TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Echtzeitsysteme und Robotik

Multiple People Tracking-by-Detection in a Multi-Camera Environment

Lili Chen

Vollständiger Abdruck der von der Fakultät der Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr.-Ing. Nils Thürey
Prüfer der Dissertation: 1. Univ.-Prof. Dr.-Ing. habil. Alois Knoll
2. Univ.-Prof. Dr.-Ing. Rüdiger Dillmann, Karlsruher Institut für Technologie

Die Dissertation wurde am 30.04.2014 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 26.11.2014 angenommen.

Abstract

Automatic vision-based people tracking system is getting ubiquitous and promises to be the key for a large variety of domains, including video surveillance, cognitive human-robot interaction, vision-based sport analysis, automotive driving safety assistance, and so on.

To develop such a system, there are two major issues involved from the perspective of computer vision: one is to accurately detect objects of interest, the other is to robustly track them across frames and maintain their corresponding correct identity. However, there are numerous challenges, for instance, varying appearance/motion/pose of people, uncontrolled environmental illumination conditions, bad weather effects, cluttered and dynamic changing background, partial or even full occlusions, long-term interaction, grouping and splitting, etc.

This thesis focuses on developing a complete system for long-term detection and tracking of an a-priori unknown number of people, walking randomly in a complex and crowded multi-camera indoor/outdoor environment with all those challenges arise, particularly aiming to reduce identity switches, miss-detections and false positives into minimum amount. We propose an unified framework for detecting and tracking people on the basis of a hierarchical grid-based, globally optimal tracking-by-detection strategy. Frame-by-frame detection is performed by means of a hierarchical grid-based methodology. We demonstrate that it can yield nice results for detecting and localizing targets with no prior knowledge in cluttered scene, while dealing fairly well with the challenges including complex interactions, mutual occlusions, illumination changes. The tracking problem then can be achieved by linking detection across frames, which is formulated as a grid-based network flow model with the discretized state-space, resulting in a convex problem casted into an Integer Linear Programming (ILP) form and solved through relaxation, therefore providing a global optimal solution while optimizing all the trajectories simultaneously. In

addition, we show that a finer analysis of the human behavior can contribute to better understanding of people attention and can be used to analyze the case of close interaction therefore enhancing the tracking performance. Thus we integrate a behavior cue (body orientation) into our tracking framework, providing valuable hints for resolving ambiguities between crossing trajectories.

A novel performance evaluation framework is also proposed to quantitatively evaluate the performance of the proposed system through experiments on a large variety of benchmark video sequences, including both indoor and outdoor scenarios. Within this evaluation framework, the ground truth of each sequence has been annotated in 3D space. The performance is evaluated based on a series of metrics, which gives out an intuitive measure of the detector and tracker's performance at detecting objects, localizing objects, keeping their identities, and so on. More importantly, with these evaluation metrics, it provides a much easier way to compare the proposed system to the state-of-the-art, so that clearly indicating respective strengths.

Zusammenfassung

Automatische Vision-basierte System zur Personenverfolgung werden allgegenwärtig. Sie sind der Schlüssel in einer Vielzahl verschiedener Anwendungen wie Videoüberwachung, kognitiver Mensch-Roboter Interaktion, bildgestützter Sportanalyse und Fahrerassistenzsysteme.

Zur Entwicklung derartiger Systeme gibt es zwei wesentliche Aspekte aus Sicht des maschinellen Sehens: Der eine ist die akkurate Erkennung von Objekten. Der andere ist die robuste Verfolgung dieser Objekte über mehrere Rahmen hinweg sowie die Erhaltung der zugehörigen, korrekten Identität. Jedoch gibt es zahlreiche, schwierige Herausforderungen, wie beispielsweise die Variation des Erscheinungsbildes/ der Bewegung/ der Pose der Personen, unkontrollierte Beleuchtungsbedingungen in der Umwelt, schlechte Wettereffekte, überfüllte und dynamische Hintergründe, partielle oder sogar volle Verdeckungen, Langzeit-Interaktionen, Gruppierungen und Zersplittungen.

Die vorliegende Arbeit konzentriert sich auf die Entwicklung eines vollständigen Systems zur Langzeiterkennung und Tracking einer vorher unbekannten Anzahl von Personen, die sich in einer komplexen und überfüllten Innen- oder Außenbereich unter Beobachtung von mehrerer Kameras bewegen. Die Arbeit behandelt dabei zahlreiche Fragestellungen, insbesondere die Reduzierung der Identitätswechsel, der inkorrekten Erkennungen und der falsch positiven Ergebnisse auf einen minimalen Betrag. Es wird ein vereinender Rahmen zur Erkennung und Tracking von Personen auf der Basis einer hierarchischen gitterbasierten, global optimierten Tracking-durch-Detektion Strategie vorgestellt. Die Frame für Frame Erkennung durch eine hierarchische, gitterbasierte Vorgehensweise durchgeführt. Es wird demonstriert, dass gute Ergebnisse zur Erkennung und Lokalisierung von Zielen ohne Vorwissen in überfüllten Szenen unter Beachtung der beschriebenen Herausforderungen inklusive komplexer Interaktionen, wechselseitigen Verdeckungen und Beleuchtungsveränderungen erzielt werden können. Das Tracking-Problem kann dann durch die

Erkennung von Verknüpfung über Frames erreicht werden. Es wird als gitterbasiertes Netzwerk-Flussmodell mit diskreten Zustandsräumen formuliert, was zu einer konvexen Problem in eine Interger Linear Programming(ILP) Form gegossen und durch Relaxation gelöst. Eine global optimierte Lösung mit simultaner Optimierung aller Trajektorien wird folglich zur Verfügung gestellt. Zusätzlich wird gezeigt, dass eine feinere Analyse des Verhaltens des Menschen zu einem besseren Verständnis der Menschen beitragen kann, und zur Analyse im Fall von direkten Interaktionen verwendet werden kann and erhöhen damit die Trackingleistung. Ein Verhaltenssignal (Körperorientierung) wird daher in den Trackingrahmen integriert, das wichtige Hinweise zur Auflösung von Mehrdeutigkeiten zwischen sich überkreuzenden Trajektorien zur Verfügung stellt.

Die Arbeit stellt auch einen neuen Evaluierungsrahmen vor, um die Leistung des präsentierten Systems durch Experimente mit einer großen Viefalt von Benchmark-Videosequenzen, inklusive Innen- und Außenszenarien, quantitativ zu evaluieren. Die Referenzdaten jeder Sequenz im 3D Raum wurden innerhalb dieses Evaluierungsrahmens beschriftet. Die Evaluierung der Leistung basiert auf Standardmetriken, die eine intuitive Messung der Leistung des Erkennungs- und Trackingsystems im Hinblick auf die Erkennung der Objekte, der Lokalisierung der Objekte, der Erhaltung der Identiät, usw., erlaubt. Noch wichtiger ist, mit diesen Standardevaluierungsmetriken, bietet es eine viel einfachere Möglichkeit, zu vergleichen unser System auf die State-of-the-Art, um so das jeweiligen Stärken deutlich zu anzeigt.

Acknowledgements

There are lots of people I would like to thank for a huge variety of reasons.

First and foremost, I would like to thank Prof. Dr. Alois Knoll for providing me with the opportunity and excellent international working environment to conduct my research in his prestigious group, supporting me the freedom to pursue my research interest. My sincere thanks go to Prof. Dr. Rüdiger Dillmann for reviewing my thesis. I would also like to express my gratitude towards Dr. Giorgio Panin, for guiding and inspiring me from the very beginning of my PhD study, for all the time we discuss and work together, providing excellent advices and comments, and also for all the time he spent on reading my papers.

I would like to thank my colleagues at the OpenTL team for their interlectual support. A special thanks to Dr. Suraj Nair, Dr. Emmanuel Dean, Dr. Claus Lenz, Thorsten Röder, Sebastian Klose, Philipp Heise, Caixia Cai, M. Ali Nasseri, Martin Eder, for fruitful discussions and the help on recording the experimental datasets. I would also like to thank Amy Bücherl, Gertrud Eberl, Marie-Luise Neitz for their kind support for all administrative affairs and the creation of nice and friendly atmosphere to work in.

My gratitude also goes to Yang Chen, Jia Huang, Gang Chen, Kai Huang, Dr. Chih-Hong Cheng, my PhD study has been made not only a great learning experience but also enjoyable by the interactions with them.

I would also like to show my great gratitude to Prof. Chunxia Zhao, who was my supervisor during my master studies, for leading me on the way of research. She knew how to encourage me and how to fulfill my potential, I do not think I could keep to the research road without her encouragement and support.

Finally, I am very grateful to my family, first of all to my parents, for all the opportunities they created for me and their continuing support. And great thanks to my husband, who has been always very kind to support and encourage me every

day for my research work, especially for all of the great discussions about research. Particularly to my girl, our little angel, brings me and the family so much enjoyable moments. Without their consistent love and support, any of my achievements would have not been possible.

Contents

List of Figures

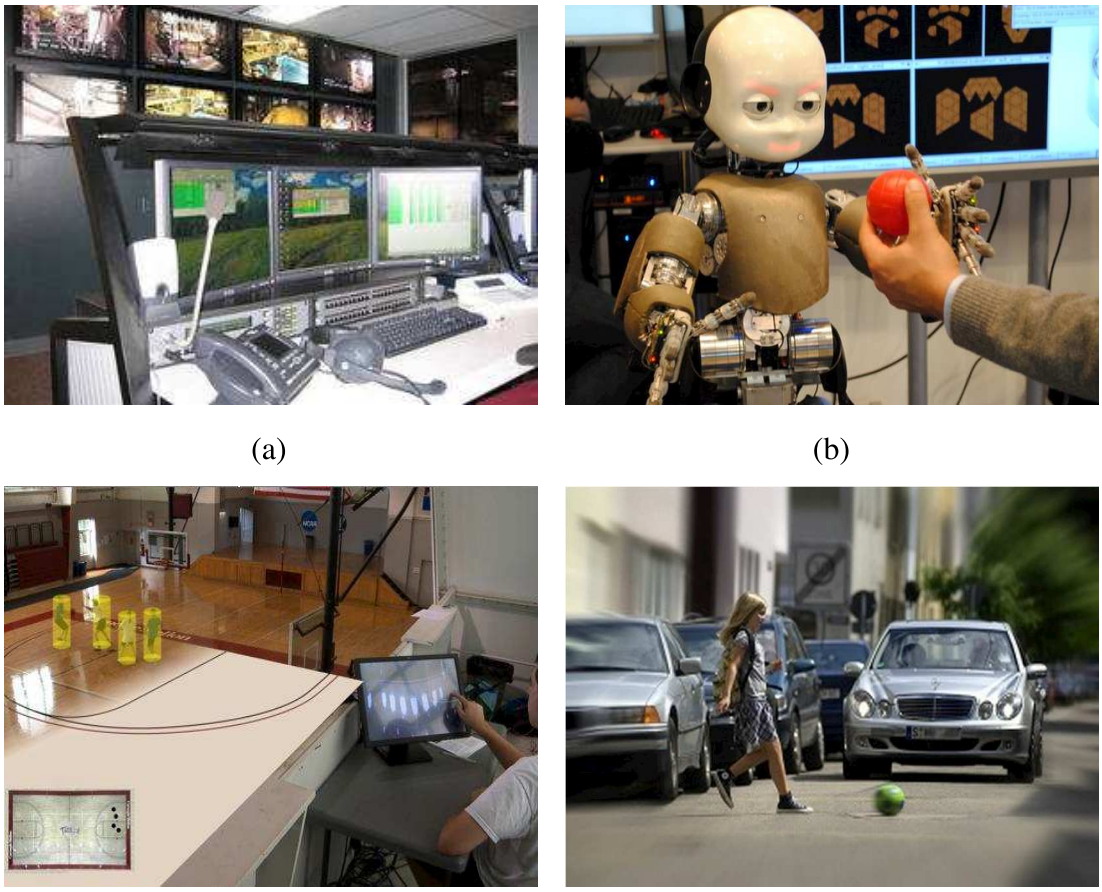
List of Tables

Chapter 1

Introduction

Due to the fast development of high powerful computers, availability of widespread high-resolution and low-cost vision sensors, such as CCD or CMOS, as well as the increasing demand for automatic video analysis, a great deal of interest in the research area of computer vision has been generated. Following the tendency, intelligent camera systems are getting ubiquitous, in particular an automatic vision-based people tracking system becomes increasingly popular and promises to be the key for a large variety of domains, including automatic video surveillance [1–7] (Fig. 1.1(a)), cognitive human-robot interaction [8–12] (Fig. 1.1(b)), vision-based sport analysis [13–17] (Fig. 1.1(c)), automotive driving assistance system [18–23](Fig. 1.1(d)), etc. As for video surveillance, reliably locating and tracking people in video can facilitate human behavior understanding for better event detection; tracking in real-time from a robot can form the basis for human-robot interaction and robot's more efficient performance in human environments; automatically finding the players, finding the paths of players' movement, distinguishing players from each other and quantifying their ability can offer significant help to the sports expert; detecting and tracking pedestrians based on a car's driving assistance system can help people drive safely, as the system can give the assessment of various of dangers, which gives the driver correct instructions for collision avoidance, so that keeping people safe in the presence of autonomous cars.

