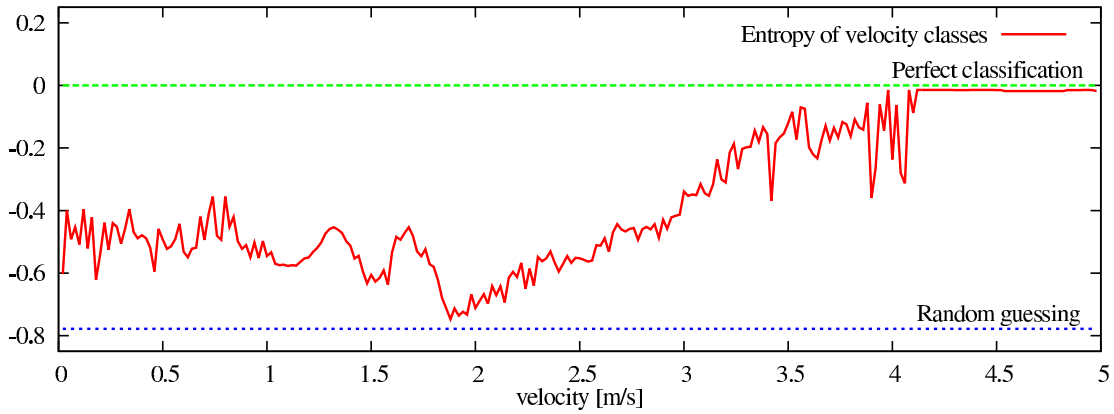


Figure 8.9: Analysis of the track velocities as alternative features for classification. While a high velocity is a strong indicator for a certain class (CYC), there is a higher confusion in the low and medium range.

for all six classes is calculated. For every velocity bin $\boldsymbol{v}_i$, the entropy

$$H(\boldsymbol{v}_i) \;=\; \sum_{j=1}^{6} \Big(p(c_j \mid \boldsymbol{v}_i) \;\log p(c_j \mid \boldsymbol{v}_i) \Big) \tag{8.13}$$

is calculated and the are result visualized in Figure 8.9. Note that the uniform distribution over six classes, which corresponds to random guessing, has an entropy of $6 \cdot (1/6 \cdot \log(1/6)) \approx -0.778$, which is shown by a straight, dashed line. As can be seen from the diagram, a high velocity is a strong indicators for a certain object class (here *cyclists* (CYC)) while there is a high level of confusion in the low and medium range.

8.7.4 Classification with a Mobile Robot

To demonstrate the practicability of the approach for a mobile sensor, an additional supervised and unsupervised experiment was carried out with a moving platform. A total of 18 tracks has been collected: 3 pedestrian tracks, 5 skater tracks, 4 cyclist tracks, and 6 suitcase tracks (kangaroo shoes and buggies were unavailable for this experiment). The robot moved with a maximal velocity of $0.75\,{}^{m}/_{s}$ and an average velocity of $0.35\,{}^{m}/_{s}$. A typical robot trajectory is depicted in Figure 8.10.

Classes	PED	SKA	CYC	SUI	BUG	KAN
Pedestrian	0.86	0	0	0	0	**0.14**
Skater	0.08	0.83	0	0	0	0.09
Cyclist	0.01	0	0.85	0.08	0.06	0.01
Suitcase	0.01	0	0	0.94	0.03	0.02

Table 8.4: Results of the supervised learning experiment using a mobile sensor platform. Whereas the rows contain the inspected object classes, the columns show the averaged classification probabilities $p(\mathcal{C}^k \mid \mathcal{T})_{k=1}^{6}$. All objects have been classified correctly. The highest confusion marked in bold occurred again between the object classes of pedestrians (PED) and people walking on kangaroo shoes (KAN).

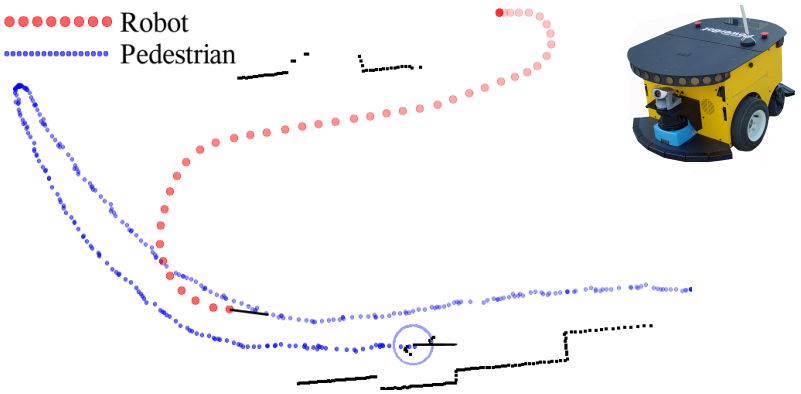


Figure 8.10: Trajectory of an ActivMedia PowerBot robot (shown as red dots) and a pedestrian (blue dots) over a sequence of 450 observations. In the unsupervised outdoor experiment the clustering result was exact: four classes were created autonomously – each containing the tracks of exactly one object category.

For the supervised experiment, the trained models from one of the supervised experiments in subsection 8.7.1 have been reused to classify the tracks collected from the moving platform. All objects were classified correctly by the moving robot. Table 8.4 contains the classification probabilities averaged over all tracks in the corresponding class. The last two columns contain the probabilities for the classes BUG and KAN, all being close to zero. The lowest classification probability in this experiment was a pedestrian track which still had the probability 0.73 of being a pedestrian.

In the unsupervised experiment, the tracks have been presented to the system in random order without prior class information. The clustering result compared to human classification was exact: four classes were created autonomously – each containing the tracks of exactly one object category.

8.8 Conclusions

In this chapter an unsupervised learning approach to the problem of tracking and classifying dynamic objects has been presented. In the proposed framework, the time-varying appearance of dynamic objects in planar 2D laser range data is represented by a probabilistic exemplar model in conjunction with a hidden Markov model for dealing with the dynamically changing appearance over time. The provided data is normalized to achieve translational and rotational invariance and thus all segments have a fixed position and orientation simplifying further processing steps. The normalization process is validated by inspecting the self-similarity of tracks from relevant object classes. Extensive real-world experiments including 500 recorded trajectories show that the model is expressive enough to yield high classification rates in the supervised case and that the unsupervised learning algorithm produces meaningful object classes consistent with the true underlying class assignments. Additionally, the system does not require any manual class labeling and runs in real-time.

In future research, it is planed to strengthen the interconnection between the tracking process and the classification module, i.e., to improve segmentation and data association given the estimated posterior over future object appearances. Furthermore, it should be studied how the behavior of the robot can be adapted given the estimated class labels of tracks. For example, a human slowly approaching the robot might want to communicate with the robot while a quickly moving human or cyclist does not.

9 On-line Learning Of Target Appearance for 3D Tracking

People tracking is a key component for robots that are deployed in populated environments. Previous works have used cameras, 2D, and 3D range finders for this task. In this chapter, state of the art in 3D people detection and tracking is advanced by combining a novel multi-cue person detector for RGB-D data with an on-line detector that learns individual target appearance models. This is a new aspect for range-based target tracking which usually deals with objects of identical appearance. The two detectors are integrated into a decisional framework with a multi-hypothesis tracker that controls on-line learning through a track interpretation feedback. Simultaneously, tracking is improved by a joint likelihood data association based on motion and appearance.

For on-line learning, a boosting approach that benefits from the richness of the data by using four types of RGB-D features is proposed. A depth-informed confidence maximization search bounded by state uncertainty predictions is introduced to find the most likely target position in 3D. The approach is general in that it neither relies on background learning nor a ground plane assumption.

Evaluation has been performed on data collected in a populated indoor environment using a setup of three Microsoft Kinect sensors with a joint field of view. Analyzed with a tracking performance metric, the results demonstrate reliable 3D tracking of people in RGB-D data with 50% less target misses. Furthermore, it is shown that the presented framework avoids drift of the on-line detectors. The overall tracking performance is increased by 16% through improved detection and data association.

This chapter is structured as follows. Introduction and related work are presented in section 9.1 and section 9.2, respectively. The a priori people detector is briefly summarized in section 9.3. Section 9.5 presents the on-line AdaBoost approach that allows to learn target-specific appearance models in RGB-D data. The integration of this learning procedure into the multi-hypothesis tracking system with track interpretation feedback to control learning is described in section 9.6. In section 9.7 the experiments and results are presented. Finally, section 9.8 concludes the chapter.

9.1 Introduction

People detection and tracking is an important and fundamental component for interactive systems, intelligent vehicles, and robots that share their space with humans. Especially latter must be able to track people in 3D space as most environments are composed of more than one level. Popular sensors for this task are stereo cameras and 3D range finders. Cameras have the advantage of capturing rich scene information in a high resolution and frame rate but work in a limited band of illumination conditions especially from a mobile observer. If detection is performed in 2D camera space the projection into 3D is another critical issue. Range finders, on the other hand, provide quite accurate detections in 3D. They typically have a large field of view, and work well with vibrations. Furthermore, they are relatively robust against illumination changes since they emit the energy necessary to perform a measurement by themselves. A major drawback of range finders is that targets usually have identical appearance. However, while both sensing modalities have advantages and

drawbacks, their distinction may become obsolete with the availability of affordable and increasingly reliable RGB-D sensors that provide both image and range data.

To take advantage of the rich RGB-D data an on-line boosting approach using four types of RGB-D features is proposed to learn target-specific appearance models. During partial occlusions or in case the a priori detector fails a depth-informed confidence maximization search in 3D space is employed to find the most likely target position. The search is bounded by the state uncertainty and centered around the target position predicted with a motion model. Learning the on-line detectors is controlled via track interpretation feedback provided by the multi-hypotheses tracker (MHT) to avoid drift. On the other hand, the learned model can be applied to guide the hypotheses generation within the MHT by a joint likelihood data association respecting motion and appearance.

9.2 Related Work

Many researchers in robotics have addressed the issue of detection and tracking people in laser range data. Early works from Fod et al. [2002] and Schulz et al. [2003] were based on 2D data in which people have been detected using ad-hoc classifiers that find moving local minima in the scan. A learning approach has been taken by Arras et al. [2007], where a classifier for 2D point clouds has been trained by boosting a set of geometric and statistical features. As there is a natural performance limit with a single layer of 2D range data, multiple co-planar 2D laser range scanners have been used in Gidel et al. [2008], Carballo et al. [2008], and Martínez Mozos et al. [2010]. In the latter the authors apply boosting on each of three layers and use a probabilistic scheme to combine the three classifiers in a flattened 2D space.

People detection and tracking in 3D range data is a rather new problem with little related work. Navarro-Serment et al. [2010] collapse the 3D scan into a virtual 2D slice to find salient vertical objects above ground. They align a window to the principal data direction, compute a set of features, and classify pedestrians using a set of SVMs. Bajracharya et al. [2009] detect people in 3D point clouds from stereo vision by processing vertical objects and considering a set of geometrical and statistical features of the cloud based on a fixed pedestrian model. Unlike these works that require a ground plane assumption, Spinello et al. [2010a] overcome this limitation with a layered approach that learns a bank of specialized part classifiers that vote into a common space. Detection hypotheses then emerge as local maxima in that space. The method was recently extended in Spinello et al. [2011] by the combination with a top-down verification procedure that learns both an optimal features and the volume tessellation. The combined approach achieved equal error rates (EER) of more than 93% using a Velodyne 3D laser range finder to detect people in up to 20 meters distance. In this chapter, the bottom-up top-down approach is deployed as people detector on RGB-D data as described in section 9.3.

In the computer vision literature, the problem of detecting, tracking, and modeling humans has been extensively studied by Dalal and Triggs [2005], Leibe et al. [2005], Felzenszwalb et al. [2008], and Enzweiler and Gavrila [2009]. A major difference to range-based systems is that the richness of image data makes it straightforward to learn target appearance models. For this reason, visual tracking systems can achieve good results with methods as simple as independent particle filters with nearest-neighbor data association. Breitenstein et al. [2011], for example, support tracking by exploiting an appearance-based likelihood term for data association. They can even replace a detector in a bootstrap fashion as shown in Ramanan et al. [2007]. Dense depth data from stereo are used by Beymer and Konolige [1999] to support foreground segmentation in an otherwise vision-based people detection and tracking system. They use a set of binary person templates to detect people in images and demonstrate multi-person tracking with learned appearance-based target models. By

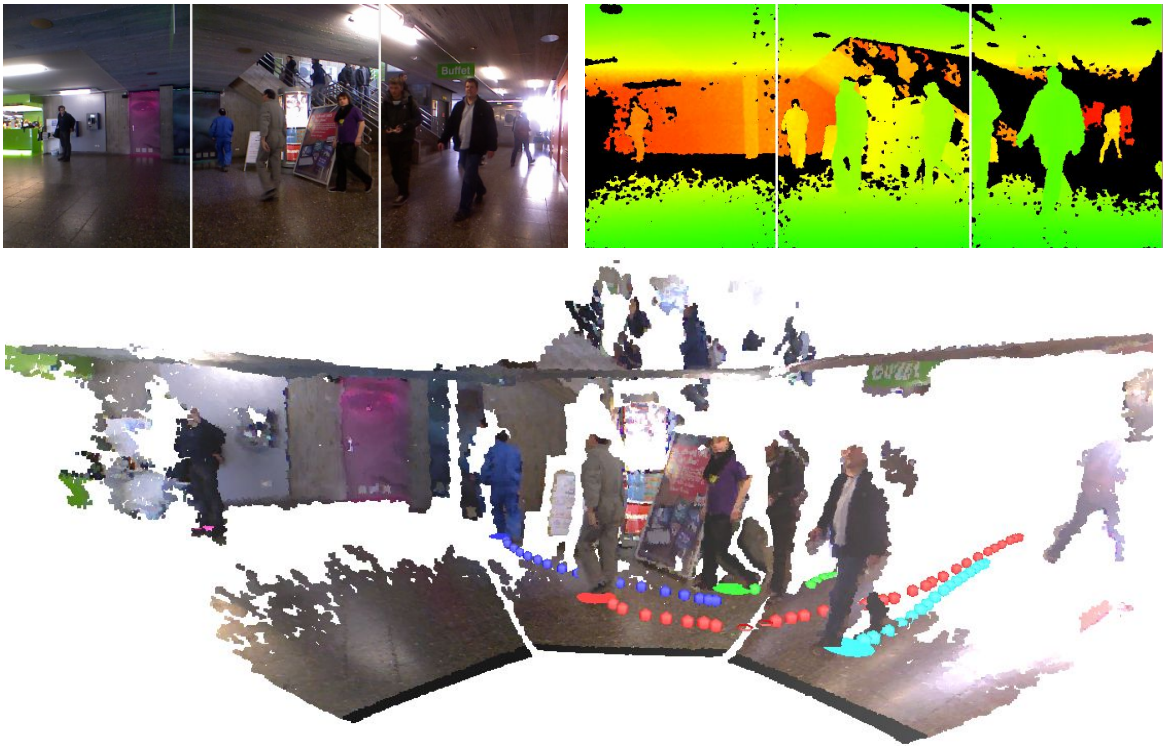


Figure 9.1: People tracking in RGB-D data. The top pictures show the three color and depth images, below the 3D point cloud. The data was collected in the lobby of a large university canteen at lunch time with a setup joining the views of three Kinect sensors. The colored disks and dots in the point cloud show the positions and trajectories of five tracked persons.

learning temporal color-based appearance models data association is supported in Bahadori et al. [2005]. Both model the background and utilize disparity information to improve segmentation. Using a neural network on dense stereo Harville et al. [1998] detect frontal faces and learn skin color and head size statistics to re-detect people in case of detector failures, occlusions, or when they have left the scene. A local search centered around the head in the previous frame is applied to find the most likely head position. The work of Leibe et al. [2008] and Ess et al. [2009a] detect people in intensity images and track them in 3D. In Enzweiler et al. [2010] a stereo system for combining intensity images, stereo disparity maps, and optical flow is used to detect people and handle partial body occlusions. It employs a mixture of local multi cue expert component classifiers for robustly fusing the different information. Multi-modal detection and tracking of people is also performed in Spinello et al. [2010b] where a trainable 2D range data and camera system is presented.