From the perspective of computer vision, the major subissues within the automatic vision-based tracking system is how to firstly detect objects of interest (i.e., find the image regions corresponding to the objects), and how is to track them across different frames while maintaining their corresponding correct identities. Due to people's huge variations in physical appearance, pose, movement and interaction, or even partially and also fully occluded for long period of time,



(a) (b)

(c) (d)

Figure 1.1: Sample of significant applications for automatic vision-based tracking system. (a) Automatic video surveillance. (b) Cognitive human robot interaction. (c) Vision-based sport analysis. (d) Automotive driving assistance system.

implementing such a system is obviously very challenging thus receives a significant amount of attention in the area of research and development.

Huge research efforts have been made to aim at increasing the performance of tracking system and making more effective use of cameras to reduce the workload of human resources. There are numerous approaches that have been proposed to tackle the problems, they differentiate from each other by different features that are used, different object representation, different modeling method of motion, appearance and shape. Even though there has been great progress during the past decades due to advances in computing power, image quality, and developed algorithms, the performance and accuracy of multiple people tracking systems is still far from being satisfactory, it still can not be able to rival the astonishing ease with how human beings accomplish the same task. It is due to three major reasons: 1) developing a robust and general object detection method still remains an open problem. In particular, people detection is a special case of generic object category detection, therefore it inherits all the difficulties common to the problem of object detection, for example, the necessity to deal with large amount of background clutter, the variability of illumination conditions and the limited number of training samples. Additionally, due to the complex structure and varying appearance of human body, it is very challenging to extract the correct image regions belong to body part of the people from background clutters. Furthermore the problem gets to be even more complicated when multiple people are present in the scene; 2) even after the problem of people detection is solved, we are still facing the ambiguities arising from maintaining consistent tracks from detections, in particular when the scenarios involves significant occlusions, such as self-occlusion, inter-object occlusion, or static occluders within the scene. Most works deal with those cases by only concentrating on a small time window along the whole trajectories and do not look for a joint global optimum, they are therefore prone to mistakes such as identity switches. Moreover if objects do not have distinctive appearance among each other then it would make tracking even more difficult; 3) almost none of the works improve the tracking performance from the perspective of human behavior, which conveys the most valuable information about the person's current and also future activities. Therefore, this important cue, compared to the common used cues such as appearance or motion, can contribute much more to enhance tracking performance, especially when discovering the cases of complex interaction and significant mutual occlusions.

Therefore, to develop such a tracking system is highly significant: that could firstly solve the detection issue efficiently (being able to detect people robustly in any situation without prior knowledge, no matter with the environment, illumination, appearance, pose, people density, etc)

and then consistently track targets over the whole tracking process with improved robustness against various challenges, particularly to exploit advantages from human behavior analysis, in order to enhance robustness against identity switches under the case of complex interaction and significant occlusions.

1.1 Objectives and Challenges

The overall objective of the thesis is to develop a framework for long-term detection and tracking of an a-priori unknown number of people with random walking in an overlapping, multi-camera indoor/outdoor environment. The obvious benefits of using multiple cameras are firstly increasing the coverage of the scene, since each of the combined field of views of all the cameras should be greater than that of any individual camera, then efficiently handle large number of occlusions, secondly increasing the possibility of preserving object identity across the region, and thirdly allowing to obtain accurate 3D localization of people. We then want to explore the benefits from behavior analysis by analyzing some representative cue (e.g. 3D human body orientation), to better understand people attention so that to disambiguate complex scenarios, therefore being able to serve for robustness enhancement of final tracking. Furthermore, to develop a global optimal scheme to look for joint global optimum among all trajectories so that improving robustness to wrong identity assignment, is also significant.

More specifically, we aim at:

- robustly detecting multiple people in spite of heavy occlusions, background clutters and environmental illumination changes in both indoor and outdoor scenarios;
- precisely locating people in 3D world coordinate;
- accurately tracking people with global optimum so that reducing identity switches, miss-detections and false positives into minimum amount;
- exploring representative cues (e.g. 3D human body orientation) of human behavior to improve tracking performance, by disambiguating complex scenarios such as long-term interactions and heavy mutual occlusions when general used features (e.g. appearance, motion) become unreliable;
- being capable of recovering from occasional tracking mistakes, so that can run over long time;

- quantitatively outputting the performance of the tracking system, aim to evaluate their precision in estimating object locations, their accuracy in recognizing object configurations, their ability to consistently label objects over time, etc.

To implement such a framework and achieve these goals is very difficult, which faces a large mount of difficulties and challenges. Fig. 1.2 shows several examples that illustrating the various challenges within indoor and outdoor scenarios. As can be seen, people's appearance could be able to vary quite largely, and their poses might change significantly. The environmental illumination conditions could be extremely terrible such as Fig. 1.2(a), the completely dark environment makes the detection very difficult. We can also see the occlusions due to target interaction and scene obstacles, for example the road sign forms the occlusion regions in Fig. 1.2(b). The targets might also exhibit very fast and unpredictable motion as in Fig. 1.2(c), then the tracking would be very difficult. Furthermore the uncontrolled outdoor illuminations and cluttered background as in Fig. 1.2(d) makes the detection and tracking ambiguous.

To be more comprehensive, the challenges are summarized as follows:

- People typically wear a large variety of clothing, can move fast and unpredictably, being capable of exhibiting numerous body poses, therefore can have much more uncertainty compared to other targets;
- Environmental illumination conditions and intensity can considerably change, resulting in drastic changes in target appearance. Particularly in outdoor environments where lighting conditions cannot be controlled and image intensities are subject to large changes in illumination variation.
- Video sequences, in particular with outdoor scenarios, might be of very low quality. For example, weather effects (e.g. dust, rain, snow), shadows can corrupt the images and make them difficult to interpret.
- Cluttered and dynamic changing background makes the foreground segmentation ambiguous.
- The tracking scenarios usually include partial or even full occlusions by other objects and obstacles.
- The targets might interact with each other for long term, and also frequently move in groups, which eventually split or re-merge with other groups. It becomes even difficult if they have similar appearance.

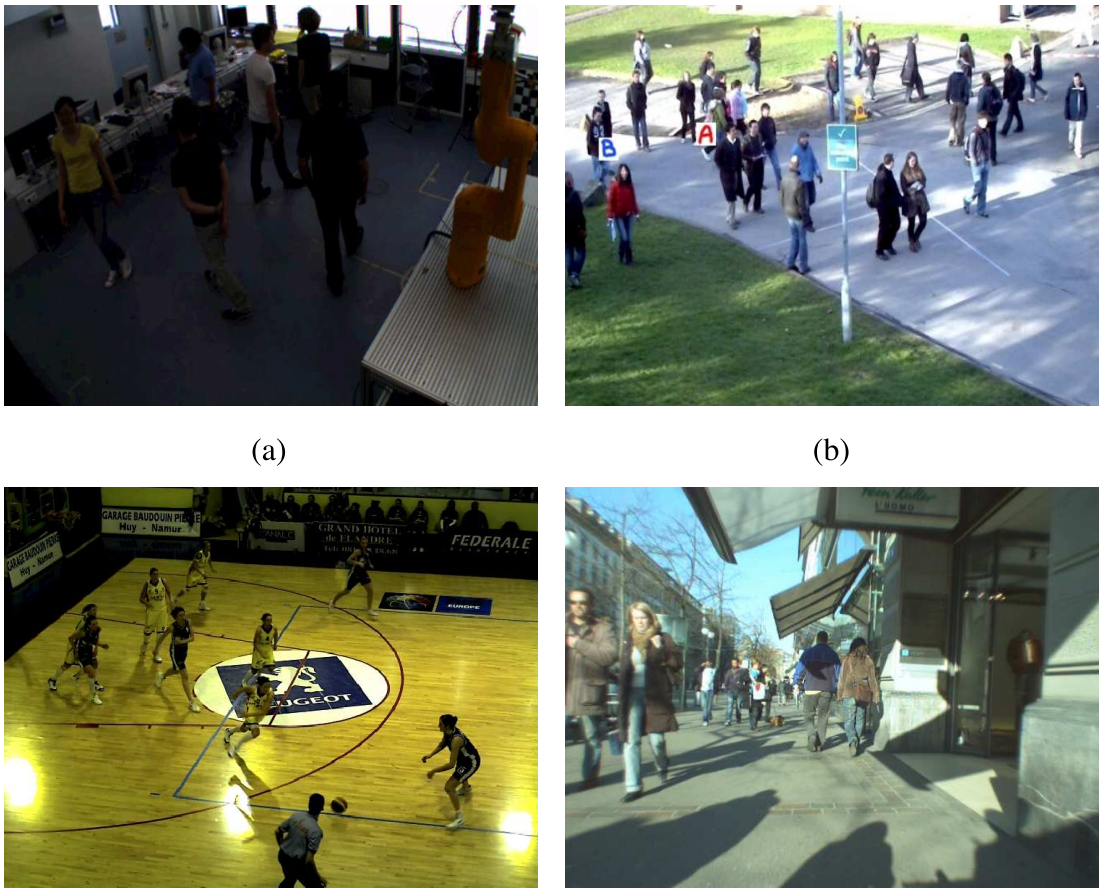


Figure 1.2: Typical examples of images from indoor and outdoor tracking scenarios that illustrate some of the challenges for people detection and tracking. (a) Extreme bad illumination. (b) High density of the targets, similar appearance with each other, frequently move in groups, self/mutual occlusions. (c) Fast and unpredictable motion. (d) Heavy cluttered background, shadows, uncontrolled illumination condition.

1.2 Main Contributions

The work presented in this thesis contributes to the research and application areas of multiple people detection, multiple people tracking as well as human body orientation estimation. The main contribution of our work is a complete system for tracking an unknown number of people within complex and crowded scenarios on the basis of tracking-by-detection strategy. Particularly, for better tracking performance, a finer analysis of human behavior (human body orientation) is integrated into this framework. The system is capable of automatically detecting people when they enter into the observing scene, estimating the 3D pose of human body orientation, and consistently tracking them over time, in particular dealing with long-term interactions and occlusions.

To summarize, the main contributions of this work thereby include:

- an unified framework for hierarchical grid-based multiple people tracking-by-detection with global optimization, with the advantage of being resistant to divergence and superior robustness against identity switches;
- an innovative tracking enhancement scheme by performing a finer analysis of individual human behavior (human body orientation), contributing to better understanding of people attention then can be used to analyze the case of close interaction;
- a hierarchical grid discretization scheme, the core for both detection and tracking;
- a hierarchical grid-based detection methodology, yielding nice results for detecting and localizing targets with no prior knowledge in cluttered scene;
- a robust edge-based background subtraction algorithm being insusceptible to illumination changes both for indoor and outdoor scenarios;
- a fast oriented distance transform, that efficiently matching foreground edges with model silhouettes, by considering not only the location of edge points but also their orientation, reducing a large amount of false alarms in presence of clutter;
- a hybrid 3D human body orientation estimation approach, dynamically combining the merits of a motion-based orientation estimator with a 3D appearance model-based orientation estimator, providing discriminative hints for the targets no matter they are moving, slowly moving or even still-standing;

- a global optimization framework, formulating the data association problem as finding the global maximum of a convex objective function, that globally optimizing all the trajectories;
- an association affinity model, incorporating the measurements on behavior cue, as well as location and appearance in a global manner, that efficiently enhance the tracking performance if any of the features becomes ambiguous;
- a performance evaluation framework based on standard metrics, that quantitatively evaluating the performance of the entire tracking system, giving out an intuitive measure of the detector and tracker's performance at detecting objects, localizing objects, keeping their identities, and so on.
- quantitative evaluation on a large variety of benchmark video sequences, including both indoor and outdoor scenarios. Comparisons to the state-of-the-art are also provided based on those standard evaluation metrics.

1.3 Outline of Thesis

The remainder of the thesis is organized as follows:

Chapter 2 provides an overview of the related work in the field of people detection, tracking, human body orientation estimation, and other works that influenced this thesis.

Chapter 3 presents the system architecture and detailed experimental set-up. We start by giving out the entire framework of the tracking system, the experimental platform that we used for implementation. Hardware setup used to capture the video sequences is described, the choice on datasets and detailed descriptions on corresponding attributes of each dataset is presented. Furthermore, a performance evaluation framework used for evaluation of the entire tracking system is discussed.

Chapter 4 describes the proposed hierarchical grid-based people detection. We investigate the use of template hierarchy and oriented distance transform in a people detection framework with a variable number of cameras, and demonstrate that it is capable of detecting and localizing people in 3D space efficiently and precisely without prior knowledge, towards the challenges of frequent mutual occlusions, cluttered background and environmental illumination changes.

Chapter 5 explains our hybrid strategy that combines a motion-based and 3D appearance-based orientation estimation approach dynamically, that being capable of working robustly

under the case of moving, slowly moving or even still-standing targets. With this strategy, we not only overcome the issue of automatic initialization of 3D appearance model, but also highly improve the runtime performance.

Chapter 6 describes our global optimization framework for long-term tracking of an a-priori unknown number of targets. And we show how to integrate the behavior cue (body orientation) into the association affinity model , efficiently provides valuable hints for resolving ambiguities between crossing trajectories. The strength of the global optimization framework is demonstrated by presenting its performance across several challenging datasets, with robustly tracking and identifying targets through complex interactions and significant mutual occlusions. Our method achieves consistent robustness and outperform state-of-the-art techniques in most cases.

Chapter 7 concludes this thesis, listing the key contributions and detailing a number of interesting issues that are left for future work. Some extensions of our approaches respect to other related computer vision issues are also discussed.

Chapter 2

Related Work

In this chapter we provide context for this research in the backdrop of previous work. Since the large amount of publications about people detection and tracking prohibits an exhaustive review of all related works, this chapter will focus on work that has had a significant impact on the field or is especially closely related to the work presented in this thesis. We present an overview of the developments and the most recent advances in the areas of people detection, human body orientation estimation and multiple people tracking. Strengths and potential deficiencies are discussed through a comparative analysis, which enable us to identify the keypoints that needed to be considered by this research.

2.1 People Detection

People detection has attracted an extensive amount of interest from computer vision community over the past few years [24–30]. Many techniques have been proposed in terms of features, models, and general architectures [31–37]. According to the number of cameras employed in the detection scenarios, these approaches can mainly be divided into two categories: monocular and multi-view approaches [38].

2.1.1 Monocular based Detection Approaches

Monocular approaches perform people detection by relying on the input of one single camera. In order to detect people, monocular based people detection methods generally scan the input image at all relevant positions and scales, and try to determine whether there is people or not. The monocular approaches that have been developed to date can be roughly classified into

two major categories. The first category is model-based approaches, for example, a model is firstly defined for the object of interest, and then the approach attempts to match the model to different parts within the image so that to find an appropriate matching [39, 40]. The second category is learning-based approaches, that learn discriminative features of a class from sets of labeled positive and negative samples [41].

Model-based Approaches Regarding the model-based approaches, among the various visual cues that can be used to match objects, shape has the advantage of powerful object discrimination capability which is relatively stable to environmental illumination changes. There are two representation methods to model the shape space: discrete and continuous approaches [42].

Within the discrete approaches, the shape manifold is represented by a set of exemplar shapes [18, 19, 43–45]. Efficient exemplar-based matching techniques based on distance transforms are combined with precomputed hierarchical structures, allowing for real-time online matching with thousands of exemplars [18, 43, 44, 46]. This technique yields nice results for locating targets with no prior knowledge in a cluttered scene. The efficiency of this method is illustrated by using about 4,500 templates to match pedestrians in images in [18]. The core idea is using a Chamfer distance measure, so that matching a template with the DT image results in a similarity measure. Meanwhile this approach enables the use of an efficient search algorithm. However, if only computing the location of edge pixels without considering their orientation when computing distance transform, it inevitably leads to a high rate of false alarms in presence of clutter. Another highlight of the work [18] is the utilization of a template hierarchy, which is generated automatically from available examples, and formed by a bottom-up approach using a partitioned clustering algorithm. It only searches locations where the distance measure is under a given threshold, so a speed-up of three orders of magnitude is demonstrated, compared to exhaustive searching. This idea was taken further by [44], that however does not build the template hierarchy (or tree) by bottom-up clustering, rather by partitioning a state-space represented with an integral grid. The grid is hierarchically partitioned as the search descends into each region, so that regions at the leaf-level define the finest partition. This method is demonstrated to be capable of covering 3D motion, even with self-occlusion. Although being demonstrated efficiently, these methods did not exhibit several important properties, for example, they did not employ any form of local gradient information, which is important for making the matching robust to image noises and clutters. Furthermore, these approaches unfortunately require a very specific model, which is only valid for specific target.

For the continuous approaches, the shape models are learned from a set of training shapes, if an appropriate manual of automatic shape registration method exists [42, 47–53]. Compared to discrete shape models, continuous generative models can fill gaps in shape representation by using interpolation. However, the online matching process proves to be much more complex as recovering an estimate of the maximum-a-posteriori model parameters involves iterative parameter estimation techniques [42].

Learning-based Approaches Different with model-based approaches, object detection is performed by learning different object views automatically from a set of examples by means of a supervised learning mechanism within the learning-based approaches. These learning based approaches rely on the fact that collected training data is representative of all relevant variations necessary to detection.

In context of object detection, the learning examples are composed of pairs of object features and an associated object class where both of these quantities are manually defined [54], *feature* and *classifier* are the two major component involved [55]. The *feature* component provides the visual appearance information of the person, and the *classifier* component determines whether there is person or not within the sliding window. The most commonly used features are Haar wavelets features [31, 32, 56, 57], shape contexts [58–60], code-book feature patches [34, 61–64], Histogram of Oriented Gradients (HOG) [33,65–68], edgelet features [69], shapelet features [70]. The classifier that are used are mainly Support Vector Machines (SVM) [31,33,71,72], AdaBoost using boosted detector cascades [32, 65, 69, 73, 74], convolutional neural network [75–77], and graphical models [34, 78–80].

Selection of features plays a key role in the performance of the classification, therefore, it is important to select the appropriate feature according to the application environment, so that discriminate one class from the other. Following we will have a short discussion on several feature and classifier representations.

Haar wavelet features have first been proposed by [31], and further adapted by [32, 40]. They introduce a dense overcomplete representation using wavelets, which represents local intensity differences at various locations, scales, and orientations. However, due to overlapping spatial shifts, the many-times redundant representation requires mechanisms to select the most appropriate subset of features out of the vast amount of possible features [42].

Code-book feature patches are also belong to the class of local intensity-based features [34,61,81,82]. A codebook of distinctive object feature patches along with geometrical relations

is learned from training data followed by clustering in the space of feature patches to obtain a compact representation of the underlying people class [42].

Histograms of Oriented Gradient (HOG) descriptors have been proposed in [33] for people detection, and obtained good results on multiple datasets. The HOG feature finds the spatial distribution of edge orientations. The HOG representation is based on the local gradient histograms, and the gradient of image intensity at each pixel contributes to each of the histograms of its adjacent cells through trilinear interpolation. Since the representation is based on the local distributions of gradient locations and orientations, so it appears to be more effective for modeling object appearance compared to Haar features, while at the same time being robust to noise and intra-class variability. And block normalization makes the HOG descriptor robust to illumination changes. However, while demonstrating excellent performance, HOG based approaches have difficulties handling detection of people in the presence of occlusions and in cases where people exhibit especially large pose variability.

As the most representative ones within various classifiers, Support Vector Machines (SVM) are demonstrated to be a powerful tool to solve pattern classification problems, it clusters data into two classes by finding the maximum marginal hyperplane that separates one class from the other. Linear SVMs have been successfully used by combining with various features [33, 83–86], while nonlinear SVMs yield further performance boosts however with much higher computernal costs [31, 40, 87–89]. Adaboost is an iterative method for finding a strong classifier based on a combination of moderately inaccurate weak classifier. It has been first introduced by [32] and then further applied in many other works [65, 69, 90–95].

However, supervised learning methods usually require a large collection of relevant training data from each object class, it is not only tedious but also often an ill-defined process because it is unclear that which part of the people class distributions is well-represented and which parts of the distribution are still insufficiently sampled [96].

Monocular approaches have the inherent advantage of simple and easy deployment, however have limited ability to handle occlusions due to several objects involved, as the single viewpoint is intrinsically not able to observe the hidden areas.

2.1.2 Multi-View based Detection Approaches

The utilization of multiple cameras provide a better solution to detect and locate multiple occluding people, as multiple camera views can be used to recover 3D structure information

and solve occlusion in crowded environments, also computing accurate 3D locations for targets in complex scenarios.

A majority of multi-view detection approaches rely on the *segment-then-locate* scheme, that detecting people by firstly obtaining foreground masks computed in multiple views [38, 97–106]. One of the key aspects within these approaches is how to define the correspondence between the foreground masks. Some works align the foreground masks by using homography constaints [107–111].

The work [107] first obtains the foreground likelihood maps in each view, then warping them from all the other views onto the reference view. These warped foreground likelihood maps are then multiplied to produce a 2D grid of occupancy likelihoods, called "synergy map". This map clearly highlights the feet regions of all the people in the scenario. Their extended work [108] localizes people on multiple planes parallel to the reference plane in the framework of plane to plane homologies.

A similar approach is used in [109], it extracts denser clusters in a ground plane occupancy map that is computed based on the projection of foreground masks. They do not consider projection on the ground plane only, but on a set of planes that are parallel to the ground plane, and cut the object to detect at different heights. And also they propose a heuristic-based method to combine the multiple projections thus generated.

The work [110] uses the homography constraint in several planes parallel to the ground plane, searching for people's heads in the higher planes. All camera views are mapped using homographies to a reference plane and intensity correlation is used to detect candidate heads. In [112], it uses the homography constraint within a particle filtering framework for locating the ground location of a person.

The homography constraint based methods to localize people on a ground plane can also be interpreted as a visual hull intersection process [107, 113–115]. They have the advantage of efficient computation, as the decision about ground plane occupancy is directly taken from the observation of the project of foreground masks from each view. However, the decision only relies on the part of the whole body (e.g. head, feet), but not the entire object silhouette, thus their methods is prone to cause many false positive errors due to shadows, reflections on the floor, and the density of the crowd.

An occupancy map is a plan-view representation of area of interest and allows for efficient aggregation of information coming from different views, usually about the presence of individuals. A Probabilistic Occupancy Map (POM) is proposed by [116], it assumes that the

objects are observed by multiple cameras at the head level. Within this work, the ground plane is firstly discretized into a regular grid and the occupancy probability of each grid cell is estimated using results from background subtraction. In particular, a simple rectangle model that approximates silhouettes is used to back-project the probabilities onto all views. The occupancy map is obtained by iterative optimization of the probability field, so that the difference between the back-projected and the input binary images is minimized. This method is mainly affected by the incorrect foreground blobs obtained by background subtraction (e.g., reflections or casted shadows). Similarly, [117] also obtains occupancy map from foreground images, by using sparsity-constrained inverse problem formulation. [118] presents a multi-modal fusion framework for simultaneous person detection and localization from multiple cameras in non-sequential manner. They encodes the multiple weak features as feature maps, which generalizes the concept of an occlusion map, in each cell the probability of occupancy is replaced with the value of a feature. [119] firstly extracts foreground objects and the silhouettes are used to compute a planar projection of the scene's visual hull. This projection is used to bound the number and possible location of people. In [120, 121], a 3-D Marked Point Process model is proposed to detect and localize people. The method extracts pixel-level features by projecting foreground silhouettes on the ground plane and the hypothetical head plane, and estimates the positions and heights of the objects by using a stochastic optimization process with geometric constraints.

In [122], instead of a *segment-then-locate* approach, they propose a *locate-then-segment* approach, which integrates available information of all cameras before any detection decision. Their method integrates the information of all parallel planes by projecting the foreground directly on the reference plane and accumulating the evidence from multiple cameras. Occlusion and people detection are solved simultaneously and instantly at each time, using the accumulated evidence from all cameras.

Compared to monocular detection, multi-view detection is capable of accurately localizing individuals on the 3D ground plane. Hence, it can be used for many other high-level vision tasks, such as multiple object tracking, people counting, scene understanding, and so on.

2.2 Human Body Orientation Estimation

Accurate estimation of human body orientation can significantly enhance the analysis of human behavior, which is important for improving tracking performance. To our knowledge, there are

few works have been conducted to visually estimate the orientation of human body. Generally, they can be divided into two categories up to the type of data they rely on: 2D based approaches and 3D based approaches.