In the work of Song et al. [2010] laser and vision data is fused to disambiguate interacting people walking close to each other. Using multiple independent particle filters pedestrian positions are projected into the image space to extract visual information used to train on-line classifiers. For interacting targets the learned classifiers support data association by weighting the particles using a cross-correlation like measure based on the trained and observed image patches. In Babenko et al. [2011] a single object is tracked in video data. Given its location in the first frame on-line multiple instance learning is used to train adaptive appearance models. By presenting sets of positive examples (called "bags") ambiguity is passed on to the learning algorithm avoiding significant drift during partial occlusions. Munaro et al. [2012] propose a multi-people tracking algorithm in RGB-D data. Assuming a ground plane 3D sub-clustering allows to efficiently detect people very close to each other

or to background. For each track an on-line classifier based on AdaBoost learning color appearance is maintained to support joint likelihood global nearest neighbor data association.

This work advances the state of the art in the relatively new problem of detecting and tracking people in RGB-D data in the following aspects. A generic a priori person detector is combined with an on-line learned person-specific detector and a multi-target multi-hypothesis tracker (MHT), able to estimate the motion state of multiple people in 3D. Learning individual target models is a new aspect to range data-based object tracking that usually deals with targets of identical appearance. To this end, the on-line learning method from Grabner and Bischof [2006] is adapted to RGB-D data employing features on all sensory cues. Furthermore, a novel framework to integrate the two detectors and the tracker that involves a track interpretation feedback to control learning is presented. It is shown, how the MHT machinery to associate and interpret observations and tracks can be used to control the on-line learning. This enables the system to bridge gaps of mis-detections of the a priori detector and to handle target occlusions while avoiding drift of the on-line detectors. Finally, quantitative results are given using the CLEAR MOT performance metric. Unlike the above mentioned related works that integrate multiple sensory modalities, image and range data are considered as equally important cues for detection, tracking, and target appearance model adaptation. Further a novel integration framework to effectively combine a tracker with on-line learned target classifiers is presented.

9.3 Detecting People in RGB-D Range Data

In this section the a priori people detector is summarized briefly. It employs a novel RGB-D person detector called Combo-HOD (Combined Histograms of Oriented Depths and Gradients) recently proposed by Spinello and Arras [2011]. The method takes inspiration from Histogram of Oriented Gradients (HOG) introduced by Dalal and Triggs [2005] and combines the HOG detector in the color image with a novel approach in the depth image called Histograms of Oriented Depths (HOD).

Since RGB-D data contains both color and depth information, the Combo-HOD detector combines the two sensory cues. HOD descriptors are computed in the depth image and HOG descriptors are computed in the color image. They are fused on the level of detections via a weighted mean of the probabilities obtained by a sigmoid fitted to the SVM outputs. HOD includes a depth-informed scale-space search in which the used scales in an image are first collected and then tested for compatibility with the respective depth. This test is made particularly efficient by the use of integral tensors, an extension of integral images over several scales. This strategy dramatically reduces the number of descriptors computed in the image at improved detection rates. In case no depth information is available, the detector gracefully degrades to a standard HOG detector.

In order to train a robust detector and to achieve a good separation between object and background, a large number of negative examples is needed. In practice this can be a prohibitive memory-expensive task. Therefore, the authors propose to organize the training phase in two separate rounds. First, sampled negative image patches combined with the positive samples are used to train an initial detector. In the second round, this initial detector is used to search for false positives, exhaustively. Adding these false positives to the original negative set enables to re-train the improved final detector. The result is a more robust people detector trained only with the negative information needed to reduce the training error.

The output of the detector in each step are the position and size of all targets in 3D space and the center and size of the bounding boxes in the color and depth images. They are the observations $\mathbf{z}_i(t)$ that constitute the set of M_t observations $\mathcal{Z}(t)$ at time index t. More information on people detection in RGB-D data can be found in Spinello and Arras [2011].

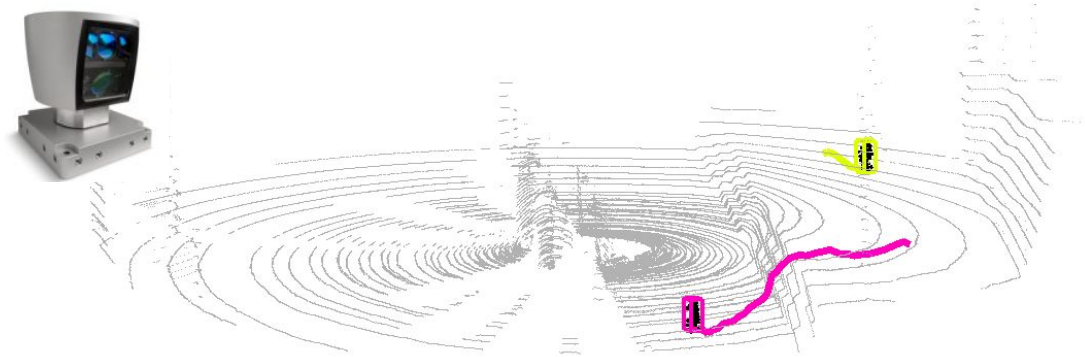


Figure 9.2: Tracking two persons in 3D using range data collected with a Velodyne HDL 64E S2 laser scanner placed in the center of the displayed scene. The person marked in magenta is descending a flight of stairs. The z axis in the visualization is magnified for clarity.

9.4 Tracking People in 3D

Tracking people in 3D requires an extended state space that integrates the z dimension, thus $\mathbf{x}_t = (\ x_t\ \ y_t\ \ z_t\ \ \dot{x}_t\ \ \dot{y}_t\ \ \dot{z}_t\ )^T$ defines the position and velocity of a pedestrian at time t and Σ_t the corresponding 6×6 covariance matrix estimate. Human motion prediction based on the previous state estimate $\mathbf{x}_{t-1}$ using the *continuous white noise acceleration model* (or *constant velocity motion model*) is now defined as

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \ = \ \mathcal{N}(\mathbf{x}_t;\ A_t\,\mathbf{x}_{t-1}, A_t\,\Sigma_{t-1}\,A_t^T + Q_t), \tag{9.1}$$

with A_t being the constant velocity state transition matrix and Q_t the process noise Matrix defined as

$$A_t = \begin{pmatrix} 1 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta t \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \text{and } Q_t = \begin{pmatrix} 1/3\Delta t^3 & 0 & 0 & 1/2\Delta t^2 & 0 & 0 \\ 0 & 1/3\Delta t^3 & 0 & 0 & 1/2\Delta t^2 & 0 \\ 0 & 0 & 1/3\Delta t^3 & 0 & 0 & 1/2\Delta t^2 \\ 1/2\Delta t^2 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 1/2\Delta t^2 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 1/2\Delta t^2 & 0 & 0 & \Delta t \end{pmatrix} \tilde{\sigma}, \tag{9.2}$$

respectively. The entries of Q_t represent the acceleration capabilities of a human modeled with the continuous-time process noise intensity $\tilde{\sigma}$. For more details on the continuous white noise acceleration model see [Bar-Shalom et al., 2002, p. 269–270].

During data association the measurement likelihood $\mathcal{N}\big(\mathbf{z}_i(t), \mathbf{x}_j(t-1)\big)$ specifies how well an observation $\mathbf{z}_i(t)$ describes an existing track $\mathbf{x}_j(t-1)$. It is assumed that this likelihood has a Gaussian pdf centered around the measurement prediction $\hat{\mathbf{z}}_j(t)$ with innovation covariance matrix $S_{i,j}(t)$, both calculated using the motion prediction in Eq. 9.1. Finally, the measurement likelihood of a single matched observation is then defined as $\mathcal{N}\big(\mathbf{z}_i(t)\big) := \mathcal{N}\big(\mathbf{z}_i(t);\ \hat{\mathbf{z}}_j(t),\ S_{i,j}(t)\big)$.

Tracking people in 3D using data from a Velodyne laser scanner as shown in Figure 9.2 is explained in more detail in Spinello et al. [2011].

9.5 On-line Boosting

The detector described above learns a generic person model from a priori data labeled in advance. However, this section describes the use of on-line boosting to learn target appearance models in RGB-D data, later used to guide data association in the tracking system.

Boosting is a widely used technique to improve the accuracy of learning algorithms. Given a set of annotated training examples $\mathcal{E} = \{(\varepsilon_1, l_1), \ldots, (\varepsilon_N, l_N)\}$ with labels $l_i \in \{-1, +1\}$, a strong classifier $H(\varepsilon)$ is computed as linear combination of a set of weighted hypotheses called weak classifiers $h(\varepsilon)$. The only requirement to the weak classifiers is that their accuracy is better than a random guessing. The discrete AdaBoost algorithm by Freund and Schapire [1997] belongs to the most popular boosting algorithms. The method trains weak classifiers from labeled training samples (ε_j, l_j), initialized with uniform weights ω_j associated to each of them. Learning is done in rounds where the weights are updated based on the mistakes of the previous weak learner. By increasing the weights of the wrongly classified samples the algorithm focuses on the difficult examples. The AdaBoost algorithm is outlined in algorithm 1.

On-line boosting, initially proposed by Oza and Russell [2001], processes each training instance "on arrival" without the need of storage and reprocessing, and maintains a current hypothesis that reflect all the training samples seen so far. The on-line boosting approach has been applied for object detection while tracking by Grabner and Bischof [2006]. The approach proposed in this chapter builds upon the latter to develop an on-line people detector in RGB-D data.

9.5.1 Updating the Weak Classifiers

Unlike the off-line approach to boosting, the on-line algorithm (shown in algorithm 5) presents positive and negative training samples only once and discards them after training. The weak classifiers have thus to be updated in an on-line fashion each time a new training sample is available. For updating the weak classifiers, any on-line learning algorithm can be employed to estimate the probability distributions of positive and negative samples and generate a hypothesis. However, as the difficulty of the samples is not known in advance the computation of the weight distribution of the samples is a critical issue. The basic idea of on-line boosting is that the weight of a sample (called the importance weight λ in this context) can be estimated by propagating it through a fixed chain of weak classifiers as proposed by Oza and Russell [2001]. If the sample is misclassified, λ is increased proportional to the error of the weak classifier. Therefore, the importance has the same effect as the adapted weights in the off-line approach (see algorithm 1). The error of the i-th weak classifiers is estimated from the summed weights of the correctly ($\lambda_i^{correct}$) and wrongly (λ_i^{wrong}) classified samples and calculated as

$$e_i \;=\; \frac{\lambda_i^{wrong}}{\left(\lambda_i^{wrong} + \lambda_i^{correct}\right)}. \tag{9.3}$$

The weak classifiers with the smallest errors are selected (see alg. 5, line 12) by the feature selectors as explained in the next subsection.

9.5.2 On-line-boosting for Feature Selection

For the purpose of learning target models during tracking, Grabner and Bischof [2006] propose to employ *feature selectors*. The main idea is to apply on-line boosting not directly to the weak classifiers but to the selectors. A selector h^{sel} selects the best weak classifier from a pool of M weak learners $\mathcal{F}$ with 'best' being defined by the lowest error e_i defined in Eq. 9.3.

Algorithm 5: On-line boosting algorithm for feature selection.

Input : On-line received new training example (ε, l) with annotation $l \in \{-1, +1\}$.
Strong classifier H^{t-1} consisting of N selectors each having M weak classifiers.
Weights of the correctly ($\lambda_{n,m}^{correct}$) and wrongly ($\lambda_{n,m}^{wrong}$) classified samples.
Output : Updated strong classifier H^t.
Variables: Importance weight λ.

$\lambda \leftarrow 1$;
`/* update selectors */`
for $n \leftarrow 1$ **to** N **do**
  `/* update weak classifiers */`
  for $m \leftarrow 1$ **to** M **do**
    $h_{n,m} \leftarrow update(\varepsilon, l, \lambda)$;
    `/* estimate errors */`
    if $h_{n,m}(\varepsilon) = l$ **then**
      $\lambda_{n,m}^{correct} \leftarrow \lambda_{n,m}^{correct} + \lambda$;
    else
      $\lambda_{n,m}^{wrong} \leftarrow \lambda_{n,m}^{wrong} + \lambda$;
    end
    $e_{n,m} \leftarrow {}^{\lambda_{n,m}^{wrong}}/_{(\lambda_{n,m}^{correct} + \lambda_{n,m}^{wrong})}$;
  end
  `/* select weak classifier with lowest error → feature selection */`
  $m^+ \leftarrow \arg\min_m(e_{n,m})$;
  $e_n \leftarrow e_{n,m^+}$;
  $h_n \leftarrow h_{n,m^+}$;
  `/* calculate voting weight */`
  $\alpha_n \leftarrow {}^1/_2 \log({}^{1-e_n}/_{e_n})$;
  `/* update importance weight */`
  if $h_n(\varepsilon) = l$ **then**
    $\lambda \leftarrow {}^{\lambda}/_{2(1-e_n)}$;
  else
    $\lambda \leftarrow {}^{\lambda}/_{2e_n}$;
  end
  `/* replace worst weak classifier */`
  $m^- \leftarrow \arg\max_m(e_{n,m})$;
  $\lambda_{n,m^-}^{correct} \leftarrow 1$;
  $\lambda_{n,m^-}^{wrong} \leftarrow 1$;
  $h_{n,m^-} \leftarrow$ sample new weak classifier;
end
`/* return updates strong classifier, the strong classifier is given by: */`
`/*` $H^t(\varepsilon) = \text{sgn}\left(\sum_{n=1}^{N} \alpha_n \ h_n(\varepsilon)\right)$, `with` $h_n(\varepsilon_n) \in \{-1, +1\}$ `*/`
return H^t;

Figure 9.3: On-line boosting algorithm for feature selection.

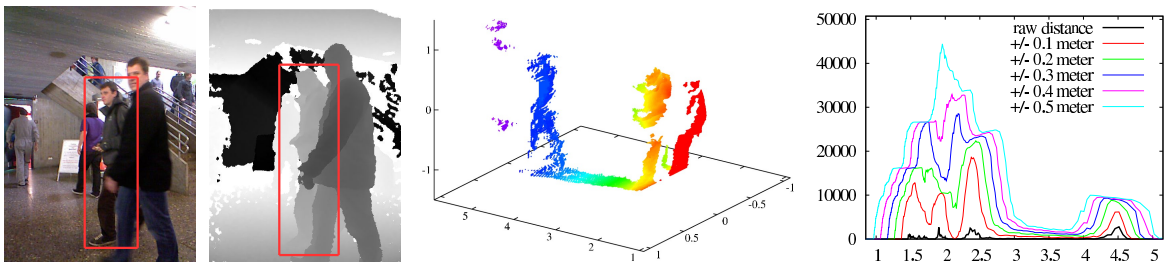


Figure 9.4: Method to estimate the correct 3D position of a person given the corresponding regions in 2D image space. *Left:* A person is marked with red bounding boxes in the RGB and depth image, respectively. *Middle:* As a rectangular region badly describes the contour of a person, thus the extracted 3D point cloud contains many points of other people and background. *Right:* Histograms of different bin size showing the number of points with a specific range are tested to estimate the correct distance of the person. As the ground truth distance is 2.3 meters the histogram with 0.1 meter tolerance performs best.

With the number of selectors N being a fix parameter, the following procedure is repeated for all selectors when a new sample (ε, l) arrives: First, all weak classifiers are updated and the best one, denoted with index m^+, is selected

$$h_n^{sel}(\varepsilon) \;=\; h_{m^+}^{weak}(\varepsilon), \tag{9.4}$$

with $m^+ = \arg\min_m(e_{n,m})$ and $e_{n,m}$ defined as in Eq. 9.3 with subscript n, m for i. Then, the voting weight $\alpha_n = \frac{1}{2} \cdot \ln(\frac{1-e_n}{e_n})$ is computed where $e_n = e_{n,m^+}$ and the updated importance weight λ is propagated to the next selector h_{n+1}^{sel}. Similar to AdaBoost, λ is increased if h_n^{sel} predicts ε correctly and decreased otherwise as shown in algorithm 5, line 16. The strong classifier is finally obtained by computing the confidence value $\kappa(\varepsilon)$ as a linear combination of the N selectors and applying the signum function,

$$\kappa(\varepsilon) \;=\; \sum_{n=1}^{N} \big(\alpha_n \cdot h_n^{sel}(\mathbf{x})\big), \qquad H(\varepsilon) \;=\; \mathrm{sgn}(\kappa(\varepsilon)). \tag{9.5}$$

Unlike the off-line version, the on-line procedure creates an always-available strong classifier in an any-time fashion. In order to increase the diversity of the classifier pool $\mathcal{F}$ and to adapt to appearance changes of the targets, at the end of each iteration, the worst weak classifier is replaced by a new one randomly chosen from $\mathcal{F}$ (see alg. 5, line 24).

9.5.3 RGB-D Features

Taking advantage of the richness of RGB-D data four different types of features that correspond to the weak classifiers are computed:

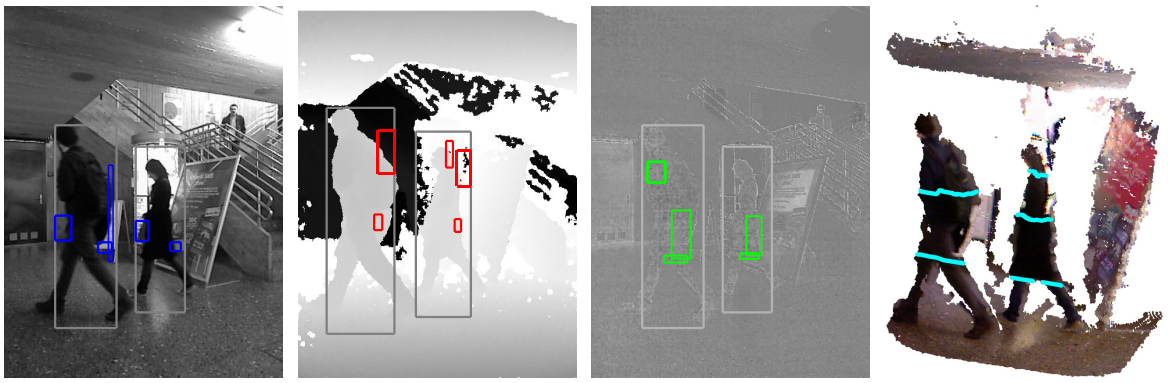
(1) Haar-like features in the intensity image[55],

(2) Haar-like features in the depth image,

(3) illumination agnostic *Lab* color features in the RGB image, and

(4) geometrical features defined in the extracted 3D point cloud.

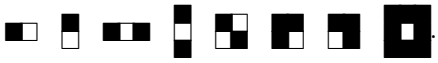
[55] The intensity image is achieved by converting the initial RGB values into grayscale using coefficients that represent human perception of colors, in particular that humans are more sensitive to green and least sensitive to blue (gray $=$ 0.299 R $+$ 0.587 G $+$ 0.114 B).



(a) Haar-like features on the intensity image. (b) Haar-like features on the depth image. (c) Color features on the *Lab* image. (d) Geometrical features on 3D point cloud.

Figure 9.5: Bounding boxes of two detected persons in the (a) RGB, (b) depth, and (c) *Lab* images. The best features of each on-line detector are marked with colored rectangles and colored dots, respectively. Haar-like features on the intensity image are shown in blue, Haar-like features on the depth image in red. The *Lab* color features are depicted in green. Geometric features are marked with cyan dots on the extracted 3D point cloud (d).

In more detail, the following Haar-like features defined by Viola and Jones [2001] have been used:



The features encode color or depth changes in the image and are calculated by summing positive (marked in black) and negative (white) pixel values within a sampled image region. To speed up the calculation integral images are employed. The *Lab* features are computed by summing up the intensity values in the $a*$ or $b*$ space under a sampled area, respectively. The advantage of the *Lab* color model is that features in $a*$ or $b*$ space can compactly and robustly subsume entire RGB histograms. The rectangular areas in which the Haar-like and *Lab* features are computed have randomized positions and scales in the bounding box associated to each target. Selected geometrical features defined in subsection 2.3.1 and subsection 2.3.2 adapted to 3D are calculated using a single line of range readings sampled in the depth image within the bounding box of the target. A total of M features is computed where the initial number of features is $M/4$ for all types. Given the above mentioned adaptation mechanism, their relative numbers can change to best describe a target dynamically. The randomized features positions and scales are kept fix over the lifetime of a target w.r.t. its bounding box that changes due to motion (up to the weak feature that get replaced by new samples). In an example, the best five features of each type are shown in Figure 9.5.

Further features can be defined to improve the description of various human appearances. Interesting approaches include the Binary Robust Appearance and Normals Descriptor (BRAND) of Nascimento et al. [2012], that efficiently combines appearance and geometric shape information from RGB-D images, and is largely invariant to rotation and scale transform. Additionally, the distribution of features in the bounding box, their preferred size, and the geometric relations among features can be learned either from training data or on-line as in the attributed relational feature graph proposed by Tang and Tao [2008].

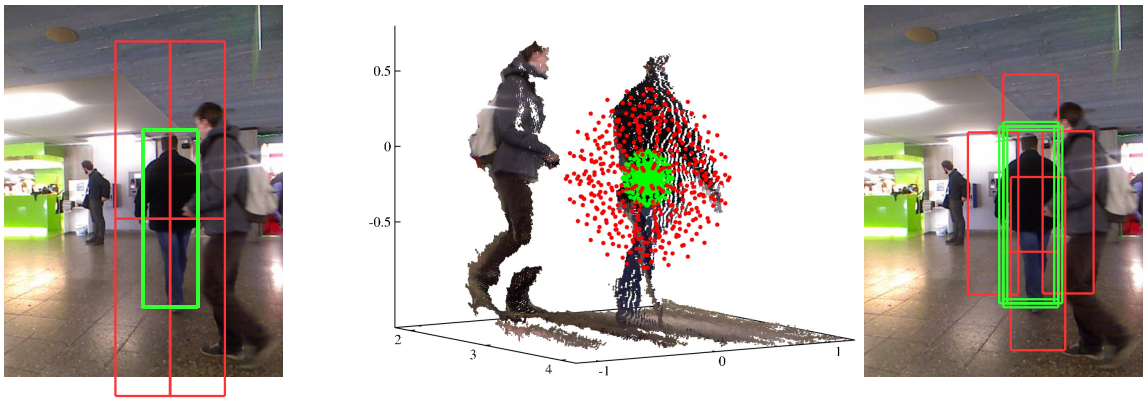


Figure 9.6: Strategies to sample bounding boxes with positive and negative features, respectively. *Left:* In Grabner and Bischof [2006] the original bounding box (shown in green) is used to compute the positive features. Four regions – shifted by half of the size (shown in red) – are used to sample negatives. *Middle:* The proposed approach first performs sampling in 3D space based on the target's position estimate and uncertainty. *Right:* The sampled 3D positions are back-projected into the camera space to obtain the bounding boxes with positive and negative features (the geometrical meaning is maintained).

9.5.4 On-line Boosting for Tracking

On-line boosting enables a tracker to continuously update a target (appearance) model to optimally discriminate it from the current background and other targets. As stated by Avidan [2004] this is a formulation of tracking as a classification problem. Briefly, classification is implemented by a confidence maximization procedure around the current tracking region. The region is obtained as the bounding box of the previous detection. All features within the region are considered the positively labeled foreground samples. The negative samples are obtained by sweeping the bounding box over a local neighborhood. The classifier is then evaluated at each sweep position of this neighborhood yielding a confidence map whose maximum is taken as the new position of the tracking region. The classifier and therefore the target model is updated in this region to adapt to appearance changes and the process is continued. An example evolution of the confidence values over time can be seen in Figure 9.13. The individual processing steps needed during the entire life-cycle of a target are explained in more detail in the following subsections.

Classifier Initialization

In most related work, object detected is performed in 2D image space. To initialize the target-specific appearance model the bounding box provided by the detector serves as positive training sample. Furthermore, negative training examples are sampled within a region of interest by shifting the bounding box around the center of detection as illustrated in Figure 9.6 (*left*). While this approach is sufficient to sample negative information no target specific information is used to guide the sampling procedure. However, in the proposed approach detection is performed in 3D and provides a Gaussian state estimate for each target. Hence, positive and negative training examples are sampled in 3D space using mixtures of sigma points. The sampled positions are finally back-projected into 2D image space to calculated the corresponding features. The method is visualized in Figure 9.6 (*middle, right*).

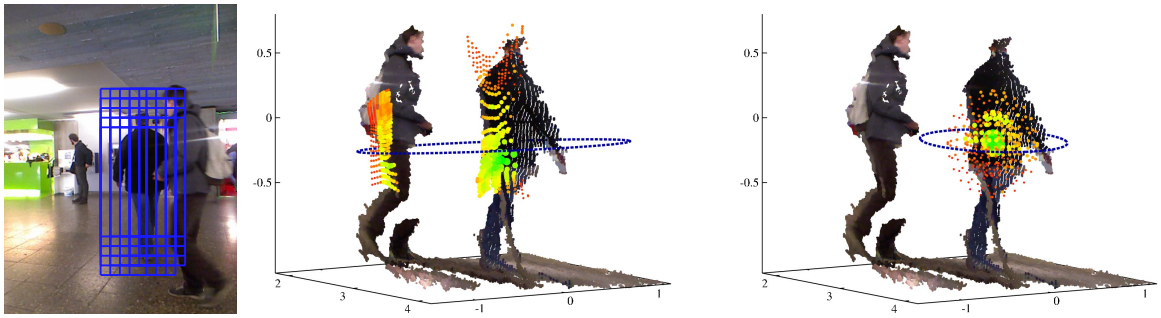


Figure 9.7: Strategies to evaluate the classifier H^{t-1} on new data received at time t. *Left:* In Grabner and Bischof [2006] the classifier is evaluated by shifting the bounding box of the previous detection in a region of interest (illustrated by blue rectangles). *Middle:* The dots display the projected 3D positions and confidence values $\kappa(\vec{\varepsilon})$ colored in red (low confidence) to green (high confidence). The dashed ellipse denotes sample mean and covariance. *Right:* In the proposed approach, the classifier is evaluated in an area defined by the predicted state and uncertainty. The sample distribution reflect potentials positions given the previous state and motion dynamics of the target.

Feature Update

Unlike Grabner and Bischof [2006] where the new region is only bootstrapped from the previous detection, in the presented approach the bounding box position of the a priori detector serves to recenter the on-line detector. This strategy avoids a key problem of on-line adaptation namely drifting of the model to background, clutter, or other targets.

Confidence Evaluation

To evaluate a learned classifier at multiple positions a so called confidence map can be created by shifting the tracking region in the 2D image space (see Figure 9.7, left). By identifying the best position[56] the target state is updated making an exact calculation of the 3D position crucial for tracking. However, computing the unique corresponding 3D position given a bounding box in 2D image is impossible. And even when using RGB-D data it is a extremely hard task as shown in Figure 9.4. Therefore, a similar strategy as described above is employed. Instead of shifting the tracking region in 2D multiple position hypotheses are sampled in 3D according to the target's predicted state and uncertainty obtained via a motion model. Subsequently, the likelihoods of the hypotheses are calculated using the classifier. Finally, instead of defining the best hypothesis as new position of the target the weighted sample mean and covariance are feed back into the tracking system. Figure 9.7 (right) demonstrates the confidence search in 3D.

Depth-Informed Confidence Evaluation

As it is impossible to know the size (also called *scale*) of a person in image space a priori many detection algorithms relying on a scale-space search to find objects at various distances in the image. In the computer vision community uninformed search heuristics such as image pyramids are employed to consider multiple scales at a fine resolution. However, in Spinello and Arras [2011] a depth-informed

[56] The best position is defined to have the feature descriptor with the highest confidence value returned by the classifier.

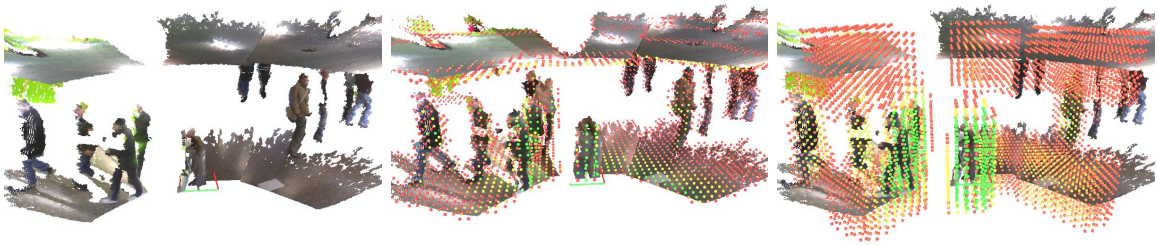


Figure 9.8: *Middle:* The density of the 3D point cloud is estimated in a three dimensional regular grid by counting the points collected in each cell. *Right:* Given the density the number of support points contributing to the volume representing a position hypothesis. For visibility reasons, cells having less than 10% of the maximum density or maximum number of support points are not drawn.

scale space search is proposed which is both more efficient and at the same time more accurate as the search is guided by depth-informed scale estimates.