2.2.1 2D based Human Body Orientation Estimation

Some works consider the task of human body orientation estimation as a classification problem. The orientation space is divided into n categories (i.e., eight categories), then specific visual features are extracted and orientation classifier are trained with these extracted features. Support Vector Machine (SVM) has been efficiently used for classifying between two or more classes, or estimating a continuous variable using regression. [123] use HoG descriptors to classify the orientation of pedestrians, recovering an estimate based on 2D low-resolution images. However, they group the orientation of a person very roughly, only covering a 45-degrees range for in-plane rotations. [84] estimates the pedestrian orientation using multiple SVM classifiers on Haar wavelet coefficients, in order to distinguish between different orientations. [124] proposes an idea of exploiting a modified Histogram of Oriented Gradients (HoG) descriptor with "bin shifting" technique to estimate the orientation of human. [125] estimates the human upper-body orientation by using a Support Vector Machine (SVM) on Histograms of Oriented Gradients (HoG) but replace the SVM by a decision tree with SVMs as binary decision makers. In [126], body pose classification is performed by using multi-level HoG features and a sparse representation technique at each frame of the sequence, and estimating body orientation in a temporal filtering framework. However, for each human region they end up with a very high dimensional feature vector, which has the drawback of computational complexity, and are limited to a single view experimental scenarios. To better fuse the body pose related features, the problem is further explored in [127], they address the issue as a joint model adaption problem in a semi-supervised framework. In the work of [128], pedestrian classification and orientation estimation problems are integrated into a set of view-related models. Their probabilistic model does not restrict the estimated pedestrian orientation to a fixed set of orientation classes but directly approximates the probability density of body orientation. However, above approaches are only based on static feature cues, ambiguities might arise up when discriminating the symmetric orientation.

In order to disambiguate these cases, the human dynamics has been considered to address this issue in some works. Some of them introduce body orientation as a link between the head pose and body movement cues [129, 130], they suppose the body orientation is similar to people's moving direction and use motion cues to estimate the orientation. The work of [131]

also assumes that the orientation is simply given by the walking direction. The estimation is based on the motion of a tracked person and the size of its bounding ellipse, however it fails in the case of people who are not moving, or slowly walking. These works, have not exploited body pose related features, therefore are problematic when the targets are static or move slowly, as the velocity becomes too noisy to provide reliable information for body pose estimation.

The work [132] proposes an approach combining Shape Context and SIFT features, initially the body orientation is calculated by matching the upper region of the body with predefined shape templates, then finding the orientation within the ranges of 22.5 degrees. The orientation is further refined by the optical flow vectors of SIFT features with a basic logic approach.

Nevertheless, 2D based human body orientation estimation approaches mainly rely on the visual features to estimate the orientation. The lack of geometry information decreases the ability of disambiguating.

2.2.2 3D based Human Body Orientation Estimation

Due to the limitations of 2D based human body orientation estimation approaches, for example the lack of geometric information, the use of 3D information is considered in some works.

Some of them reply on the information of silhouette shape, [133, 134] perform an analysis of the shape of the silhouette in images from ceiling mounted cameras, however, they only give a coarse estimation and very sensitive to arm movement. In the work of [135], body orientation coefficient vectors are extracted by performing multilinear analysis using binary silhouette images obtained from multiple cameras. The body orientation estimation is casted into a non-linear least square problem by using a one-dimensional manifold, which is constructed from the body orientation vectors. However, a disadvantage of this approach is low speed. Similarly, in the work of [136], they also estimate the upper-body orientation by using silhouette information only, independent of the target's appearance. The silhouette information is encoded using shape context descriptors, then classify the extracted feature vector to obtain a hypothesis for the orientation on each camera view and then fuse the results with a Bayesian filter framework. In the work of [137], they employ a 3D body model consisting of three elliptic cylinders, instead of binary silhouette, to represent human body, allowing to introduce a spatial color layout to discriminate the tracked person from potential distractors. In their work, a dynamical model is presented that models the coupling between people orientation and motion direction. This work exploits a loose coupling at low speed, but they did not have an explicit observation model for body pose estimation, resulting in an similar problem when people is moving slowly.

In [138], a texture sampled from a cylinder surrounding the person is shifted along the rotational axis to find the best matched orientation and form an panoramic appearance map simultaneously. In [139], the overall human body orientation from overlapping cameras is estimated. They jointly estimate the orientation with a 3D shape and texture representation of a person's torso-head, under a rigidity assumption. The overall body orientation is estimated by minimizing the difference between a learned texture model in a canonical orientation and a texture sampled using the current 3D shape estimate.

Compared to 2D based human body orientation estimation approaches, the additional geometry information improves the estimation results in 3D based approaches. However, almost none of these 3D based approaches have combined all the advantages from the components of 3D shape, appearance and motion in an unified framework.

2.3 Multiple People Tracking

Tracking multiple people has been an active research topic in computer vision area. Its primary goal is to generate accurate trajectories and assign consistent identity labels to corresponding people at each frame. There exists extensive work in this domain, which mainly includes two categories.

One line of research is based on sequential techniques, that relies on the recursive update of tracks with the most recent detections, such as Kalman filtering [140–148] and particle filtering [149–158], such trackers are causal, they consider only information from previously processed frames. It is suitable for time-critical applications when the number of objects remains small since no clues from future are required. In the same spirit, [159] proposes a general framework to sample the data association hypothesis using a Markov Chain Monte Carlo (MCMC) approach, which forms a track based on both current and past observations. [160] introduces a probabilistic model to associate merged and splitted measurements using a MCMC-based particle filter. [161–164] further extend the framework in [159] to identify the best spatial and temporal association of regions with a Data-Driven MCMC sampling approach. However, due to the recursive nature of these methods, identity switches get to be much more frequent and are difficult to correct when the number of objects increases. In addition, relying on recursive tracking may result in irrecoverable errors when a person fails to be detected in a frame or when two detections made at different frames are incorrectly linked. Errors tend to propagate and multiply in the subsequent

frames. Eventually, such unbounded error propagation leads a tracker to fail and manual re-initialization is required. Furthermore, these sampling-based methods typically require careful tuning of several meta-parameters, reducing the generality of its application. And they are only able to look at small time window, since their state space grows exponentially with the number of frames.

The other line has formulated tracking as frame-by-frame association of detections, it is a widely used paradigm for multi-person tracking [165, 166]. Tracking-by-detection approaches have become increasingly popular, driven by the recent improvements in object detection performance. Such methods have demonstrated impressive results in addressing the various challenges [35, 38, 69, 91, 167–176], they first apply an object detector to generate target hypotheses in each frame, and then seek to transitively associate the detections in a coherent manner, and thus maintain their unique identities. The transitive linking is difficult in the face of (potentially numerous) false positives and miss detections. This is usually addressed by learning an affinity model between detection responses or tracklets in terms of their intrinsic properties (e.g., appearance, location, motion, size) [177]. These methods consider both past frames and future frames, and then perform a global optimization, mitigating the false positive and missing detections that occur in individual frame, thus it is more likely to give an improved results.

Due to the advances of tracking-by-detection approaches, we will follow this line in our work. To this end, people tracking-by-detection methodologies will be focused in the following. As above mentioned, such methods involve two parts: detection of people in individual frame, and data association between frame-by-frame detections. As the related work on people detection has been reviewed in previous section, we now focus on the detailed review of the work on data association.

Classic data association approaches such as the Hungarian algorithm [178], is used to find the best assignment of possible detection-tracker pairs, whose runtime that is cubic according to the number of targets. A simple nearest-neighbor approach [179] uses only the closest observation to any predicted state to perform the measurement update, it is commonly used for MTT systems because of its low computational cost. Global Nearest Neighbor (GNN) [180] is based on the idea of bipartite matching, which formulates the single-scan observation-to-track association as a two-dimensional assignment, choosing the one with the highest joint probability as final association for current scan among all possible assignments. Although it has low computational complexity, it suffers from severe drawbacks in dense and noisy environments. Other approaches, such as Joint Probabilistic Data Association Filters (JPDAFs) [181] and

Multi-Hypothesis Tracking (MHT) [182] jointly consider the data association from sensor measurements to multiple overlapping tracks. In particular, JPDAF combines all of the potential measurements into one weighted average, before associating it to the track, in a single update. While MHT calculates every possible update hypothesis with a track, formed by previous hypotheses associated to the target. Both methods are known to be quite complex, and require a careful implementation in terms of parameters. In particular, the latter can not avoid the drawback of an exponentially growing computational complexity, with the number of targets and measurements involved in the resolution situation. Moreover, a global optimal solution cannot be guaranteed in sub-exponential time although they attempt to model the joint trajectories of all objects.

Recent works show that global optimization approaches of using Dynamic and Linear Programming have appeared to be powerful alternatives. Berclaz et al. [38] studies an efficient approximate dynamic programming scheme over individual trajectories. Greedy strategies are utilized to combine trajectories and handle potential conflicts. This approach tends to mix trajectories when targets are densely located, as occlusions are not explicitly modeled because of separate optimization.

By contrast, Linear Programming seeks to optimize all trajectories simultaneously over the whole sequence. Jiang et al. [183] tackles multiple people tracking problem with the use of Integer Linear Programming, in which the problem is formulated as multi-path searching by explicitly modeling the track interaction and objects' mutual occlusion. The metric for inter-object interaction term is convex while the intra-object term quantifying object state continuity through sequence. This scheme explores a large search space efficiently and gives a near-global optimality, because of the specific structure of the formulation. However, its state-space only consists of observations, not able to interpolate trajectories smoothly in case of the false alarms. Moreover, it requires *a priori* knowledge of the number of targets, which severely limits its applicability in real tracking situations.

Similarly, Berclaz et al. [184] formulates multi-people tracking problem as a constrained flow optimization, resulting in a convex problem that can be solved by standard Linear Programming techniques. Their method does not need *a priori* knowledge of target numbers, and the model is far simpler. Nevertheless, they haven't incorporate appearance features into data association process, thus making it prone to ID-switches in complicated scenarios. While dynamic model is also discarded in this work.

Shitrit et al. [185] addresses the appearance limitation by exploiting the global appearance constraints, as the extension of [184], in which the total number of tracked person is partitioned into L groups, and a separate appearance is assigned to each group. It reduces the number of ID switches for overlapping tracks. However, the appearance templates are selected manually through bounding boxes corresponding to members of each group.

Andriyenko et al. [186] proposes a similar work for multi-target tracking with global optimization approach. Compared with [184], they sample the location space on a hexagonal lattice to achieve smoother trajectories. In contrast to purely discrete approaches for multi-target tracking [184,186], the work [187] presents individual trajectories in continuous space and uses cubic B-splines for that purpose. It performs data association using discrete optimization with label costs, and trajectory estimation is posed as a continuous fitting problem.

Some other methods, like Quadratic Boolean Programming (QBP) [170,188], min-cost flow [172], have also been tailored to simultaneously optimize all tracks in polynomial time, are in fact closely related to ILP. The works [170, 188] couple detection and estimation of trajectory hypotheses by QBP, where a redundant set of putative trajectories is pre-computed, and the optimization takes place at the trajectory level by pruning to an optimal subset, thus such approaches can only optimize over a limited time window. While Zhang et al. [172] defines data association as a maximum-a-posteriori (MAP) problem, and models trajectory hypotheses as disjoint flow paths in a cost-flow network. There is also some work linking the trajectories through a set of pre-computed tracklets [189], the global optimal solution can be found by linking the tracklets via max-flow computation.

Despite of intensive studies, robust and efficient tracking of multiple targets with complex interactions and significant mutual occlusions remains a problem. Meanwhile, the proposed different ways of handling the data association problem do not take advantage of any behavior cue, such as body orientation, which provides the direct evidence of what the person is going to do and where the person is facing at. In particular, body orientation can provide valuable insight into the dynamics in case of social interaction and mutual occlusion. Although some works couple the dynamic model into the affinity model [186], however, such dynamic models mostly suppose steady heading, steady velocity or steady acceleration. It is problematic when a person is static or in low speed, as the velocity becomes too noisy to provide reliable information.

2.4 Conclusion

In this chapter we have provided a detailed literature review in the fields of people detection, human body orientation estimation, multiple people tracking while focusing on the works which have had a significant impact on these fields or are closely replated to our works. We have discussed their advantages and potential deficiencies according to the problems that we aim to solve, therefore those keypoints which have not been considered are clearly indicated, which inspires us to develop new approaches with improved robustness.

The next chapter will introduce that how we design the system architecture, to uniformly integrate the modules on people detection, human body orientation estimation and people tracking. It will provide a detailed description on experimental platform, system framework, hardware setup, test datasets and evaluation framework.

Chapter 3

System Architecture and Experimental Set-up

This chapter comprehensively presents the system architecture and detailed experimental set-up. We start by giving out the entire framework of the tracking system, the experimental platform that used for implementation. Hardware setup used to capture the video sequences is described, the choice on datasets and detailed descriptions on corresponding attributes of each dataset is presented. Furthermore, we present the performance evaluation framework used for evaluation of the entire tracking system, which includes the process of annotating ground truth data, and the various evaluation metrics that are used for assessing the different aspects of detection and tracking performance.

3.1 System Architecture and Implementation Backgrounds

In this section, we will systematically describe the structure of proposed multiple people tracking system and some background on the implementation platform.

Fig. 3.1 illustrates the unified tracking framework on multi-view hierarchical grid-based people tracking-by-detection under global optimization. It mainly consists of:

1. Tracking pipeline: sensor input module for acquiring source images, hierarchical grid-based detector for people detection and 3D localization, hybrid human body orientation estimation based on the local association of measurements from our detector, global optimal data association based on association affinity model which uniformly integrates the output from the previous two modules, and the final output module, which is the

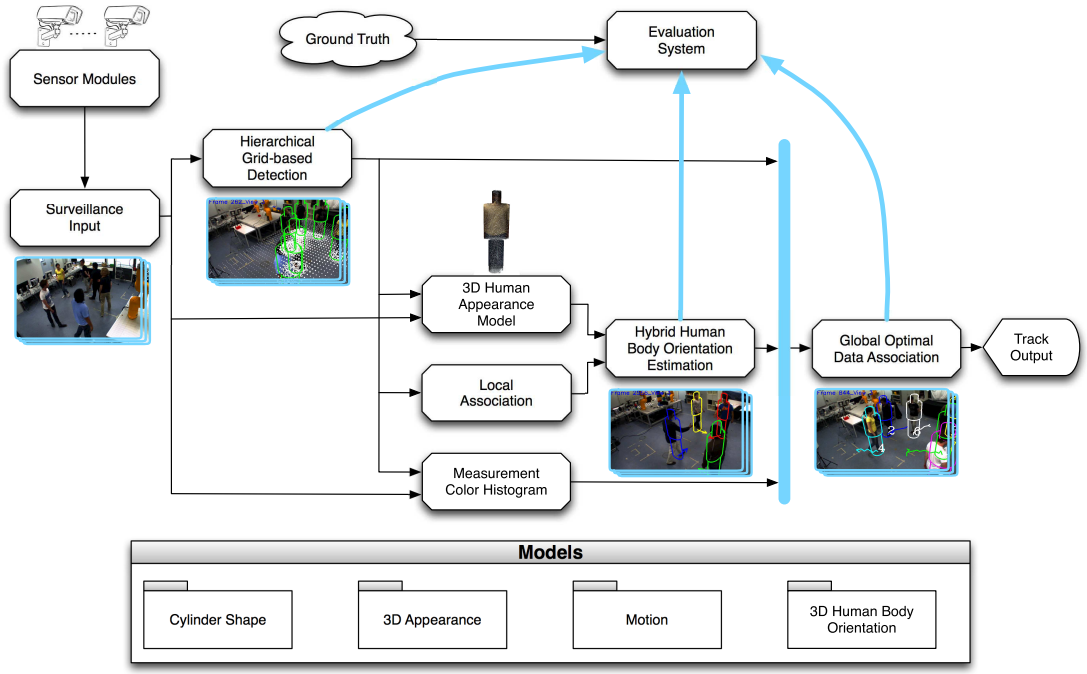


Figure 3.1: Framework of the proposed multiple people tracking system. It consists of tracking pipeline, evaluation framework and models that are used in the system.

determined people's trajectories associated with unique identity, in a world coordinate system.

2. Evaluation framework: we evaluate each subpart and the entire system through a performance evaluation framework, together with ground truth data and evaluation metrics. The evaluation is performed by comparing the results with the ground truth data through diverse aspects, such as accuracy, precision, robustness and so forth.

3. Models: They are usually prior information about targets, which consist of cylinder shape, 3D appearance model, motion and 3D human body orientation.

The software structure of our tracking system is designed and implemented partly on *OpenTL* [1] [190, 191], which is a structured, general purpose software library for model-based and marker-less visual tracking. It provides a user-friendly high-level application programming interface (API) for the widest variety of computer vision algorithms and applications. This library integrates multiple, heterogeneous visual modalities (edges, color, texture, motion, etc) in

[1] `http://www.opentl.org`

a seamless way, and handles information from multiple sensors, simultaneous targets, different object models [190]. Our global optimal multi-people tracking-by-detection system is a great extension, both in functionalities and applications, of the *OpenTL* library. The system is implemented with a hiearchical, object-oriented architecture, the main implemented functionalities can be briefly summarized as below,

1. Visual pre-processing module:
    - A new background subtraction algorithm based on the edge modality;
    - An automatic template hierarchy generation strategy, the levels of hierarchy and grid size could be defined by users;
    - A novel fast oriented distance transform algorithm, the DT map is then generated on background-subtracted edge maps;
2. Measurement processing module:
    - Data association by global nearest neighbor in feature level;
    - Global likelihood computation by fusing from multiple camera views;
    - Likelihood clustering on the finest grid level;
3. Detector Module:
    - Hierarchical grid-based detection with multiple cameras;
4. Tracker Module:
    - Affinity association model construction;
    - Tracking by detection approach with global optimal data association;
5. Visual modalities module:
    - Intensity edges for the hierarchical grid-based detection;
    - Hierarchical cylinder templates for matching with intensity edges;
    - Color histograms for the affinity model in global optimal data association;
6. Scene models module:
    - Cylinder shape consists of three cylinders, correspondingly represents the head, body and leg of a person;

- 3D appearance model constructed for 3D appearance-based human body orientation estimation;
- Dynamic motion model for motion-based human body orientation estimation;
- Edge-based background model, which is learned for edge-based background subtraction in hierarchical grid-based detector;

7. Data storage and GPU-computing module:

- Storage of generated template hierarchy for both position and orientation of all the points;
- Visibility test for planar reprojection of 3D appearance model;

8. Output module:

- Output visualization by cylinder model rendering for people detection, body orientation estimation and people tracking;
- Performance evaluation with ground truth data and a variety of metrics.

3.2 Test Datasets and Hardware Setup

In order to thoroughly validate our system, we collect a variety of datasets, including both indoor and outdoor scenarios. For the indoor sequences, we record in our own laboratory, while the outdoor ones are the publicly available datasets.

To be straightforward, all the video sequences used in our work, with its corresponding attributes and challenges, are briefly summarized in Table 3.1. For each sequence, a short description about the main challenges is provided in upper row. And the lower row indicates corresponding main attributes of each sequence, including the number of used cameras, image resolution, maximum number of targets involved, if some targets walk across others, if targets interact with each other, if occlusions happen, if targets walk with rotational motion, if targets run in the scene, if the illumination conditions change, if there are casted shadow. Note that, "Y" indicates the respective case exists in the scenario, whereas "N" means not. Every test sequence is chosen with regard to constraints and challenges that are slightly different, so that to test each module of our system with different emphasis.

We will give further details on the acquisition of the test sequences in our laboratory. The sequences are recorded from four, synchronized *uEye* USB cameras, with a resolution of 752 ×

Table 3.1: Corresponding attributes and challenges of each dataset.

Dataset	N_{Cam}	$Size_{Image}$	$N_{Targets}$	*Cross*	*Inter*	*Occlu*	$Motion_{rotate}$	*Run*	Ill_{change}	*Shadow*
Lab 2 Targets	The targets closely interact with each other by shaking hands for very long time.									
	4	752×480	2	N	Y	Y	N	N	N	N
Lab 3 Targets - Boxmotion	The targets walk with notable speed in most case, however, sometimes they walk with rotational motion, and they are mutually occluded.									
	4	752×480	3	Y	Y	Y	Y	N	N	N
Lab 3 Targets	Closely interacting with each other, and being occluded, frequently walking across others, and pose changes significantly.									
	4	752×480	3	Y	Y	Y	N	N	N	N
Lab 4 Targets - Interaction	Very long term interaction between targets, and with very slow motion even standing still.									
	4	752×480	4	N	Y	Y	N	N	N	N
Lab 4 Targets - Crossing	Most of the targets wear very dark clothing, with ambiguous appearance compared to each other.									
	4	752×480	4	Y	N	Y	N	N	N	N
Lab 6 Targets	Highly crowded in a small observing area, high density and heavy occlusion, large variability of clothing.									
	4	752×480	6	Y	Y	Y	N	N	N	N
EPFL - Campus	Featured with typical outdoor challenges, including illumination changes, shadows, reflections and moving objects in the background.									
	3	360×288	3	N	N	Y	N	Y	Y	Y
EPFL - Terrace	Much more challenging and crowded outdoor scenario, including large number of occlusions, interactions, significant scale changes, very long and sharp shadows, reflections, extreme close proximity.									
	4	360×288	9	Y	Y	Y	Y	N	Y	Y
PETS - S2L1	Additional challenges are posed due to monocular view, frequent grouping and splitting, far from camera view, dynamic motion, frequent occlusion by others or traffic sign.									
	1	720×576	8	Y	Y	Y	N	N	Y	N

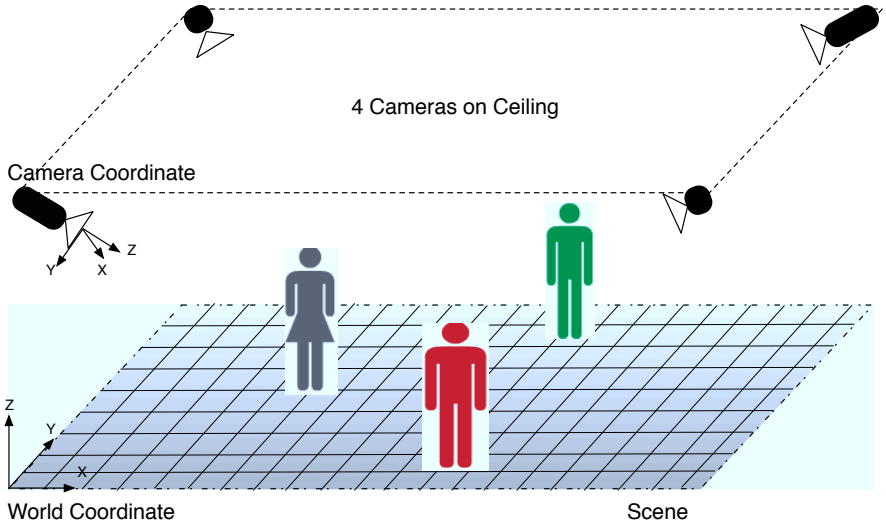


Figure 3.2: Laboratory camera setup. The four uEye usb cameras are mounted overhead on the corners of the ceiling, each of them observing the same 3D scene synchronously.

480 and a frame rate of 25 fps, as depicted in Fig. 3.2. The four cameras are mounted overhead on the corners of the ceiling, each of them observing the same 3D scene synchronously. Furthermore, all the cameras are connected to one multi-core PC. The videos are recorded in an unconstrained environment, and present significant variability of image quality. Note that the system is scalable with respect to the number of cameras, no modifications is required for the entire framework if adding a new camera.

A necessary step before being able to get accurate 3D information, is calibration of the intrinsic and extrinsic camera parameters, that we perform with the Matlab Calibration Toolbox [1], with respect to a *world* coordinate system placed on the floor. A chessboard pattern is used to acquire coplanar points, and perform the full calibration.

Within our laboratory datasets, from 2 up to 6 people are involved in the scenarios. In order to imitate the real world scenarios better, a wide diversity of poses are included in our dataset, they might shake hands, hug, or talk with each other for long time, and also might stand still, being occluded, or closely interacting with others, and so on. The observing area is about $4m \times 6m$, with the increasing number of targets, the scenario is getting much crowded and occlusions between targets from one or more views are getting more serious. In all of the sequences, multiple people freely enter and leave the observation area, as well as closely

[1] `http://www.vision.caltech.edu/bouguetj/calib_doc/`

interact with each other for long periods, with significant mutual occlusions and even with extreme similar appearance.

For outdoor datasets, we choose several typical sequences which have representative outdoor attributes. The first two sequences are from the public "EPFL" dataset [116, 192]: Sequences *campus* is recorded outside in a campus from 3 different camera viewpoints, with a resolution of (360×288) and a frame rate of 25 fps, consists of around 5800 frames. Up to four people are simultaneously walking randomly. Sequence *terrace* is recorded outside on a terrace for around 3 1/2 minutes, up to 9 people involved in front of 4 DV cameras, with a very high density therefore featuring a quite challenging outdoor scenario. EPFL terrace sequence consists of 5010 frames, covering an area of $7.5m \times 12m$. The last sequence is from S2.L1 subset of PET2009 [193], it is publicly available and well documented since it is used as part of the PETS tracking challenge. There are 795 frames showing up to 8 people moving in a street, with a resolution of 768×576 at 7 fps. The PETS2009 S2L1 dataset is around 2 minutes, using 7 cameras covering an area of approximately $100m \times 30m$, in particular, four DV cameras are places about 2 meters above the ground, while three video surveillance cameras are placed around 3 to 5 meters high, and significantly far from the scene. Due to the elevated viewpoint of this sequence, it is more suited for monocular tracking. The outdoor scenarios pose much more challenges, including daylight changing, dynamic background, large variability of clothing, fast motion of people, a large number of occlusions, interactions, significant scale changes as well as extreme close proximity, and so on. Note that calibration data is available for all the public datasets.