Summarized briefly, search is improved by discriminating compatible scales at each position in the depth image. Therefore, the average human height is computed from the training data set, in which ground position and height of each sample is accurately annotated. This information is used to compute a scale-depth regression for each pixel of the depth image to generate a *scale map* from which a list of scale hypotheses is generated. The list contains only those scales that are compatible with the presence of people in the image.

This approach is adapted to derive a depth-informed confidence evaluation. Given a 3D point cloud the density of points in the monitored environment is approximated in a three dimensional grid representation called *density map*. The map is generated by counting the points within the volume of each cell. As visualized in Figure 9.8 (middle) static structures – like floors, walls, and ceilings – and also people cause a high density of range readings. But due to partial occlusions areas that are obviously covered by people are empty. To account for partial occlusions, the map is smoothed using a kernel representing the average size of a person learned from training data. The new representation shown in Figure 9.8 (right) encodes the number of points supporting a position hypothesis. During the confidence search explained above positions with empty support are rejected. Similar to the improved scale space search using depth information proposed by Spinello and Arras [2011] the sampled position of the 3D confidence map are analyzed in advance to reject impossible position hypotheses.

9.6 Integration into the Multi-Hypothesis Tracker

This section describes how the on-line detector is integrated into a Kalman filter based multi-hypothesis tracking framework (MHT). An extensive introduction into the MHT is given in Chapter 3. Hereafter, a brief summary is provided and the aspects that change are discussed in more details.

In short, the MHT algorithm hypothesizes about the target states by considering all statistically feasible assignments between observations and tracks and all possible interpretations of observations as false alarms or new track and tracks as matched, occluded or obsolete. Thereby, the MHT handles the entire life-cycle of tracks from creation and confirmation to occlusion and deletion.

Formally, let $\mathbf{x}_t = (\ x_t \ \ y_t \ \ z_t \ \ \dot{x}_t \ \ \dot{y}_t \ \ \dot{z}_t\ )^T$ be the filtered state of a track $\mathbf{x}_j(t)$ at time t with position and velocity information in 3D and Σ_t its associated 6×6 covariance. Let $\mathcal{Z}(t) = \{\mathbf{z}_i(t)\}_{i=1}^{M_t}$ be the set of M_t observations which in the concrete case is the set of detected people in RGB-D data.

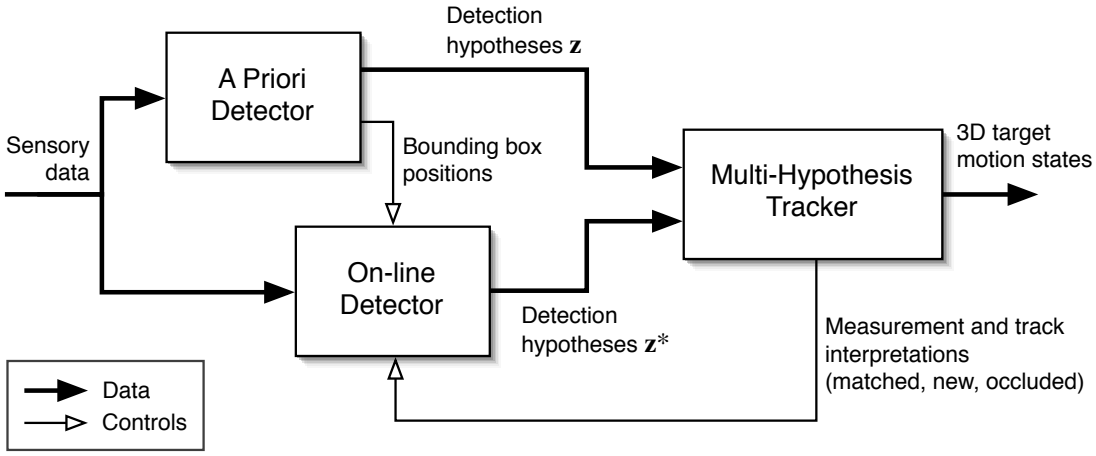


Figure 9.9: The decisional framework to integrate both detectors and the tracking system.

Observations consist in a 3D position provided by the a priori detector $\mathbf{z}_i(t)$ and a positive training sample $\vec{\varepsilon}_i(t)$ from the on-line detector. The sample $\vec{\varepsilon}_i(t)$ is a vector $\{\varepsilon^1, \ldots, \varepsilon^M\}$ of stacked feature values[57] computed in the rectangular bounding boxes obtain by back-projecting the 3D position $\mathbf{z}_i(t)$ into 2D image planes of the RGB and D images (see for example Figure 9.6).

Let Ω_l^t be the l-th hypothesis at time t and $\Omega_{p(l)}^{t-1}$ the parent hypothesis from which Ω_l^t has been derived. Let further $\psi_l(t)$ denote a set of assignments which associates predicted tracks to observations in $\mathcal{Z}(t)$ and $\mathcal{Z}^t$ the set of all sensor readings up to time t. In each cycle, the MHT method tries to associate the known tracks from the parent hypotheses of the previous step to the set of new observations $\mathcal{Z}(t)$, producing all possible assignment sets $\psi_l(t)$ that each give birth to a child hypothesis that branches off its parent. This results in an exponentially growing hypothesis tree as it can be seen in Figure 1. Most practical MHT implementations prune the tree by Murty's algorithm explained in subsection 3.8.2 able to generate and evaluate the current k best hypotheses in polynomial time.

9.6.1 Joint Likelihood Data Association

Each multi target tracking algorithm has to address the data association problem to solve the ambiguities in target to observation associations. The measurement likelihood $p(\mathbf{z}_i(t) \mid \psi_l(t), \Omega_{p(l)}^{t-1})$ in the regular MHT consists in two terms, one for observations interpreted as new tracks and false alarms (which is left unchanged) and a second one for matched observations $\mathbf{z}_i(t)$ (see Eq. 3.6). Latter follows the Gaussian likelihood model centered on the measurement prediction $\hat{\mathbf{z}}_j(t)$ with innovation covariance matrix $S_{ij}(t)$, thus

$$p(\mathbf{z}_i(t) \mid \psi_l(t), \Omega_{p(l)}^{t-1}) \;=\; \mathcal{N}(\mathbf{z}_i(t)\,; \hat{\mathbf{z}}_j(t), S_{ij}(t)). \tag{9.6}$$

This likelihood quantifies how well an observation matches a predicted measurement based on position and velocity.

Here, the on-line classifier H_{t-1} learned on the target-specific observations up to time $t-1$ adds an appearance likelihood that expresses how much the observed target's appearance matches the learned model. Thus a joint likelihood is derived that accounts for both motion state and appearance. With

[57] The vector of stacked feature values $\vec{\varepsilon}_i(t)$ is also called feature descriptor of target or observation i at time t.

$\vec{\varepsilon}_i(t)$ being the feature descriptor of $\mathbf{z}_i(t)$, defining $\mathbf{z}_i(t) = (\mathbf{z}_i(t), \vec{\varepsilon}_i(t))$ in the remainder of the chapter, and assuming independence between the two terms, Eq. 9.6 yields

$$p(\mathbf{z}_i(t) \mid \psi_l(t),\ \Omega_{p(l)}^{t-1},\ H^{t-1}) \ = \ p(\mathbf{z}_i(t) \mid \ \psi_l(t),\ \Omega_{p(l)}^{t-1}) \ \ p(\vec{\varepsilon}_i(t) \mid H^{t-1}). \tag{9.7}$$

The appearance likelihood is also modeled to be a Gaussian pdf centered on the maximum confidence of the strong classifier (which is 1.0), hence

$$p(\vec{\varepsilon}_i(t) \mid H^{t-1}) \ = \ \mathcal{N}(\kappa(\vec{\varepsilon}_i(t))\,; 1.0,\ \sigma_a^2), \tag{9.8}$$

where σ_a^2 is the variance of the Gaussian and a smoothing parameter to trade off the two likelihoods.

9.6.2 Feeding Data Association Back to On-line Boosting

In each cycle, the tracker produces assignments of observations to tracks (also called matches) and interpretations of observations as new tracks or false alarms and of tracks as occluded or deleted. This information directly serves the on-line boosting algorithm to initialize, update, pause update, and delete the strong classifiers:

- **Initialization:** When an observation $\mathbf{z}_{new}(t)$ is declared as a new target, a new track $\mathbf{x}_{new}(t)$ is initialized and a new strong classifier H_{new} is created at the bounding box position of the hypothesis of the a priori detector. To train the appearance model both sets of positive and negative samples (feature descriptors) are generated using the position uncertainty provided by the detector and assumed to have a Gaussian pdf. Samples with positions inside the interval defined by $\sigma\theta^+$ are marked as positive, samples outside $\sigma\theta^-$ are treated as negatives.

- **Update:** When an existing target $\mathbf{x}_j(t)$ is associated to an observation $\mathbf{z}_i(t)$, the strong classifier H_j^t is updated using the feature descriptor $\vec{\varepsilon}_j(t)$ calculated within the new bounding box of the a priori detector. The on-line detector is centered at this new bounding box position. The update procedure takes feature descriptors sampled in the same way as described above.

- **Confidence search:** When the MHT declares a track as occluded, there are two possible reasons: an occlusion or a misdetection. To cope with both cases, it is proceeded as follows: Given the on-line learned model, for each target without valid observations[58] the depth-informed confidence evaluation is applied to search for the target within an area centered around the motion prediction of the Kalman filter. The map size is proportional to the uncertainty of the prediction, the confidence values are calculated using the projections of the 3D positions into 2D image space. Once the confidence values are known the most likely target position is calculated using sample mean and covariance. This is unlike the approach of Grabner and Bischof [2006] in which this search is carried out in image space and with a fixed-size search window. Furthermore, the region with highest confidence value is treated as the new target position increasing the risk of drift, dramatically. However, if a high-confidence match with a likelihood above $\theta_{\mathbf{z}^*}$ was found, the occlusion event is interpreted as a misdetection. The mean position of the confidence search is considerer as "virtual" observation $\mathbf{z}^*(t)$. Otherwise, the event is treated as a (full) target occlusion and on-line learning of the corresponding strong classifier is stopped until the target reappears. This strategy does not only support detection it also avoids drifting of the model to background, clutter, or other targets. All observations $\mathbf{z}^*(t)$ from the on-line detectors are treated like regular observations for the MHT for the exception

[58] Targets without valid observation are found by inspecting the 99.9% uncertainty region around their predicted position. If this area contains no observation the confidence search is performed to find the target position.

Figure 9.10: The setup consisting in three vertically mounted Kinect sensors offering a joint field of view of $130° \times 50°$ and supplying RGB-D data with a resolution of 1440×640 pixels at $30\,Hz$. They are mounted at $1.2\,m$ height.

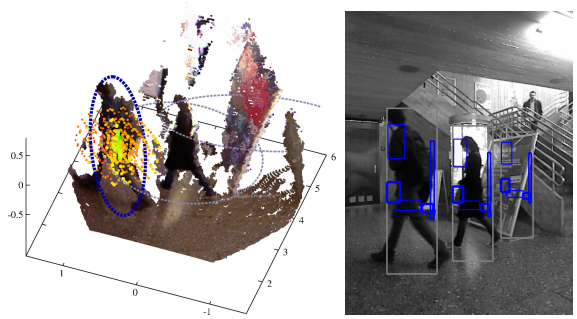
that they cannot create new tracks[59]. To avoid that hypotheses with unassigned virtual observations $\mathbf{z}^*(t)$ obtain a low probability the false alarm probability $\lambda_{fal}(\mathbf{z}^*(t))$ is set to a value close to 1.0.

- **Deletion:** If a target disappears from the sensor field of view its track is deleted by the MHT. Simultaneously, the corresponding strong classifiers becomes obsolete and is removed. Currently, the appearance information encoded in the on-line learned model is lost completely. The further usage of the models to re-detect people or to refine the a priori detector is left to future work.

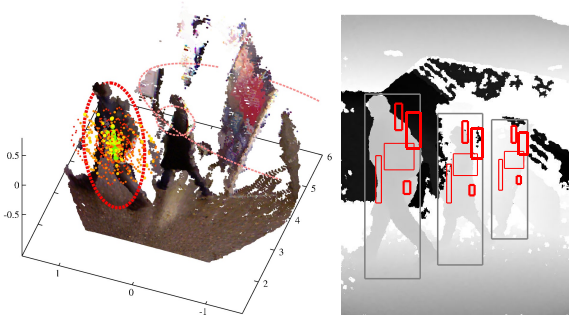
9.7 Experiments

To evaluate and compare the different features and to analyze the tracking accuracy of the proposed on-line learned detector approaches, a large-scale indoor data set with unscripted behavior of people was collected. The data set has been taken in the lobby of a large university canteen at lunch time. The a priori detector has been trained with an additional background data set collected in another, visually different university building. This is to avoid detector bias towards the visual appearance of the canteen lobby, especially since the data is acquired from a stationary sensor. The data set has been manually annotated on a 3D point basis to include the bounding boxes in 2D color and depth image space, the visibility of subjects (fully visible/partially occluded), and the data association ground truth of the tracks. A total of 3021 instances of people in 1133 frames and 31 tracks have been annotated. The sensory setup for data collection is shown in Figure 9.10. It consists in three vertically mounted Kinect sensors that jointly extend the field of view to $130° \times 50°$. Measures have been taken to calibrate the intrinsic and extrinsic parameters of the setup and to guarantee synchronized acquisition of the three images at frame rate.

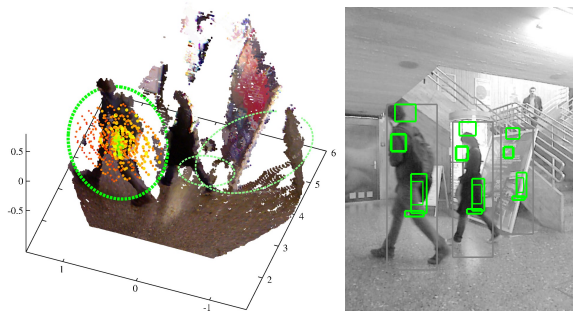
[59] The limitation that observations $\mathbf{z}^*(t)$ are not allowed to initialize new tracks is obvious as they are found by the on-line detectors trained on already known targets.