3.3 Evaluation Framework

In order to systematically evaluate the performance of entire multiple people tracking system, it is instructive to have a quantitative performance evaluation framework. By using the framework, the evaluation is usually performed by comparing the results with the ground truth data which is normally manually annotated by person. And the tracking results can be evaluated through diverse aspects, such as accuracy, precision, robustness and so forth.

3.3.1 Groundtruth Annotation

Obtaining the groundtruth of a video sequence is often a difficult, monotonous and a time wasting process, however it is a crucial step during quantitative evaluation. Most conventional ground truth annotation methods are based on generating 2D bounding boxes around the person

Table 3.2: Details on all annotated sequences.

Dataset	Number of Frames	Labelled Interval
Lab 2 Targets	1700	Every frame
Lab 3 Targets - Boxmotion	3000	Every frame
Lab 3 Targets	576	Every frame
Lab 4 Targets - Interaction	3160	Every frame
Lab 4 Targets - Crossing	1800	Every frame
Lab 6 Targets	1520	Every frame
EPFL - Campus	5800	Every 10 frames
EPFL - Terrace	5000	Every 10 frames
PETS - S2L1	792	Every 5 frames

within the images, however the main disadvantage of image plane based evaluation approach is that they provide very limited information on the real accuracy of the model, as distances and overlapping areas measured in pixels depend on the distance of the target from the viewpoint as well as extrinsic and intrinsic camera parameters.

In contrast, we measure and evaluate the 3D ground positions of the persons by locating the 3D geometrical model onto a predefined grid, as will be introduced in Chapter 4.3.1, and rendering the model onto all camera views according to calibration parameters so that to coincide with the target area within image. Thus we can provide both 3D people location on the ground and their corresponding bounding boxes in camera views, which makes our annotation technique be able to compare different methods against each other, no matter they estimate the 3D ground position of people or 2D position within image plane. Additionally, by annotating in 3D space with multiple camera views, the number of missed persons and inaccurate ground truth position data caused by occlusions are reduced significantly.

In our ground truth annotation, all targets are manually labeled in all sequences, even in case of total occlusion. Each target will acquire a new unique ID when it enters into the observing area. Note that if a target leaves the area and reenters again, then it would be assigned a new ID. Overall, the datasets that we have labelled are presented in Table 3.2. For most of the sequences, the ground truth is labelled every frame, and for some very long ones, it is labelled every 5/10 frames and each target's trajetory between labeled keyframes is then interpolated.

Table 3.3: Evaluation metrics used throughout the thesis.

Name	Definition
MODA	Multiple Object Detection Accuracy: It assess the accuracy of a detection system. The higher the better.
MODP	Multiple Object Detection Precision: The spatial overlap between a detection and its corresponding ground truth. The higher the better.
Recall	Correctly matched objects / total ground truth objects.
Precision	Correctly matched objects / total output objects.
MOTA	Multiple Object Tracking Accuracy: It combines all error types and is normalized with the total number of targets. The higher the better.
MOTP	Multiple Object Tracking Precision: The normalized distance between the objects and tracker hypotheses. The higher the better.
FN	False Negatives: Number of objects that are mis-tracked by the tracking algorithm. The smaller the better.
FP	False Positives: Number of objects that are tracked by the system which does not have a matching ground truth. The smaller the better.
IDS	ID Switches: Number of times that a tracked trajectory switches its matched ground truth identity. The smaller the better.
GT	Groundtruth trajectories: Number of groundtruth trajectories.
MT	Mostly tracked: Percentage of ground truth trajectories that are tracked for more than 80% in length. The higher the better.
ML	Mostly lost: Percentage of ground truth trajectories that are tracked for less than 20% in length. The smaller the better.
FM	Fragments: Number of times that a ground truth trajectory is interrupted. The smaller the better.

3.3.2 Evaluation Metrics

The problem of evaluating tracking system has been addressed by the computer vision community [194]. The consensus is that there is no single metric that could indicate sufficiently the performance of entire system. Therefore for a complete evaluation, it is essential to use diverse metrics quantifying different aspects of the system performance. A proper set of evaluation metrics would be able to allow optimizing the parameters of algorithms, quantitatively comparing different algorithms, supporting improvement of the algorithm.

To this end, we use two sets of metrics: detection metrics and tracking metrics. For part of them, we choose the standard CLEAR MOT metrics [195], since they have been widely employed in the literature as the main judge of performance of detection and tracking methods, and can intuitively express a detector and tracker's overall strengths. To more thoroughly evaluate the performance, we further use the metrics on precision and recall, additionally, we apply part of metrics which are presented in the work of [173].

Table 3.3 gives a summary and short description of all the metrics used throughout this thesis. We give here a short overview of the metrics, for a detailed description and motivations of the metrics, please refer to [195–197].

3.3.2.1 Detection Metrics

The detection performance is measured by four metrics: Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP), Precision, and Recall.

MODA accounts for all possible errors such as miss detection and false alarms. MODP measures the relative accuracy of alignment between ground truth and the predicted bounding box on image plane. Precision and Recall are based on the number of correct and false matches between detection results and groundtruth.

With the set of detected person at a given timestep denoted as $D = \{D_1^{(t)}, D_2^{(t)}, \cdots, D_m^{(t)}\}$, while the annotated ground truth data as $G = \{G_1^{(t)}, G_2^{(t)}, \cdots, G_n^{(t)}\}$, we apply the Hungarian algorithm to find the maximum match between detections and ground truth data, in order to determine the number of missed detections, false alarms as well as the accuracy of position errors.

In particular, MODA is used to assess the accuracy aspect of the detection system performance. Within this metric, the number of missed detections m_t and false alarms fp_t is utilized. Let g_t be the number of ground truth objects at frame t, then the MODA score can be computed as:

$$\text{MODA} = 1 - \frac{\Sigma_t(c_m(m_t) + c_f(fp_t))}{\Sigma_t g_t} \ , \tag{3.1}$$

where c_m and c_f are the cost functions for the missed detections and false positives. We use them as scalar weights here, varying up to the focus of the application. For instance, if missed detections are more critical than false positives, then we can increase c_m and reduce c_f. In our case, both are set to 1.

For metric MODP, the spatial overlap information between a detection and its corresponding ground truth is used to compute the Overlap Ratio. As mentioned before, $G_i^{(t)}$ denoted the annotated i^{th} ground truth object at frame t, $D_i^{(t)}$ denotes the corresponding detected object for $G_i^{(t)}$, thus we have

$$\text{Overlap Ratio} = \sum_{i=1}^{N_{mapped}^t} \frac{\left|G_i^{(t)} \cap D_i^{(t)}\right|}{\left|G_i^{(t)} \cup D_i^{(t)}\right|} \ , \tag{3.2}$$

where N_{mapped}^t is the number of mapped object pairs in frame t. Note that, the Overlap Ratio is summed over all camera views and normalized in our multiple view scenario.

Using the assignment sets, the final MODP that gives the detection precision for the entire sequence is computed as:

$$\text{MODP} = \frac{\Sigma_{t=1}^{N_{frames}} \frac{\text{Overlap Ratio}}{N_{mapped}^t}}{N_{frames}} \ . \tag{3.3}$$

And for the metric - Precision, it is defined as

$$\text{Precision} = TP/(TP + FP) \ , \tag{3.4}$$

where TP, FP are the numbers of True Positive and False Positive. And

$$\text{Recall} = TP/(TP + FN) \ , \tag{3.5}$$

where FN is the number of False Negative. Intuitively, precision value gives information on the amount of false alarms, and recall value tells how many of the objects have been detected. Note that for all metrics, the larger their values, the better the performance.

3.3.2.2 Tracking Metrics

The tracking performance is measured by six metrics: Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), Groundtruth Trajectories (GT), Mostly Tracked (MT), Mostly Lost (ML) and Track Fragments (FM).

In particular, MOTA accounts for three types of errors: false negatives (FN), false positives (FP), identity switches (IDS) over all frames, giving an intuitive measure of a tracker's performance at detecting objects and keeping their trajectories, independently on the precision with what the object locations are estimated.

$$MOTA = 1 - \frac{\Sigma_t(FN_t + FP_t + IDS_t)}{\Sigma_t g_t} \ , \tag{3.6}$$

where g_t is the actual number of targets at time t.

Conversely, MOTP evaluates the average distance between estimated and true target locations, demonstrating the capability of a tracker on estimating precise object positions, regardless of its skills at detecting objects, keeping consistent trajectories and so on.

$$MOTP = \frac{\Sigma_{i,t} d_t^i}{\Sigma_t c_t} \ , \tag{3.7}$$

where d_t^i is the localization error between object o_i and its corresponding hypothesis, and c_t the number of matches for time t.

In addition, MT and ML are computed on whole trajectories and measure what's the percentage of ground-truth trajectories that are successfully tracked, respectively, for at least 80% of their lifespan or for less than 20%. For FM, it is defined as the total number of times that a true trajectory is broken during tracking.

3.4 Conclusion

This chapter systematically provides an insight into the multiple people tracking system in terms of system framework, implementation platform, hardware setup, test dataset. A proper and thorough evaluation of the tracking system is an important issue which seems to be neglected in many existing works. We address this problem by annotating all the test sequences and utilizing a set of metrics, to evaluate the detection and tracking performance of the entire system. This evaluation framework can not only be used for people tracking applications, but also could be used for more general object trackers.

We now will go to details on each module of the tracking system in the following three chapters, including how to accurately detect and localize targets in cluttered scene; how to efficiently estimate the human body orientation with 3D based approaches, by combining all the advantages from the components of 3D information, appearance cue and motion information into a unified framework; and finally how to address the people tracking problem by efficiently utilizing the output from detection and orientation estimation.

Chapter 4

Hierarchical Grid-based People Detection and 3D Localization

4.1 Introduction

Robustly detecting people in real world scenes is a fundamental and challenging task, which has attracted an extensive amount of interest from the computer vision community over the past decades. Prominent challenges for this task are the high intra-class variability, high degree of articulation, and huge variation in physical appearance, shape, movement. Many techniques have been proposed in terms of features, models, and general architectures. Although some are successful towards the challenging task of detecting moving targets in both indoor and outdoor environments, in particular in scenes which are with relatively few people. It still severely remains difficult to detect multiple people precisely in the presence of cluttered environment, illumination changes and partial or fully occlusions.

In this chapter, we focus on tackling the issue of reliably detect people in real-world environment and localize them in 3D space, so that to provide reliable input for the component of people tracking afterwards. The problem is addressed by employing a hierarchical grid-based detection methodology, to detect and localize people in 3D space while providing a robust 3D output towards frequent mutual occlusion between interacting people, cluttered and dynamically changing background and illumination changes of environment.

To firstly segmenting out objects of interest, background subtraction as a commonly used technique has achieved a significant success in fixed camera scenarios. Most of methods work by comparing color or intensities of pixels in the incoming video frame to the reference image [198–

202]. However, it has the drawback of being susceptible to illumination changes, and provides a less precise localization. In contrast, we propose here an edge-based background subtraction, which employs Canny [203] edge map together with Sobel gradients [204,205], because edges are more precisely and stably localized, to a better extent in presence of illumination changes. To match objects, among the various visual cues shape has the advantage that providing a powerful object discrimination capability, which is relatively stable to illumination conditions. Template-based shape matching approaches yield nice results for locating targets with no prior knowledge in a cluttered scene. The efficiency of this method is illustrated by using about $4,500$ templates to match pedestrians in images [18]. The core idea is using a Chamfer distance measure, so that matching a template with the DT image results in a similarity measure. However, if only computing the location of edge pixels without considering their orientation when computing distance transform, it inevitably leads to a high rate of false alarms in presence of clutter. Thus, we propose here a new methodology on oriented distance transform, which matches not only the location of edge points but also their orientation, significantly reducing false alarms in cluttered environment. Additionally, we propose a hierarchical likelihood grids scheme, taking the advantage of multi-resolution grids that can, on one hand precisely and efficiently locate targets in cluttered scenes, while on the other hand providing a powerful speedup of three orders of magnitude compared to exhaustive searching.

Our detector has two important properties that making it particularly suitable for detecting multiple people in crowded scenes: firstly, it allows to detect people in the presence of variations in physical appearance, shape and environmental lighting condition, and secondly, it is capable of locating targets in cluttered scenes precisely and efficiently without prior knowledge of their position.