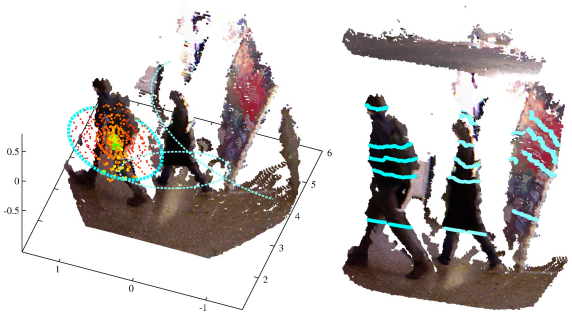
(a) Haar-like features computed on the intensity (grayscale) image. $\kappa(\vec{\varepsilon}) = 0.921 \,/\, 0.653 \,/\, 0.377$.



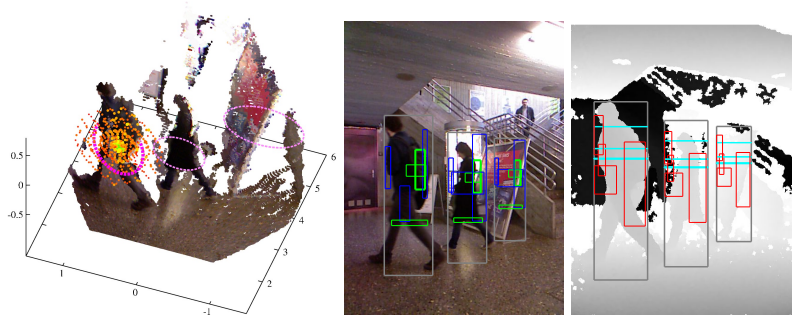
(b) Haar-like features computed on the depth image. $\kappa(\vec{\varepsilon}) = 0.853 \,/\, 0.167 \,/\, 0.026$.



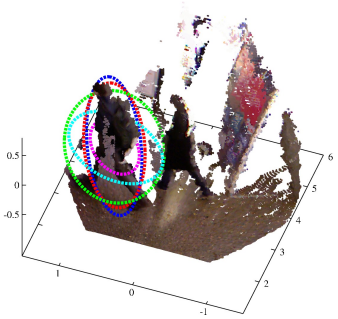
(c) *Lab* color features computed on the *Lab* image. $\kappa(\vec{\varepsilon}) = 0.892 \,/\, 0.238 \,/\, 0.293$.



(d) Geometrical features computed in real 3D space. $\kappa(\vec{\varepsilon}) = 0.997 \,/\, 0.879 \,/\, 0.474$.



(e) Mixture of all features. $\kappa(\vec{\varepsilon}) = 0.954 \,/\, 0.231 \,/\, 0.131$.



(f) Comparison.

Figure 9.11: Comparison of the proposed features employed to learn target-specific appearance models. In the left sub-figures in (a) to (e) the classification results of the positions sampled in 3D are indicated by colored dots. Green denotes high and red low confidence values, respectively. The sample covariances are shown as dashed ellipses colored differently to allow a comparison in (f). The right figures visualize the input data, the bounding boxes of the targets, and the patches of the features. The maximum confidence values achieved for the targets from left to right are given below the figures.

9.7.1 Evaluation of 2D and 3D Features

The first experiment investigates how the proposed features influence the ability of the on-line learned, target-specific appearance model to distinguish the correct person from others and the background. Additionally, the 3D confidence search that estimates the state of a target by evaluating sampled

positions is analyzed. To perform the analysis, image, depth, and geometrical features are employed individually and in combination. The analysis is performed as follows, first an appearance model of a person is trained on a single frame using 50 selectors each having 50 weak classifiers using only one type of feature[60]. Thereafter, three independent confidence searches as explained in subsection 9.5.4 are performed on the next frame at the ground truth positions of the original (correct) person, a second (wrong) person, and on background. As background object an billboard has been taken as it slightly resembles the appearance of a person. The test conditions, estimated positions and covariances of the confidence searches, and maximum confidence values are shown in Figure 9.11. Additionally, the best features are depicted. The results are analyzed in more detail below.

The **Haar-like features on the intensity image** as employed in Grabner and Bischof [2006] yield very high confidence values when the correct person is analyzed (0.921). Unfortunately, they poorly separate different people thus the maximum confidence value of the wrong person is still 0.653. Also background with horizontal structures slightly resembling the contour of a person – that is where these feature work best – still attain medium confidence values (0.377). Another drawback is the large uncertainty in the z-dimension of the camera (shown in Figure 9.11a) as the corresponding image patches are very stable to small distance changes. However, these feature seem to be sufficient when tracking single targets, especially when their change in appearance is rather small. For the purpose of multi-target tracking the distinction between people is to small.

Haar-like features on the depth image are well suited to distinguish people in front of different background as shown in Figure 9.11b. The maximum confidence value of the correct person is 0.853 while the wrong person gets 0.167 and the billboard 0.026. When people walk in front of the same background (especially with the same distance) this difference slightly declines. However, these features well support tracking due to the joint data association but suffer from the same problem as the Haar-like features on the intensity image which is strong the large uncertainty in the z-dimension of the camera.

The **Lab color features** yield very stable results when people are dressed in different colors. The maximum confidence value of the correct person is 0.892, the wrong person is classified with 0.238, and background with 0.293. As people usually wear single color clothing the covariance of the position estimate as shown in Figure 9.11c is very large in general thus tracking with these features probably leads to inaccurate trajectories.

When employing the **Geometrical features** computed on the 3D point cloud the state estimate is accurately centered at the correct position of the person and has a small covariance (see Figure 9.11d). Especially, the formerly mentioned problems of estimating the correct position in the z-dimension of the camera do not occur. The accurate position estimate comes to the expense of a very poor discrimination of people. While the correct person is classified with a maximum confidence value of 0.997 the wrong person still obtains 0.879. The billboard still gets a relatively high confidence of 0.474 as most of the horizontal structures in the background. Tracking approaches employing these features in isolation are presumed to produce very accurate 3D trajectories but run into difficulties when data association becomes hard.

Employing a **mixture of all features** as shown in Figure 9.11e achieves the most accurate and stable results. The proposed approach benefits from all advantages of the individual features described above. An very exact discrimination between the correct and wrong person with maximum confidence values of 0.954 and 0.231, respectively, is achieved. Furthermore, the predicted state is very accurate with precise position estimate and low covariance. A comparison of the state estimates calculated using the confidence search based on appearance model classifiers – employing the proposed features in isolation or combination – is shown in Figure 9.11f.

[60] The feature positions and sizes are still sampled within the the bounding boxes of the targets.

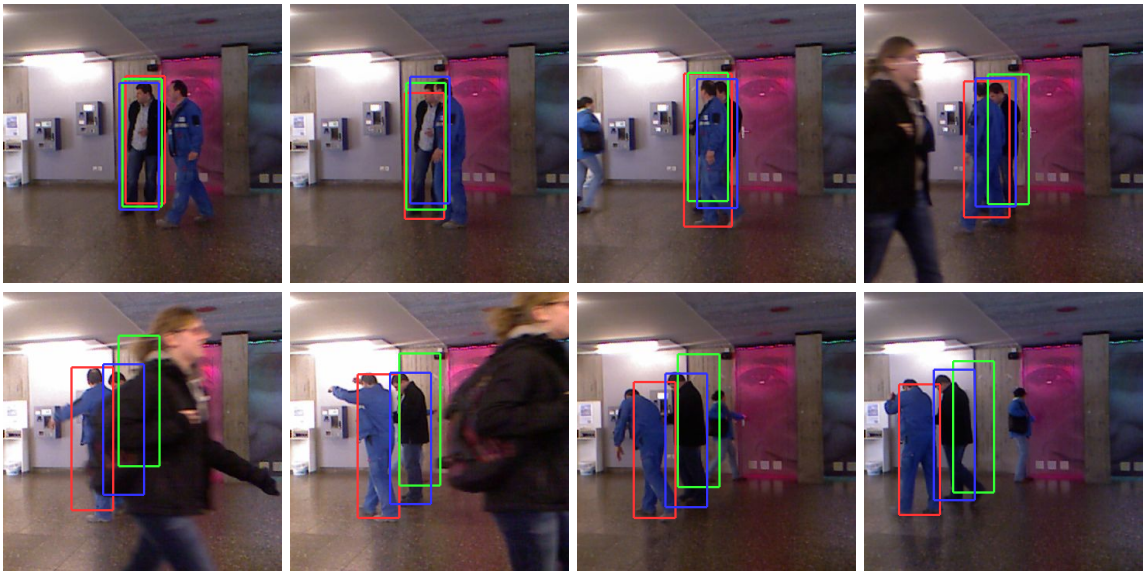


Figure 9.12: Tracking results using 2D confidence search (baseline, shown as red bounding box), 3D confidence search (green), and 3D search with controlled on-line learning via track interpretation events (blue). During occlusions updating the appearance model with the most likely target position found by a confidence search leads to drift. Using track interpretation feedback to pause on-line learning, drift is avoided.

9.7.2 Controlling On-line Learned through Tracking Feedback

The second experiment compares the known confidence search in 2D and the proposed search in 3D. Additionally, the benefits of controlled appearance learning using track interpretation events from a multi-hypothesis tracker are investigated. The analysis is performed by tracking a single person that is frequently occluded by others. Track and appearance model are initialized and updated using detections from the a priori detector. During occlusions the a priori detector fails and the confidence search serves to find the most likely target position that is in turn used to update position and appearance model. The compared methods differ in the way that during occlusions either the 2D confidence search (baseline) or 3D confidence search are employed to find the most likely position. In the third case – when on-line learning is controlled by the MHT – learning is stopped when the target is marked as occluded. The tracking scenario and the bounding boxes of the tracking result are shown in Figure 9.12. Maximum confidence values and tracking errors are presented in Figure 9.13.

Continuously updating the target-specific appearance model completely ignores the important information that the target might be occluded by others. In such a case the image regions given by the most likely target position contain wrong information that pollute the model and leads to drift to other targets or background. The confidence search in 2D further suffers from a systematic error that is caused by a wrong estimate of the distance between the sensor and the target (see Figure 9.4). The confidence search in 3D solves that problem and reduces the drift as the search area is limited by the current state uncertainty of the target predicted with a motion model. However, during lengthy occlusion events the appearance model adapts to the wrong data and the state prediction starts to drift. Additionally, once a model is polluted it hardly reaches very high confidence values of the correct target again. This is shown in Figure 9.13 starting at frame 765.

Feeding the information of the multi-hypotheses tracking framework back to control model learning solves both problems: drift and wrong confidence values. During occlusions the on-line learning is

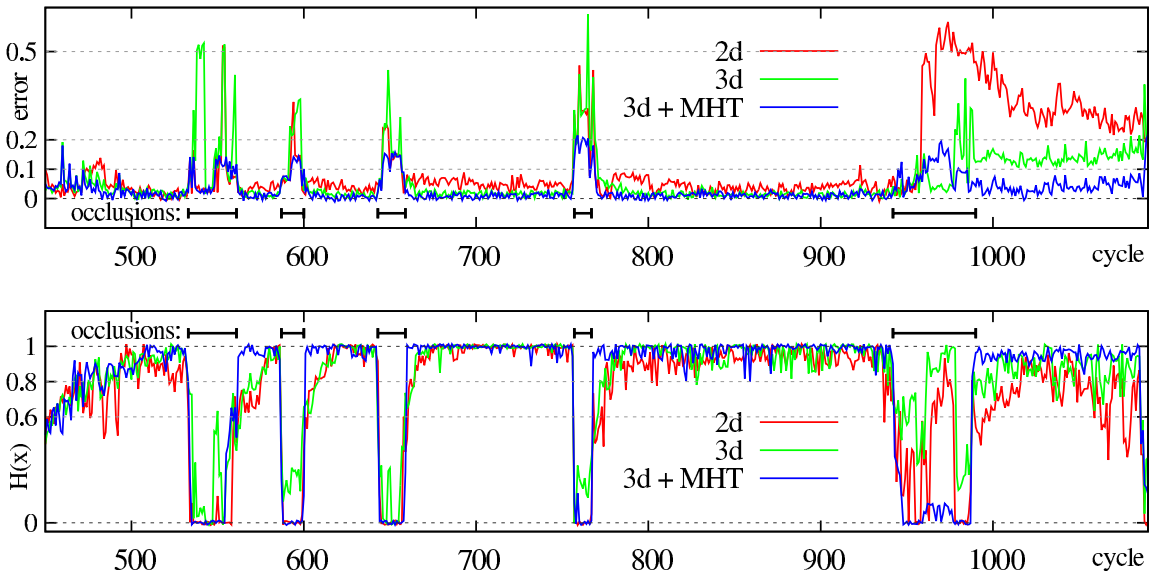


Figure 9.13: Tracking errors in meters (top) and maximum confidence values (bottom) when tracking a single target that is frequently occluded by others. During occlusion the a priori detector fails and a confidence search in 2D (red lines) or 3D (green lines) serves to find the target. Controlling on-line learning through track interpretation events of the MHT (blue lines) increases the confidence values and lowers the tracking error.

paused, thus the appearance model remains unchanged. The tracking error caused by the drifted models is reduced significantly. After the occlusion the confidences of the 3D search quickly recover to very high values as only small adaptation to the slightly changed appearance of the original target are necessary.

9.7.3 Tracking with On-line Learned Appearance Models

The last experiment evaluates the tracking accuracy of the proposed on-line learned appearance models integrated into the MHT on a large-scale indoor data set with unscripted behavior of people. The parameters of the MHT have been learned from a training data set over 600 frames. The detection probability is set to $p_{det} = 0.99$ and the termination likelihood to $\lambda_{del} = 30$. The average rates of new tracks and false alarms are determined to be $\lambda_{new} = 0.001$ and $\lambda_{fal} = 0.005$, respectively. Further, the maximal number of hypothesis N_{Hyp} is set to 100. The strong classifiers of the targets are based on 50 selectors which are trained with 50 weak hypotheses.

To assess the impact of the on-line boosting onto the tracking performance a tracker using the a priori detector only is used to obtain a baseline. All following runs are then compared using the CLEAR MOT metrics defined by Bernardin and Stiefelhagen [2008]. The metric counts three numbers with respect to the ground truth that are incremented at each frame: misses (missing tracks that should exist at a ground truth position, FN), false positives (tracks that should not exist, FP), and mismatches (track identifier switches, ID). The latter value quantifies the ability to deal with occlusion events that typically occur when tracking people. From these numbers MOTA (average number of times of a correct tracking output with respect to the ground truth) is determined. The tracking results are shown in Table 9.1.

The results show a clear improvement of all values except for the number of false positives, especially the overall tracking accuracy MOTA increases by 16% which is remarkable. A manual

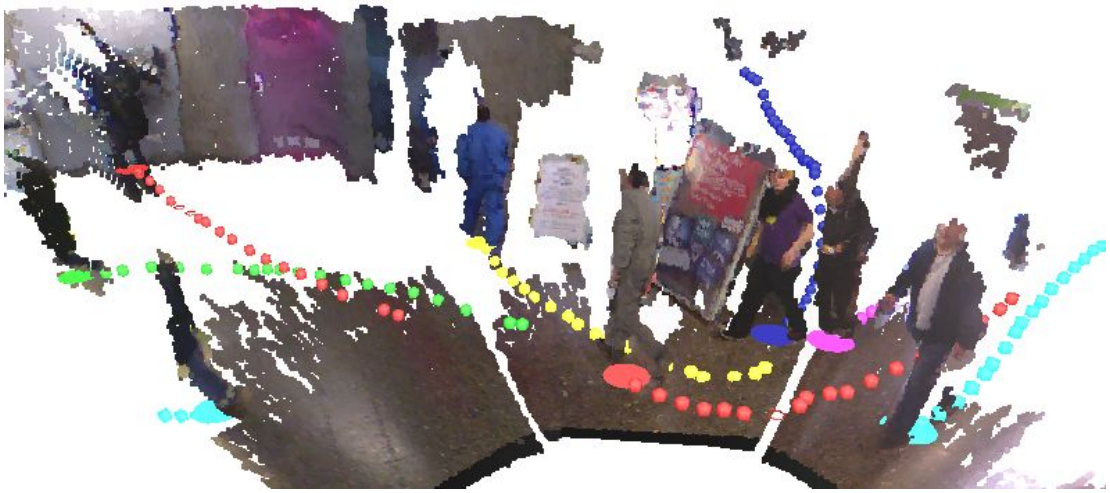


Figure 9.14: Visualization of the 3D point cloud produced by the three MS Kinect sensors including the positions and trajectories of eight of 31 tracks in the data set. The colored disks mark the current Kalman filter estimates of the target positions, the small dots show their past trajectories. The tracker maintains full 3D estimates as it can be seen by the dark blue trajectory of the subject coming down the stairs.

Approach	**FN**	**FP**	**ID**	**MOTA**
Baseline	1502	168	42	62%
On-line boosting	751	201	32	78%
Improvement	50%	-19%	24%	16%

Table 9.1: CLEAR MOT results using on-line learned appearance models.

inspection of the behavior of the tracker gained detailed insights that are discussed below.

The strongest impact of the presented approach is the reduction of the number of missed targets by 50%. This improvement is caused by the on-line found observations $\mathbf{z}^*(t)$. When the a priori detector fails to detect an existing track in several consecutive frames, the best MHT hypothesis will eventually (and wrongly) declare the track as deleted. When this happens, the miss count (FN) is increased at each frame until the detector finds the target again and creates a new track. This is where the $\mathbf{z}^*(t)$ observations come into play by detecting the target from the on-line learned models using confidence searches in 3D. Given a specific $\mathbf{z}_j^*(t)$, the MHT can match the target $\mathbf{x}_j(t-1)$ and correctly continue the track. The misses are avoided.

This benefit comes at the expense of a delayed deletion of tracks that are incorrectly created from wrong false positives of the a priori detector. In this case, the on-line detector tries to continue the track with the same strategy leading to a increase of the number of false positives (FP) by 19%. The occurrence of the behavior was observed for recurring false positive detections on static objects – like the billboard – on which the on-line detector can particularly well adapt. Systematic failures of the a priori detector are the major problem of the proposed approach as the on-line detectors assume the targets to be of the correct type.

The improvement in the number of identifier switches (ID) is achieved by the joint likelihood model that guides data association in situations of interacting and thus occluding targets. The fact that

this number is not higher is due to the unscripted behavior of people in the recorded data set. At the particular place of data collection, subjects mainly walked past rather than creating situations that stress the occlusion handling capability of the tracker.

9.8 Conclusions

In this chapter a novel approach on 3D people detection and tracking in RGB-D data is presented. It combines on-line learning of target appearance models using four types of RGB-D features with multi-hypothesis tracking. A decisional framework to integrate the on-line person detector, an off-line learned a priori detector, and a multi-hypothesis tracker is proposed. The framework enables the tracker to support the on-line classifier in training only on the correct samples and to guide data association via a joint motion and appearance likelihood. It also avoids the key problem of on-line adaptation namely drifting of models to background, clutter, or other targets by resetting the detection window at the location of the a priori detector and pausing adaptation in case of occlusions. The framework further allows to fill gaps of false negatives from the a priori detector by observations of the on-line detectors found by a depth-informed confidence maximization search in 3D space.

The experiments show a clear overall improvement of the tracking performance, particularly in the number of missed tracks and also in the number of identifier switches. They demonstrate that the on-line classifier contributes to find the correct observations in cases when the a priori detector fails. This reduces the number of missed tracks by 50%. Further, the joint data association likelihood helps to decrease the number of track identifier switches by 24%. The overall tracking accuracy (MOTA) is improved by 16%.

Future work will focus on the collection and annotation of more RGB-D data sets containing a variety of challenging social situations that stress more aspects of this approach. Finally, the target models are currently learned for each track in isolation. Extending the on-line detector to learn the models jointly over all tracks promises to even better distinguish them from each other.

Discussion and Outlook

10 Discussion

Human oriented robotics focuses on the development of state-of-the-art robotic systems made to operate in crowded human environments. As the generation of safe, efficient, and socially acceptable behavior requires precise knowledge about the presence and motion states of surrounding individuals, people detection and tracking become key technologies. During the last decades extensive robotic research led to many different approaches on people detection using various sensor modalities. Further, goal oriented motion prediction and advanced probabilistic data association techniques have been invented. However, the developed methods barely consider that human coexistence is based on (unwritten) social rules and normative behavior. People share social relations and respect the needs and desires of each other. This thesis shows, that the accuracy and robustness of people detection and tracking can be increased by taking these kinds of information into account.

In more detail, this thesis proposes various methods to learn, model, and integrate *spatial*, *temporal*, and *social* information into people detection and tracking. Analyzed on extensive real world experiments it is shown, that these models enhance people detection, ease the interpretation of detection events, improve motion prediction, and guide data association. All methods increase the accuracy of people detection and tracking, respectively, measured with state-of-the-art metrics. Their computational complexity is analyzed and it is shown, that despite expensive algorithms, people tracking can still be applied in real-time.

10.1 Conclusion

This section draws conclusions and presents results of the methods and models developed in this thesis. Furthermore, potential directions of future research are outlined.

People Detection

The presented work shows, that spatio-temporal information supports people detection by two aspects. First, location-specific classifiers can be trained to increase the detection accuracy in general. And second, learned spatio-temporal priors enable to produce refined interpretations of detection events. Additionally, on-line learned target-specific appearance models can be used to improve people detection, for example, when the a priori detector fails.

Regular approaches for people detection learn generic classifiers from manually annotated training data that are employed to detect people at all locations of the sensed environment. But for most sensor modalities, the appearance of people depends on the distance and the angle to the sensor. Thus, learning an accurate and robust but at the same time generic detector is a very challenging task. In Chapter 2, an extended approach on feature-based people detection which takes spatial information into account is introduced. Accounting for range-dependent appearance a cascade of range-specific detectors is trained. Each detector uses a set of boosted features. The approach leads to superior robustness compared to state-of-the-art methods and is transferable to new environments. From these results we conclude, that the integration of spatial information is fundamental to achieve the goal of robust and accurate people detection. However, future work includes the simultaneous

learning of spatial partitionings and detectors. In addition, temporal information may be integrated to account for changed appearances due to daytime or seasonal conditions.

However, detection only will never achieve perfect accuracy, thus false positive detections need to be recognized and filtered out. If confidence values are provided by the detection system track initialization can be controlled. But the choice of the correct confidence threshold remains is difficult task. In this work, spatial priors on detection events are either derived from a map of the environment or modeled by hand (Chapter 5) or learned by tracking people and observing the classification failures of the detector (Chapter 6). While the former still requires sensor and environment-specific knowledge in advance, the latter applies on-line learning without any prerequisites. However, from the presented results we conclude, that reasoning on spatio-temporal affordances increases tracking accuracy in terms of fewer false positive tracks. Future work may include the prediction of locations with high false positive likelihoods based on semantic information.

In addition to the use of spatial information, target-specific detectors can be learned on-line to support people detection in case of partial occlusions. Especially cameras, 3D range finders, or RGB-D sensors provide rich information for modelling the appearance of individuals. This work presents the first approach on reliable 3D people detection and tracking combining a novel multi-cue person detector for RGB-D data with an on-line detector that learns individual target appearance models on the fly. For on-line learning, we propose a boosting approach using four types of RGB-D features and a depth-informed confidence maximization search in 3D space. The approach is general in that it neither relies on background learning nor a ground plane assumption. The two detectors are integrated into a decisional framework with a multi-hypothesis tracker that controls on-line learning through a track interpretation feedback that avoids drift (Chapter 9). The results show a clear improvement of the tracking performance, particularly in terms of fewer track identifier switches and fewer missed tracks. We conclude, that appearance information can resolve ambiguities and bridge the gaps of false negatives from the a priori detector. In the future, the human-specific information encoded in the on-line learned appearance models could also be integrated into the a priori people detector to decrease the likelihood of false positive detections.

Motion Prediction

Many people tracking approaches make weak assumptions on human motion. But even over a short period, human behavior is complex and influenced by many factors. This work shows, how the integration of social rules, social grouping behavior, and environment-specific spatial constraints lead to refined motion predictions that translate into a more robust tracking behavior and better occlusion handling.

Human motion is complex and follows non-random, non-linear patterns influenced by inner motivation, social rules in the presence of other people, and the physical and social constraints of the environment. This thesis considers computational models to describe and predict individual and collective pedestrian behavior, precisely. The models have been developed in the cognitive and social science communities, for example, for the task of crowd behavior analysis. The social force model (investigated in Chapter 4) offer a concept well suited to describe all these aspects in a sound and common framework. We demonstrate, that integrating social force based motion prediction into a multi-hypothesis target tracker reduces the number of data association error and the number of false positive tracks, significantly. From this we conclude, that understanding social aspects in human behavior aids to resolve ambiguities. False positive detections, that conflict with social or physical constraints, can be recognized and filtered out. Although the same set of model parameters can be shared over different environments, in the future, on-line learned parameters depending on spatio-temporal and social aspects may be investigated.

In many situations, people are encountered in groups formed by social relations between individuals. Typically, groups form certain stable patterns, called spatial intra-group relations, that remain largely stable over time. In this work, social relations are learned, detected, and tracked over time to infer group affiliations. This information is then used to on-line learn spatial relations between people in groups (Chapter 7). Additionally, we propose a motion model for maneuvering groups. The model is informed by priors from the social science community and predicts human motion jointly over the intra-group constraints using a particle-based approach. We show, that on-line learned social and spatial relations improve people tracking by significantly fewer track identifier switches and fewer false negative tracks. From these results we conclude, that learning social and spatial relations are key technologies to resolve ambiguities of targets in groups and to bridge gaps of occlusions. Especially latter occur frequently, when tracking people from a first-person perspective. Currently, social relations are defined by coherent motion features. In the future, additional cues – like age, gender, or family affiliation – may be integrated to improve social relation recognition.

The probabilities and frequencies at which people appear, disappear, walk, or stand in an environment are not uniform but vary over space making human behavior strongly place-dependent. For example, people turn around convex corners, maneuver around obstacles, stop in front of doors, and especially they do not go through walls. In this thesis, we present a novel approach to model and encode on-line learned spatial priors on human behaviors (Chapter 6). In detail, we propose a non-homogeneous spatio-temporal Poisson process to encode the spatially and temporally varying distribution over relevant human activity events. From that model, we derive a novel approach on place-dependent motion predictions, that follow the space usage patterns of humans. Integrated into the multi-hypothesis tracker, even highly maneuvering people can be tracked at an observation frequency as low as 0.5 Hz. So we conclude, that learned environment-specific constraints and affordances are important to robustly predict the motion of humans. Future work could investigate methods to predict the space usage behavior and motion patterns of humans based on semantic information learned on known environments.

Data Association

The models developed in this thesis provide an amazing amount of information available while tracking people. Thus, we also propose a framework to integrate *spatial*, *temporal*, and *social* information into data association. Employing multi-hypothesis tracking, we show that by using informed models people tracking can be made substantially more accurate without compromising efficiency. The theoretical background of the regular MHT and the proposed extentions are presented in Chapter 3.

The regular MHT reasons about incoming observations emanating from previously known targets, from new targets, or from false alarm detections in clutter. Occurrences of new tracks and false alarms are usually modeled using Poisson distributions with fixed rates of expected numbers of events. However, we prevent these general assumptions and take into account, that people do not appear uniformly in space and time. Humans typically enter the environment at specific locations like doors, elevators, or corners and appear more often at daytime than at night. These spatio-temporal priors are either modeled (Chapter 5) or learned (Chapter 6) for each environment using Poisson processes with spatio-temporal dependent rate functions. The theory, presented in this thesis, provides a mathematically sound framework to integrate such spatio-temporal-dependent rate functions into MHT tracking. With extensive experiments, we show, that they reduce the amount of track identifier switches and false positive detections, significantly. We conclude, that place-dependent models – seamlessly integrated into the MHT framework – serve to detect false positives in clutter. Additionally, they overcome the usual fixed Poisson rate assumptions that are not well suited when

tracking people.

When a person disappears from the sensor field of view, it is either occluded or it has left the monitored area. Unfortunatelly, both situations look similar – since in both cases the person is not visible – but they differ in their information content. In the first case, the person is just occluded and its state must be maintained by the tracker. In the second case, the person is outside the surveillance area and the corresponding track can be deleted from the tracking system. However, modeling occlusion and deletion events with uniform and constant probabilities are poor assumptions often made in people tracking. In this work, we propose a physical model that predicts occlusions to occur behind static obstacles or other people. Additional, we investigate a place-dependent deletion model, that uses an exponential function to simulate the decay in the probability of detecting a target (Chapter 5). Both model contribute to lower the number of track identifier switches, significantly. The proposed occlusion model employs an expensive sampling strategy. Future work may investigate heuristics or other strategies to decrease this computational effort.

Frequent occlusions of people walking in groups are main reasons for detection failures. However, in this work we learn social relations between individuals and track the social group formations to better cope with such situations (Chapter 7). Based on coherent motion features, relations between individuals are learned and detected using a linear SVM classifier. We infer group affiliations from social group formations tracked with a multi-model multiple hypothesis data association framework. Using adaptive track-specific occlusion probabilities we demonstrate, that our approach is able to track occluded group members, reliably. In experiments with large-scale outdoor data sets, we show that our approach improves people tracking by significantly fewer track identity switches. From these results we conclude, that social information is especially important when tracking many people in every day environments. Future work may include to incorporate additional information into social relation recognition, such as age or gender.