The remainder of this chapter is organized as follows: Chapter 4.2 describes the general overview for algorithmic flow. In Chapter 4.3, we provide the detailed detection procedure, including models, edge-based background subtraction, hierarchical grid evaluation as well as model-based contour matching and state-space filtering. The experimental results are discussed in Chapter 4.4. Finally, Chapter 4.5 summarizes the chapter.

4.2 Overview of the Approach

This section presents an overview diagram of our approach, as illustrated in Fig. 4.1, that consists of two main processing modules: offline and online procedures.

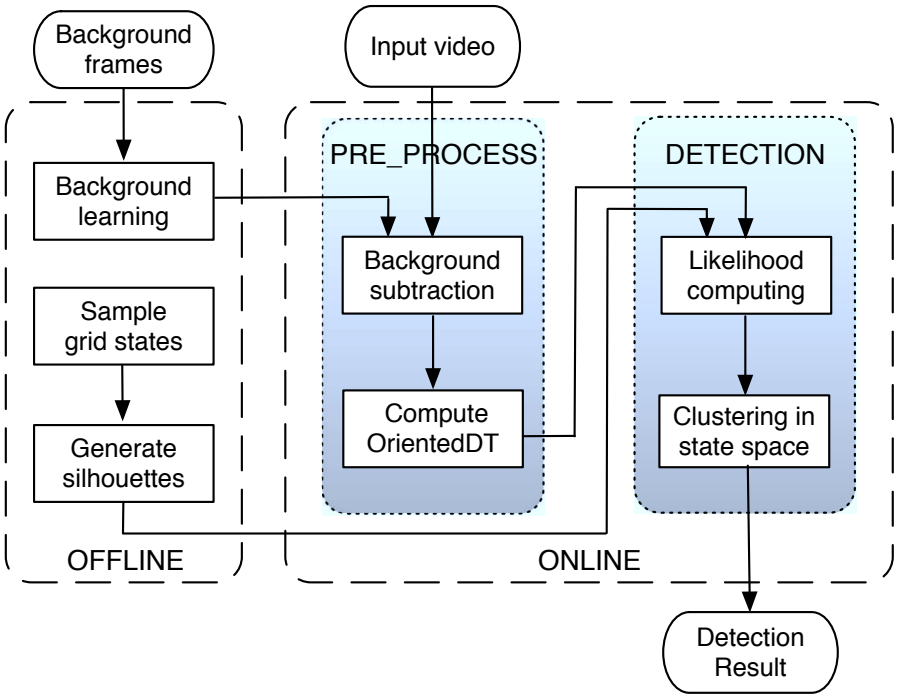


Figure 4.1: Schematic diagram of our hierarchical grid-based detection approach. It mainly contains two processing module: offline and online. The background learning and silhouette generation is done offline, and the online module includes the pre-process part and hierarchical detection.

In the offline phase, edge detection is performed on a certain number of pictures of the environment (without people) and used to learn a background map containing, for each camera view, the average distribution of edge pixels (position and orientation). Meanwhile, the 3D grid of states is sampled at each level, and silhouette templates are generated by projecting a simple geometry model, computing the external boundary, and collecting model edges pixel-wise (also storing position and orientation).

In the online phase, we have two submodules: pre-process and detection. Within the pre-process part, for each camera view foreground contours are segmented by edge-based background subtraction, using the learned model. Afterwards, we compute an oriented distance transform onto this image, in order to match, for each template, both the location and the orientation of its contours. In particular, the oriented DT is efficiently computed over a finite set of orientations, so that the image is sampled over parallel scan lines that are pre-computed. The advantage of using both edge position and orientation, during background subtraction as well as template matching, is a strong reduction of false alarms.

Detection part firstly computes the likelihoods by matching projected templates and oriented DT for each camera view, where the likelihoods are computed on the coarse grid firstly, then refined on the next resolution only the locations where the likelihood is higher than a given threshold, the joint likelihoods can simply be multiplied then. The object-level measurements, or target hypotheses, are obtained by means of likelihood grid clustering, performed by Gaussian filtering of the high-resolution grid and local maxima detection.

4.3 Hierachical Grid-based People Detection

In this section, we provide full details about people detection based on shape templates and hierarchical grids, that serves as one of our key building blocks for our entire system. In short, the detection is performed by means of hierarchical likelihood grids, by matching shape templates through an oriented distance transform over foreground intensity edges, followed by clustering in pose-space.

4.3.1 Construction of Template Hierarchy

The idea to construct a template hierarchy is inspired by the paper [44], as well as by the system developed by [18], here extended to multiple views, multiple targets, and with a more general template.

A graphical illustration is shown in Fig. 4.2. Assuming there are L levels of search, the state space is partitioned with a coarse-to-fine strategy. Each discrete region $\left\{R^{i,l}\right\}_{i=1}^{N_l}$, where

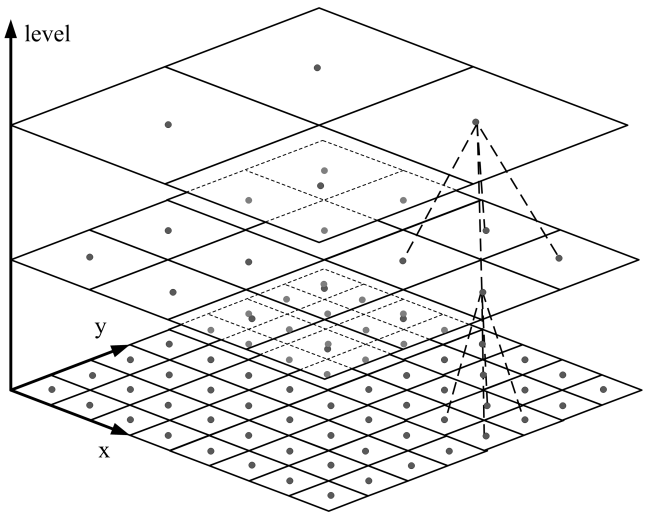


Figure 4.2: Grid based state space with hierarchical partition through a coarse-to-fine art. The grid size of each child level is doubled as the previous parent level. All the regions at a child level are connected to its parent cell.

N_l is the number of cells at level l, is sampled at its center, before the template hierarchy is generated. Meanwhile, we connect regions at a child level with its parent cell, by computing the nearest-neighbor in state-space, as well as its nearest neighbors within the same level, as it will be described in Chapter 4.3.6, in order to smooth the grid likelihoods.

After sampling the grid, templates are generated by rendering the 3D model at each state, under the respective camera projection. To more precisely match our target, the model chosen here is composed of 3 cylinders, where one cylinder is for the head, one for the torso, and one for the legs. The model undergoes (x,y) translation on the floor, while its silhouette is generated by projecting the external contour. An example of the model and a partial view of the hierarchy of silhouettes are shown in Fig. 4.3.

For each silhouette, the position of each point as well as its normal is collected, as it will be described further in Subsection 4.3.3. As already emphasized, both grid sampling and template hierarchy generation are performed offline.

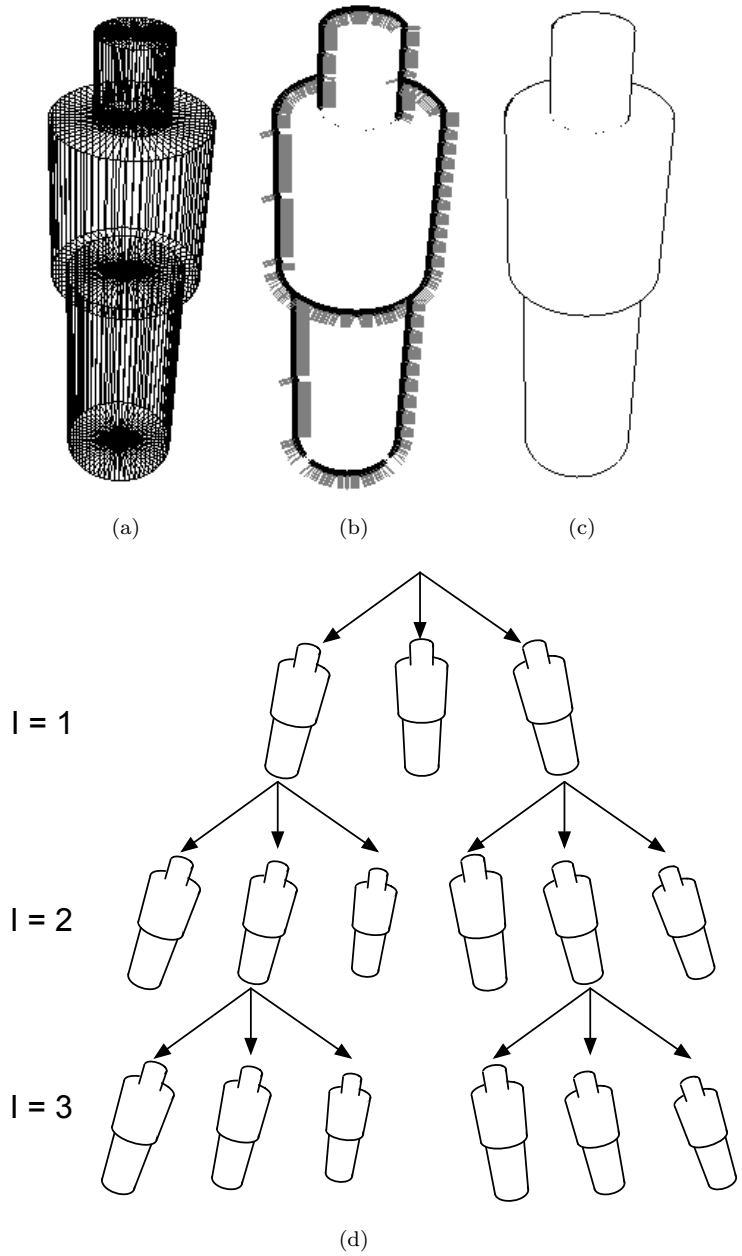


Figure 4.3: Illustration of our shape model. (a) Discretized cylinder. (b) Silhouette with normals. (c) Silhouette without normals. (d) Hierarchy of the silhouettes.

4.3.2 Edge-based Background Subtraction

In order to match the templates against image data, we propose here an edge-based background subtraction approach, that can be divided into three phases: background learning (offline), foreground segmentation and post-processing (online). In the first phase, we utilize a certain number T of frames without people to learn the background model. Let $\Theta_b(t), G_{bx}(t), G_{by}(t)$ respectively be the Canny edge map and the Sobel (x,y)-gradient images, detected at frame $I_b(t)$. The edge map Θ_b is accumulated from frame 1 to T as a simple binary OR between $\Theta_b^{(I)}(1), \ldots, \Theta_b^{(I)}(T)$, while Sobel gradients are accumulated as a running average. At the end, we also normalize the average gradient, so that $G_{bx}^2 + G_{by}^2 = 1,\ \forall\,(x, y)$ wherever the gradient is not zero.

Subsequently, a standard distance transform is applied to the accumulated edge map, and thresholded to a few pixels in order to provide a binary mask $\Theta_{DT} \in \{0, 1\}$, where potential background edges can be detected in the incoming images.