Probabilistic data association techniques rely on similarity measures to evaluate the likelihood of track to observation assignments. When tracking targets of identical appearance, these measures usually depend on positions and motion states estimates. However, additional cues can be applied to guide data association. By employing a multi-cue person detector and on-line learning of appearance models, we introduce a joint likelihood data association that accounts for both motion state and appearance conformity. The on-line learned person classifier – based on on-line learned target-specific appearance models – adds an appearance likelihood that expresses how much the observed target's appearance matches the model learned from previous observations. The appearance likelihood is modeled to be a Gaussian pdf centered on the maximum confidence of the classifier (Chapter 9). Combining the a-priori with the on-line detector results in the first robust and reliable people detection and tracking system in 3D. In the future, also other schemes on fusing motion and appearance information may be analyzed.

Appearance and Appearance Dynamics-based Classification

Tracking and classifying various dynamic objects – like humans, animals, vehicles – makes it hard to manually design suitable models for their appearance and dynamics. Thus, this work presents an unsupervised learning approach for representing the time-varying appearance of objects in 2D laser range data using probabilistic exemplar-based models. Employing a clustering procedure that builds a set of object classes from given observation sequences the system is able to autonomously learn useful models for, e.g., pedestrians, skaters, or cyclists without being provided with any external class information (Chapter 8).

All techniques presented in this thesis have been implemented and thoroughly tested. The experiments have been carried out using static sensors as well as our mobile robot DARYL. To observe people, 2D and 3D range sensors, cameras, and RGB-D sensors have been employed. The experiments have been carried out in indoor and outdoor environments. Indoor scenarios include office buildings, laboratories, and an university canteen. Outdoors, we collected data in the inner city and at the main station of Freiburg, in an urban environment in Zurich, and at the university campus. Furthermore, we implemented the proposed methods. All algorithms are running in real-time.

10.2 Future Work

Despite the encouraging acchievements presented in this thesis, there are aspects that could need more investigation. The most interesting research direction on improving people detection and tracking in general is the integration of further social and contextual information. Social information includes – besides grouping behavior – age, gender, special needs, as well as the cultural background. Context information provided, for example, by situation recognition systems might provide clues about the intended behavior of people. If goal locations and preferred walking routes are known in advance, people tracking can be made more robust against occlusions and detector failures.

In this theses, we provide mathematical frameworks to integrate temporal information into detection and tracking. But so far, our detectors are trained independently of any temporal aspects. However, people look different at various daytimes and seasons. Thus, time-dependent detectors – similar to the proposed and evaluated space-dependent detector – should be studied. The integration of social information into detection could also provide valuable information. For example, people in groups can be detected with lower confidence thresholds as they usually stay together.

The field of motion prediction is well explored, for example, learning goals and motion patterns. However, in the people tracking community, interactive multiple models (IMM) have hardly gained attention for predicting human motion. Current approaches treat everyone the same by using a fixed motion model for each target. Given additional information, like age, profession, or group affiliation, the employed models can be switched accordingly or, at least their parameters can be adapted. The social force model used in this thesis considers only repulsive forces of other people and static obstacles. Especially for people walking in groups, additional attractive forces could be incorporated.

Besides the integration of further information, the exchange of data between the three main components of tracking – namely detection, motion prediction, and data association – needs to be investigated in more detail. The extension of the usual Tracking-By-Detection framework proposed in this thesis includes feedback of short-term tracking information into on-line learned target-specific detectors. However, future methods could try to postpone detection. Using segmentation only, tracked objects could be classified based on a delayed detection step and statistical information.

List of Figures

List of Tables

Bibliography

[Ali and Shah, 2008] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 1–14, Marseille, France, 2008.

[Arras, 2003] K. O. Arras. *Feature-based robot navigation in known and unknown environments*. PhD thesis, Swiss Federal Institute of Technology Lausanne (EPFL), These No. 2765, 2003.

[Arras et al., 2008] K. O. Arras, S. Grzonka, M. Luber, and W. Burgard. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 1710–1715, Pasadena, USA, 2008.

[Arras et al., 2007] K. O. Arras, O. Martínez Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 3402–3407, Roma, Italy, 2007.

[Ashbrook and Starner, 2003] D. Ashbrook and T. Starner. Using GPS to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Computing*, 7(5):275–286, October 2003.

[Avidan, 2004] S. Avidan. Support vector tracking. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 26(8):1064–1072, 2004.

[Babenko et al., 2011] B. Babenko, Ming-Hsuan, and Y. S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 33(8):1619–1632, 2011.

[Bahadori et al., 2005] S. Bahadori, G. Grisetti, L. Iocchi, G. R. Leone, and D. Nardi. Real-time tracking of multiple people through stereo vision. In *Proceedings of the IEE International Workshop on Intelligent Environments*, 252–259, Colchester, UK, 2005.

[Bajracharya et al., 2009] M. Bajracharya, B. Moghaddam, A. Howard, S. Brennan, and L. Matthies. Results from a real-time stereo-based pedestrian detection system on a moving vehicle. In *Proceedings of the ICRA 2009 Workshop: People Detection and Tracking*, Kobe, Japan, 2009.

[Bar-Shalom et al., 2002] Y. Bar-Shalom, T. Kirubarajan, and X.-R. Li. *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[Bar-Shalom and Li, 1995] Y. Bar-Shalom and X.-R. Li. *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS Publishing, Storrs, USA, 1995.

[Bennewitz et al., 2005] M. Bennewitz, W. Burgard, G. Cielniak, and S. Thrun. Learning motion patterns of people for compliant robot motion. *International Journal of Robotics Research (IJRR)*, 24(1):31–48, 2005.

[Bernardin and Stiefelhagen, 2008] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008.

[Best and Norton, 1997] R. Best and J. Norton. A new model and efficient tracker for a target with curvilinear motion. *IEEE Transactions on Aerospace and Electronic Systems*, 33(3):1030–1037, 1997.

[Beymer and Konolige, 1999] D. Beymer and K. Konolige. Real-time tracking of multiple people using stereo. In *ICCV Workshop on Frame-Rate Applications*, Kerkyra, Greece, 1999.

[Bishop, 2006] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[Blackman, 2004] S. S. Blackman. Multiple hypothesis tracking for multiple target tracking. *IEEE Transactions on Aerospace and Electronic Systems*, 19(1):5–18, 2004.

[Breitenstein et al., 2009] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1515–1522, Kyoto, Japan, 2009.

[Breitenstein et al., 2011] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online multi-person tracking-by-detection from a single, uncalibrated camera. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 33(9):1820–1833, 2011.

[Bresenham, 1965] J. Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems Journal*, 4(1):25–30, 1965.

[Bruce and Gordon, 2004] A. Bruce and G. Gordon. Better motion prediction for people-tracking. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, New Orleans, LA, USA, 2004.

[Burstedde et al., 2001] C. Burstedde, K. Klauck, A. Schadschneider, and J. Zittartz. Simulation of pedestrian dynamics using a 2-dimensional cellular automaton. *Physica A*, 295:507–525, 2001.

[Carballo et al., 2008] A. Carballo, A. Ohya, and S. Yuta. Fusion of double layered multiple laser range finders for people detection from a mobile robot. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 677–682, Seoul, Korea, 2008.

[Carballo et al., 2010] A. Carballo, A. Ohya, and S. Yuta. Laser reflection intensity and multi-layered laser range finders for people detection. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*, 379–384, Viareggio, Italy, 2010.

[Chang et al., 2010] S. Chang, R. Sharan, M. T. Wolf, N. Mitsumoto, and J. W. Burdick. People tracking with UWB radar using a multiple-hypothesis tracking of clusters (MHTC) method. *International Journal of Social Robotics*, 2(1):3–18, 2010.

[Chis and Grosan, 2006] M. Chis and C. Grosan. Evolutionary hierarchical time series clustering. In *6th International Conference on Intelligent Systems Design and Applications (ISDA)*, 451–455, Washington, DC, USA, 2006.

[Choi and Savarese, 2012] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 215–230, 2012.

[Choi et al., 2011] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3273–3280, 2011.

[Cochran et al., 1967] W. Cochran, J. Cooley, D. Favin, H. Helms, R. Kaenel, W. Lang, G. Maling, D. Nelson, C. Rader, and P. Welch. What is the fast Fourier transform? *IEEE Transactions on Audio and Electroacoustics*, 15(2):45–55, 1967.

[Corvee and Bremond, 2010] E. Corvee and F. Bremond. Body parts detection for people tracking using trees of histogram of oriented gradient descriptors. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 469–475, Boston, MA, USA, 2010.

[Cox and Hingorani, 1996] I. J. Cox and S. L. Hingorani. An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 18(2):138–150, 1996.

[Cristani et al., 2011] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino. Towards computational proxemics: Inferring social relations from interpersonal distances. In *Proceedings of the IEEE International Conference on Social Computing (SocialCom)*, 290–297, Boston, MA, USA, 2011.

[Cui et al., 2005] J. Cui, H. Zha, H. Zhao, and R. Shibasaki. Tracking multiple people using laser and vision. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2116–2121, Alberta, Canada, 2005.

[Cui et al., 2006a] J. Cui, H. Zha, H. Zhao, and R. Shibasaki. Laser-based interacting people tracking using multi-level observations. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1799–1804, Beijing, China, 2006.

[Cui et al., 2006b] J. Cui, H. Zha, H. Zhao, and R. Shibasaki. Robust tracking of multiple people in crowds using laser range scanners. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 857–860, Washington, DC, USA, 2006.

[Cutler and Davis, 2000] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):781–796, August 2000.

[Dalal and Triggs, 2005] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 886–893, San Diego, CA, USA, 2005.

[Danchick and Newnam, 1993] R. Danchick and G. Newnam. A fast method for finding the exact n-best hypotheses for multitarget tracking. *IEEE Transactions on Aerospace and Electronic Systems*, 29(2):555–560, 1993.

[Day and Edelsbrunner, 1984] W. H. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24, December 1984.

[Ding and Yilmaz, 2011] L. Ding and A. Yilmaz. Inferring social relations from visual concepts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 699–706., Barcelona, Spain, 2011.

[Drumwright et al., 2004] E. Drumwright, O. C. Jenkins, and M. J. Mataric. Exemplar-based primitives for humanoid movement classification and control. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 140–145, New Orleans, LA, USA, 2004.

[Duffner and Odobez, 2011] S. Duffner and J.-M. Odobez. Exploiting long-term observations for track creation and deletion in online multi-face tracking. In *IEEE Conference on Automatic Face and Gesture Recognition (FG)*, 525–530, Santa Barbara, CA, USA, 2011.

[Elfes, 1989] A. Elfes. Using occupancy grids for mobile robot perception and navigation. *IEEE Computer*, 22(6):46–57, 1989.

[Enzweiler et al., 2010] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 990–997, San Francisco, CA, USA, 2010.

[Enzweiler and Gavrila, 2009] M. Enzweiler and D. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 31 (12):2179–2195, 2009.

[Ess et al., 2009a] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. Robust multi-person tracking from a mobile platform. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 31(10):1831–1846, 2009.

[Ess et al., 2009b] A. Ess, K. Schindler, B. Leibe, and L. Van Gool. Improved multi-person tracking with active occlusion handling. In *Proceedings of the IEEE ICRA 2009 Workshop on People Detection and Tracking*, Kobe, Japan, 2009.

[Fathi et al., 2012] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1226–1233, Providence, RI, USA, 2012.

[Fei-Fei et al., 2003] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1134–1141, Nice, France, 2003.

[Felzenszwalb et al., 2008] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–8, Anchorage, AK, USA, 2008.

[Fod et al., 2002] A. Fod, A. Howard, and M. Mataríc. Laser-based people tracking. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 3024–3029, Washington, DC, USA, 2002.

[Foka and Trahanias, 2002] A. F. Foka and P. E. Trahanias. Predictive autonomous robot navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 490–495, Lausanne, Switzerland, 2002.

[Fortuna and Capson, 2004] J. Fortuna and D. Capson. Ica filters for lighting invariant face recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 334–337, Washington, DC, USA, 2004.

[Freund and Schapire, 1997] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119–139, August 1997.

[Ganapathi et al., 2010] V. Ganapathi, C. Plagemann, S. Thrun, and D. Koller. Real time motion capture using a single time-of-flight camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 755–762, San Francisco, CA, USA, 2010.

[Ge et al., 2009] W. Ge, R. T. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 1–8, Snowbird, UT, USA, 2009.

[Ghahramani, 2004] Z. Ghahramani. *Unsupervised Learning*. Springer, 2004.

[Gidel et al., 2008] S. Gidel, P. Checchin, C. Blanc, T. Chateau, and L. Trassoudaine. Pedestrian detection method using a multilayer laserscanner: Application in urban environment. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 173–178, Nice, France, 2008.

[Gidel et al., 2010] S. Gidel, P. Checchin, C. Blanc, T. Chateau, and L. Trassoudaine. Pedestrian detection and tracking in an urban environment using a multilayer laser scanner. *Transactions on Intelligent Transportation Systems*, 11(3):579–588, 2010.

[Gonzalez-Barbosa and Lacroix, 2002] J.-J. Gonzalez-Barbosa and S. Lacroix. Rover localization in natural environments by indexing panoramic images. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 1365–1370, Washington, DC, USA, 2002.

[Grabner and Bischof, 2006] H. Grabner and H. Bischof. On-line boosting and vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 260–267, New York, NY, USA, 2006.

[Groh et al., 2010] G. Groh, A. Lehmann, J. Reimers, M. R. Friess, and L. Schwarz. Detecting social situations from interaction geometry. In *Proceedings of the IEEE International Conference on Social Computing (SocialCom)*, 1–8, Minneapolis, MN, USA, 2010.

[Hartigan, 1975] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, 1975.

[Harville et al., 1998] D. G. Harville, T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *International Journal of Computer Vision (IJCV)*, 601–608, 1998.

[Helbing et al., 2002] D. Helbing, I. J. Farkás, P. Molnár, and T. Vicsek. Simulation of pedestrian crowds in normal and evacuation situations. In M. Schreckenberg and S. D. Sharma, editors, *Pedestrian and Evacuation Dynamics*, 21–58. Springer, Berlin, Germany, 2002.

[Helbing et al., 2000] D. Helbing, I. J. Farkás, and T. Viscek. Simulating dynamical features of escape panic. *Nature*, 407:487–490, 2000.

[Hoogendoorn and Bovy, 2000] S. P. Hoogendoorn and P. H. L. Bovy. Gas kinetic modelling and simulation of pedestrian flows. *Transportation Research Record*, 1710:28–36, 2000.

[Ilg et al., 2004] W. Ilg, G. H. Bakir, J. Mezger, and M. A. Giese. On the representation, learning and transfer of spatio-temporal movement characteristics. *International Journal of Humanoid Robotics (IJHR)*, 1(4):613–636, 2004.

[Jacobi, 1865] C. G. J. Jacobi. De investigando ordine systematis aequationum differentialum vulgarium cujuscunque. *Borchardt Journal für die reine und angewandte Mathematik*, LXIV(4): 297–320, 1865.

[Jenkins and Matarić, 2004] O. C. Jenkins and M. J. Matarić. A spatio-temporal extension to Isomap nonlinear dimension reduction. In *International Conference on Machine Learning (ICML)*, 441–448, Banff, Alberta, Canada, 2004.

[Jonker and Volgenant, 1987] R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

[Kass and Raftery, 1995] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:733–795, 1995.

[Katz et al., 2008] R. Katz, J. Nieto, and E. M. Nebot. Probabilistic scheme for laser based motion detection. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 161–166, Nice, France, 2008.

[Khan et al., 2006] Z. Khan, T. Balch, and F. Dellaert. MCMC data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 28(12):1960–72, December 2006.

[Kleinhagenbrock et al., 2002] M. Kleinhagenbrock, S. Lang, J. Fritsch, F. Lömker, G. Fink, and G. Sagerer. Person tracking with a mobile robot based on multi-modal anchoring. In *IEEE International Workshop on Robot and Human Interactive Communication (ROMAN)*, 423–429, Berlin, Germany, 2002.

[Kluge et al., 2001] B. Kluge, C. Köhler, and E. Prassler. Fast and robust tracking of multiple moving objects with a laser range finder. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 1683–1688, Seoul, Korea, 2001.

[Ko et al., 2012] M. Ko, T. Kim, and K. Sohn. Calibrating a social-force-based pedestrian walking model based on maximum likelihood estimation. *Transportation*, 1–17, May 2012.

[Kruger et al., 2006] V. Kruger, S. Zhou, and R. Chellappa. Integrating video information over time. example: Face recognition from video. In *Cognitive Vision Systems*, 127–144. Springer, 2006.

[Kulic et al., 2008] D. Kulic, W. Takano, and Y. Nakamura. Incremental learning, clustering, and hierarchy formation of whole body motion patterns using adaptive hidden markov chains. *International Journal of Robotics Research (IJRR)*, 27(7):761–784, 2008.

[Kurien, 1990] T. Kurien. Issues in the design of practical multitarget tracking algorithms. In Y. Bar-Shalom, editor, *Multitarget-Multisensor Tracking: Advanced Applications*, 43–83. Artech House, 1990.

[Kwok and Fox, 2005] C. Kwok and D. Fox. Map-based multiple model tracking of a moving object. In *RoboCup 2004: Robot Soccer World Cup VIII*, 18–33, 2005.

[Lau et al., 2009] B. Lau, K. O. Arras, and W. Burgard. Tracking groups of people with a multi-model hypothesis tracker. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 3487–3492, Kobe, Japan, 2009.

[Lau et al., 2010] B. Lau, K. O. Arras, and W. Burgard. Multi-model hypothesis group tracking and group size estimation. *International Journal of Social Robotics*, 2(1):19–30, March 2010.

[Lawrence, 2005] N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[Leal-Taixé et al., 2011] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *1st Workshop on Modeling, Simulation and Visual Analysis of Large Crowds (ICCV)*, 120–127, 2011.

[Leibe et al., 2008] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, (10):1683–1698, 2008.

[Leibe et al., 2005] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 878–885, San Diego, CA, USA, 2005.

[Liao et al., 2003] L. Liao, D. Fox, J. Hightower, H. Kautz, and D. Schulz. Voronoi tracking: Location estimation using sparse and noisy sensor data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 723–728, 2003.

[Liao et al., 2007] L. Liao, D. J. Patterson, D. Fox, and H. A. Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, 171(5-6):311–331, 2007.

[Lin et al., 2004] L. Lin, Y. Bar-Shalom, and T. Kirubarajan. New assignment-based data association for tracking move-stop-move targets. *IEEE Transactions on Aerospace and Electronic Systems*, 40 (2):714–725, 2004.

[MacLachlan and Mertz, 2006] R. MacLachlan and C. Mertz. Tracking of moving objects from a moving vehicle using a scanning laser rangefinder. In *Intelligent Transportation Systems*, 301–306. IEEE, 2006.

[Mallick et al., 2011] M. Mallick, S. Rubin, and J. Laval. N-body filtering for road tracking using a car following model. In *Proceedings of the International Conference on Information Fusion (FUSION)*, 1–8, 2011.

[Martínez Mozos et al., 2009] O. Martínez Mozos, R. Kurazume, and T. Hasegawa. Multi-layer people detection using 2D range data. In *Proceedings of the ICRA 2009 Workshop: People Detection and Tracking*, Kobe, Japan, 2009.

[Martínez Mozos et al., 2010] O. Martínez Mozos, R. Kurazume, and T. Hasegawa. Multi-part people detection using 2d range data. *International Journal of Social Robotics*, 2(1):31–40, 2010.

[Mazor et al., 1998] E. Mazor, A. Averbuch, Y. Bar-Shalom, and J. Dayan. Interacting multiple model methods in target tracking: A survey. *IEEE Transactions on Aerospace and Electronic Systems*, 34(1):103–123, Jan 1998.

[McPhail and Wohlstein, 1982] C. McPhail and R. T. Wohlstein. Using film to analyze pedestrian behavior. *Sociological Methods & Research*, 10(3):347–375, 1982.

[Merzbach and Nualart, 1986] E. Merzbach and D. Nualart. A characterization of the spatial poisson process and changing time. *Annals of Probability*, 14(4):1380–1390, 1986.

[Miller et al., 1997] M. L. Miller, H. S. Stone, , I. J. Cox, and I. J. Cox. Optimizing murty's ranked assignment method. *IEEE Transactions on Aerospace and Electronic Systems*, 33:851–862, 1997.

[Montemerlo and Thrun, 2002] M. Montemerlo and S. Thrun. Conditional particle filters for simultaneous mobile robot localization and people tracking. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 695–701, 2002.

[Moravec and Elfes, 1985] H. Moravec and A. E. Elfes. High resolution maps from wide angle sonar. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 116 – 121, St Louis, MO, USA, 1985.

[Moussaïd et al., 2010] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PLoS ONE*, 5 (4):e10047, April 2010.

[Mucientes and Burgard, 2006] M. Mucientes and W. Burgard. Multiple hypothesis tracking of clusters of people. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 692–697, Beijing, China, 2006.

[Munaro et al., 2012] M. Munaro, F. Basso, and E. Menegatti. Tracking people within groups with RGB-D data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2101–2107, Vilamoura, Portugal, 2012.

[Munkres, 1957] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

[Murty, 1968] K. G. Murty. An algorithm for ranking all the assignments in order of increasing cost. *Operations Research*, 16, 1968.

[Nascimento et al., 2012] E. R. Nascimento, G. L. Oliveira, M. F. M. Campos, A. W. Vieira, and W. R. Schwartz. BRAND: A robust appearance and depth descriptor for RGB-D images. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1720–1726, Vilamoura, Portugal, 2012.

[Navarro-Serment et al., 2010] L. E. Navarro-Serment, C. Mertz, and M. Hebert. Pedestrian detection and tracking using three-dimensional LADAR data. *International Journal of Robotics Research (IJRR)*, 29(12):1516–1528, October 2010.

[Ng and Lee, 1996] H. T. Ng and H. B. Lee. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the Association for Computational Linguistics*, 40–47, Santa Cruz, CA, USA, 1996.

[Oh et al., 2004] S. Oh, S. Russell, and S. Sastry. Markov chain monte carlo data association for general multiple-target tracking problems. In *Proceedings of the IEEE Conference on Decision and Control*, 735–742, Atlantis, Bahamas, 2004.

[Oza and Russell, 2001] N. C. Oza and S. Russell. Online bagging and boosting. In *Artificial Intelligence and Statistics*, 105–112, 2001.

[Pedersen et al., 2008] C. R. Pedersen, L. Relund Nielsen, and K. A. Andersen. An algorithm for ranking assignments using reoptimization. *Computers and Operations Research*, 35(11):3714–3726, November 2008.

[Pellegrini et al., 2009] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 261–268, Kyoto, Japan, 2009.

[Pellegrini et al., 2010] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 452–465, Heraklion, Greece, 2010.

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

[Plagemann et al., 2005] C. Plagemann, T. Müller, and W. Burgard. Vision-based 3d object localization using probabilistic models of appearance. In *Pattern Recognition, 27th DAGM Symposium, Vienna, Austria*, volume 3663, 184–191. Springer-Verlag, 2005.

[Platt, 2000] J. C. Platt. *Advances in Large-Margin Classifiers: Probabilities for SV Machines*, 61–74. MIT Press, 2000.

[Qin and Shelton, 2012] Z. Qin and C. R. Shelton. Improving multi-target tracking via social grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1972–1978, Providence, RI, USA, 2012.

[Ramanan et al., 2007] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 29 (1):65–81, 2007.

[Reid, 1979] D. B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 24(6), 1979.

[Robin et al., 2009] T. Robin, G. Antonini, M. Bierlaire, and J. Cruz. Specification, estimation and validation of a pedestrian walking behavior model. *Transportation Research Part B: Methodological*, 43(1):36–56, 2009.

[Rong Li and Jilkov, 2003] X. Rong Li and V. P. Jilkov. Survey of maneuvering target tracking. Part I: Dynamic models. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1333–1364, 2003.

[Schadschneider et al., 2009] A. Schadschneider, W. Klingsch, H. Klüpfel, T. Kretz, C. Rogsch, and A. Seyfried. Evacuation dynamics: Empirical results, modeling and applications. In *Encyclopedia of Complexity and Systems Science*, 3142–3176. 2009.

[Schapire and Singer, 1999] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[Schulz, 2006] D. Schulz. A probabilistic exemplar approach to combine laser and vision for person tracking. In *Robotics: Science and Systems*. The MIT Press, 2006.

[Schulz et al., 2001] D. Schulz, W. Burgard, D. Fox, and A. Cremers. Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 1665–1670, Seoul, Korea, 2001.

[Schulz et al., 2003] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. People tracking with mobile robots using sample-based joint probabilistic data association filters. *International Journal of Robotics Research (IJRR)*, 22(2):99–116, 2003.

[Smith et al., 1990] R. Smith, M. Self, and P. Cheeseman. Estimating uncertain spatial realtionships in robotics. In *Autonomous Robot Vehicles*, 167–193. Springer Verlag, 1990.

[Song et al., 2010] X. Song, H. Zhao, J. Cui, X. Shao, R. Shibasaki, and H. Zha. Fusion of laser and vision for multiple targets tracking via on-line learning. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 406–411, Anchorage, AK, USA, 2010.

[Spinello and Arras, 2011] L. Spinello and K. O. Arras. People detection in RGB-D data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3838–3843, San Francisco, CA, USA, 2011.

[Spinello et al., 2010a] L. Spinello, K. O. Arras, R. Triebel, and R. Siegwart. A layered approach to people detection in 3D range data. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Atlanta, USA, 2010.

[Spinello et al., 2011] L. Spinello, M. Luber, and K. O. Arras. Tracking people in 3D using a bottom-up top-down detector. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 1304–1310, Shanghai, China, 2011.

[Spinello et al., 2010b] L. Spinello, R. Triebel, and R. Siegwart. Multiclass multimodal detection and tracking in urban environments. *International Journal of Robotics Research (IJRR)*, 29(12): 1498–1515, 2010.

[Streit and Luginbuhl, 1995] R. Streit and T. Luginbuhl. Probabilistic multi-hypothesis tracking. Technical report, Naval Underwater Systems Center, Newport, SI, USA, 1995.

[Tang and Tao, 2008] F. Tang and H. Tao. Probabilistic object tracking with dynamic attributed relational feature graph. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8): 1064–1074, 2008.

[Tasoulis et al., 2006] D. K. Tasoulis, N. M. Adams, and D. J. Hand. Unsupervised clustering in streaming data. In *International Conference on Data Mining - ICDM Workshops*, 638–642, Washington, DC, USA, 2006.

[Taylor and Kleeman, 2004] G. Taylor and L. Kleeman. A multiple hypothesis walking person tracker with switched dynamic model. In *Proceedings of the Australasian Conference on Robotics and Automation*, Canberra, Australia, 2004.

[Tenenbaum et al., 2000] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

[Thrun, 2001] S. Thrun. An online mapping algorithm for teams of mobile robots. *International Journal of Robotics Research (IJRR)*, 20(5):335–363, 2001.

[Topp and Christensen, 2005] E. Topp and H. Christensen. Tracking for following and passing persons. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2321–2327, Alberta, Canada, 2005.

[Toyama and Blake, 2002] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision (IJCV)*, 48(1):9–19, June 2002.

[Treptow and Zell, 2004] A. Treptow and A. Zell. Real-time object tracking for soccer-robots without color information. *Journal of Robotics and Autonomous Systems*, 48(1):41–48, 2004.

[Vasquez et al., 2009] D. Vasquez, T. Fraichard, and C. Laugier. Incremental learning of statistical motion patterns with growing hidden markov models. *Transactions on Intelligent Transportation Systems*, 10(3):403–416, 2009.

[Verschae et al., 2008] R. Verschae, J. R. del Solar, and M. Correa. A unified learning framework for object detection and classification using nested cascades of boosted classifiers. *Machine Vision and Applications*, 19(2):85–103, 2008.

[Viola and Jones, 2001] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:511–518, 2001.

[Viola and Jones, 2002] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2002.

[Wang et al., 2010] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: recognizing people and social relationships. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 169–182, Heraklion, Greece, 2010.

[Wang et al., 2006] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models. In *Advances in Neural Information Processing Systems 18*, 1441–1448. The MIT Press, 2006.

[Wang and Han, 2005] Y. Wang and J.-Q. Han. Iris recognition using independent component analysis. *Machine Learning and Cybernetics, 2005.*, 7:4487–4492, 2005.

[Wieneke and Willett, 2008] M. Wieneke and P. Willett. On track-management within the PMHT framework. In *Proceedings of the International Conference on Information Fusion (FUSION)*, 1–8, Cologne, Germany, 2008.

[Wolf et al., 2005] J. Wolf, W. Burgard, and H. Burkhardt. Robust vision-based localization by combining an image retrieval system with monte carlo localization. *IEEE Transaction on Robotics and Automation (TRO)*, 21(2):208–216, 2005.

[Wolf and Burdick, 2009] M. T. Wolf and J. W. Burdick. Multiple hypothesis tracking using clustered measurements. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 3955–3961, Kobe, Japan, 2009.

[Wren et al., 1997] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentl. Pfinder: Real-time tracking of the human body. *IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)*, 19: 780–785, 1997.

[Xavier et al., 2005] J. Xavier, M. Pacheco, D. Castro, and A. Ruano. Fast line, arc/circle and leg detection from laser scan data in a player driver. In *Proceedings of the International Conference on Robotics & Automation (ICRA)*, 3930–3935, Barcelona, Spain, 2005.

[Yu et al., 2009] T. Yu, S. N. Lim, K. A. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1462–1469, Miami Beach, FL, USA, 2009.

[Yücel et al., 2012] Z. Yücel, F. Zanlungo, T. Ikeda, T. Miyashita, and N. Hagita. Modeling indicators of coherent motion. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2134–2140, Vilamoura, Portugal, 2012.

[Ziebart et al., 2009] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa. Planning-based prediction for pedestrians. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3931–3936, St. Louis, MO, USA, 2009.