Online, edge map and image gradients $\Theta_f(t)$, $G_{fx}(t)$, $G_{fy}(t)$ are computed at frame $I_f(t)$, and the position and orientation of each edge pixel is tested: any pixel with $\Theta_f(t) \neq 0$ lying close to a background edge ($\Theta_{DT} \neq 0$) is a candidate for removal. These edges are further tested for orientation using image gradients, and if the scalar product is higher than a given threshold Th_s

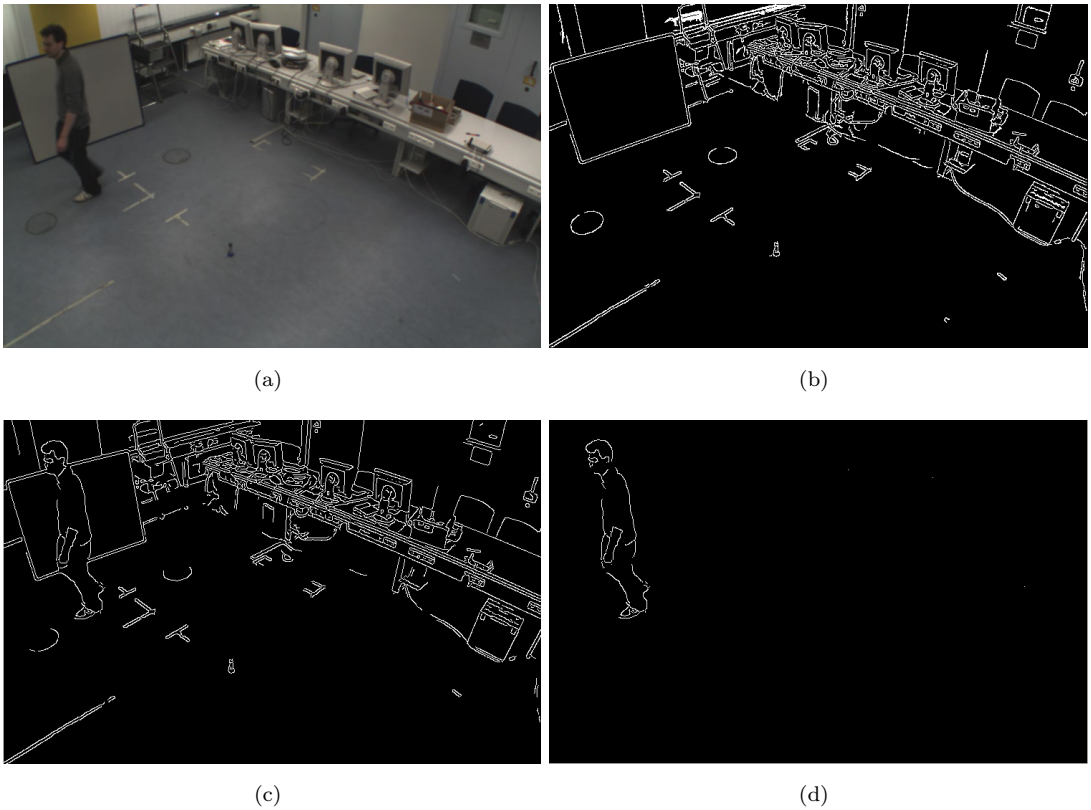
$$\frac{G_{bx}G_{fx} + G_{by}G_{fy}}{\sqrt{G_{fx}^2 + G_{fy}^2}} > Th_s, \tag{4.1}$$

the point is removed from $\Theta_f(t)$.

Fig. 4.4 shows an example of this procedure: as it can be seen, the resulting edge map robustly preserves the person contours while discarding most of the background edges, despite the relatively cluttered scenario.

However, we notice that some small gaps may result in the foreground edges. Therefore, we perform some morphological postprocessing, in order to close gaps by analyzing the singular edge points (referred to as endpoints), without appreciably increasing the overall computing time.

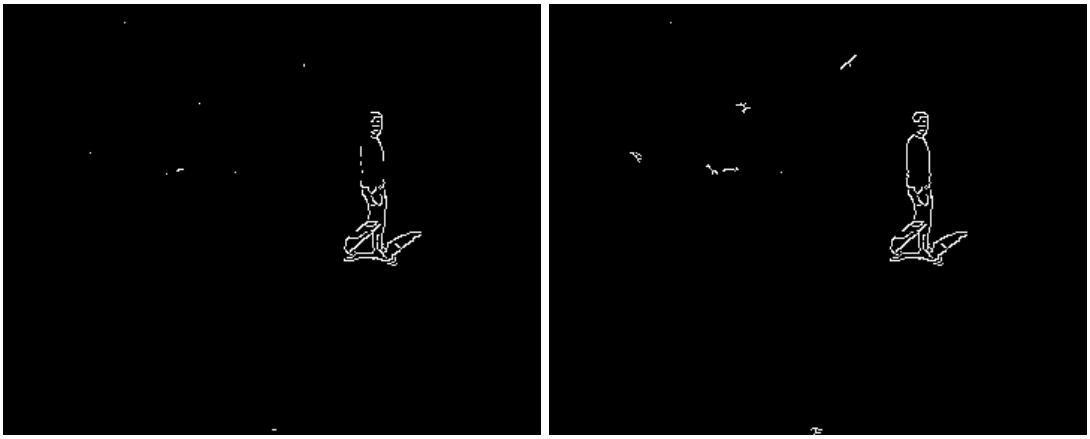
A straightforward edge linking approach consists of detecting endpoints of contours in the new edge map, and linking them through the neighborhood of endpoints in the original edge map. Following this heuristics, we firstly detect the endpoints in $\Theta_f(t)$ as a set $\left\{x_{i_{ep}}, y_{i_{ep}}\right\}_{i_{ep}=1}^{N_{ep}}$. Then, we check the 8-neighborhood of an endpoint within the unsegmented foreground map $\Theta_{f_u}(t)$ (Fig. 4.4c), and eventually extend the segmented contour by adding back a previously



(a) (b)

(c) (d)

Figure 4.4: Edge-based background subtraction. (a) Original frame, in which the background is quite cluttered. (b) Learned background model. (c) Unsegmented foreground edge. (d) Segmented foreground edge.



(a) (b)

Figure 4.5: Illustration of the postprocessing result. (a) Edge map before postprocessing. (b) Edge map after postprocessing, the gaps are closed mostly within the unconnected edges.

detected neighbor pixel. Afterwards, the endpoint is replaced with the new one and the procedure is repeated until the number of updated endpoints falls below a threshold Th_{ep}.

Fig. 4.5 illustrates the effect of morphological postprocessing. This scheme proved to be efficient enough to close gaps between closely spaced, unconnected edges, at the price of slightly increasing the size of (erroneously detected) background edges.

4.3.3 Oriented Distance Transform

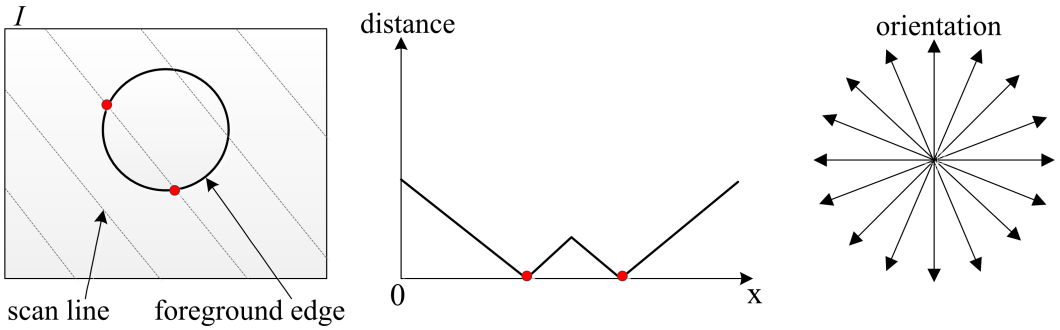
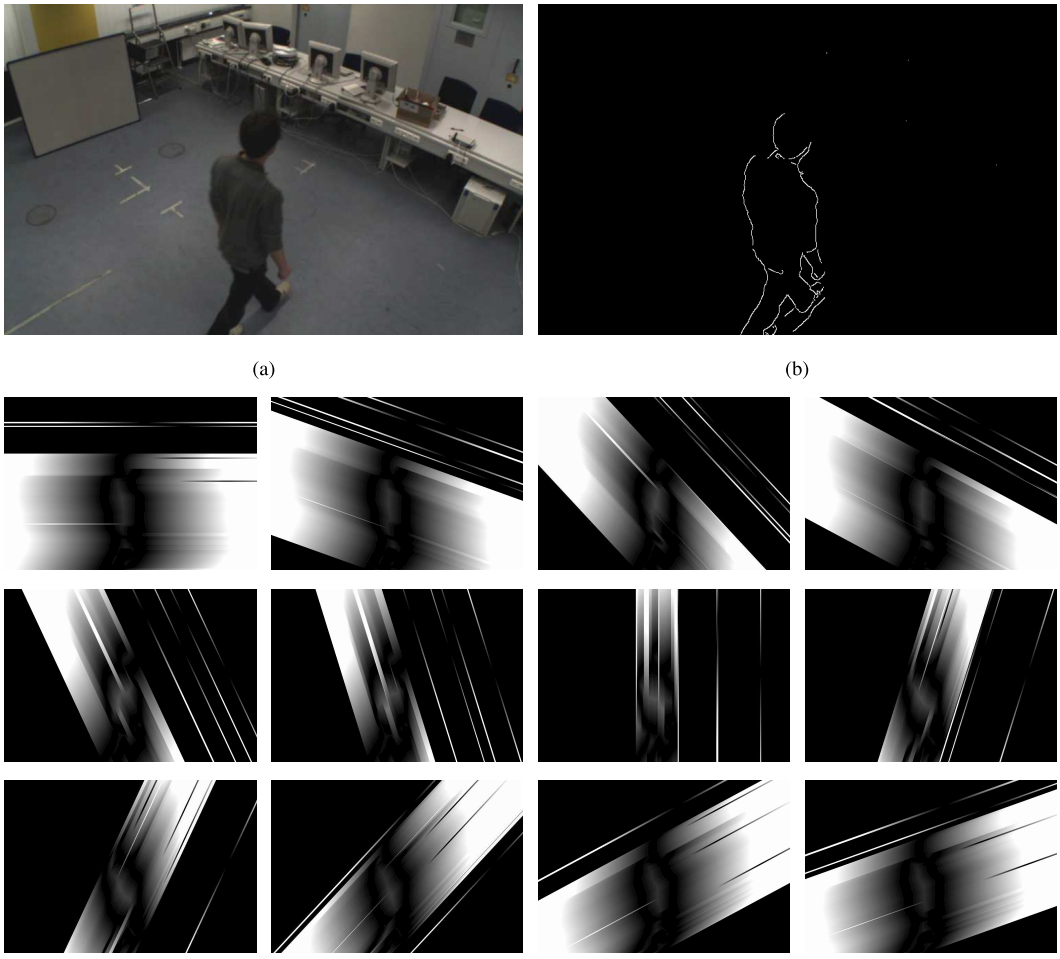


Figure 4.6: Scanning single line for one direction. From left to right: Multiple single line scanning; Distance value to the nearest edge point on the line; Multiple scanning directions.

The next step is to match foreground edges with the model silhouettes. One possibility would be to apply the standard Chamfer distance to the edge map, that is tolerant to small shape



(a) (b)

(c)

Figure 4.7: Results of oriented distance transform. (a) Input image. (b) Foreground edge map. (c) Oriented DT (at 12 discrete orientations).

variations and has already been applied in several works, such as [18, 206]. However, despite the background edge subtraction, in case of images with significant clutter a high rate of false alarms could be present. This problem can be reduced by matching not only the location of edge points, but also their orientation [207]. For this purpose, we consider an *oriented* distance transform.

However, integration of edge orientation and position can be defined in many ways. In particular, we choose the most conservative approach of searching for corresponding edges only along the *template normals*, (that is chosen among a discrete set of angles $\gamma \in \Gamma$), leading to several DT images that are obtained by scanning the edge map through all parallel raster lines

having the orientation γ.

We build the oriented DT efficiently by scanning the edge image along parallel lines $\mathcal{L}_\gamma(a)$ through pixel $a = (x, y)$ for a given orientation γ, and repeat it for a finite set of N_γ directions $\Gamma = \{\gamma_i\}_{i=1}^{N_\gamma}$. The algorithm is illustrated in Fig. 4.6.

In particular, for each direction and each scan line, the oriented DT is a mono-dimensional function, looking for the nearest edge point b on either direction

$$b = DT_\gamma(a) = \min_{b \in \mathcal{L}_\gamma(a)} \|a - b\| \tag{4.2}$$

An example of oriented distance transform is shown in Fig. 4.7.

4.3.4 Optimized Fast Search Strategy

The proposed oriented DT method needs two scans for each raster line: one for finding edge pixels on the line, and the other for writing the DT values in the output image. In particular, all of the image pixels on each line must be read, before deciding whether any edge pixel is present, and then assign them DT values. However, if the line crosses no edge, no one of these pixels will have a valid DT value. Moreover, even for a valid scan line, most pixels have a DT which is higher than the validation gate, and therefore have no valid DT as well, but the line iterator can only proceed one pixel at a time, therefore wasting computational resources.

We propose here an optimized fast search strategy on oriented DT, which instead performs line scans starting directly from the *edge pixels*, and proceeding in the two opposite directions, until a maximum distance that corresponds to a pre-defined *validation gate* $D_{\max}$.

Thus, each pixel in the DT image holds a counter, telling how many pixels away, along the given direction, the closest edge is found[1]. Notice that, unlike the standard DT, this image is *sparse*, because many points do not find a corresponding edge along the desired direction, or the closest edge may be beyond the validation gate. Those points are therefore considered outliers, and their DT is set to ∞.

Going into more detail, scanning is obtained by maintaining a double-linked, circular list of *exploring units*, two for each foreground edge pixel (which are in a limited amount, after background subtraction), that keep trace of the current DT value and perform a single pixel move in each direction, executed one after the other through the circular list.

[1] To be precise, one pixel move corresponds to a distance tha depends on the orientation, e.g. along the diagonal it would be $\sqrt{2}$. However, we simply multiply each DT by this constant factor when computing the likelihood.

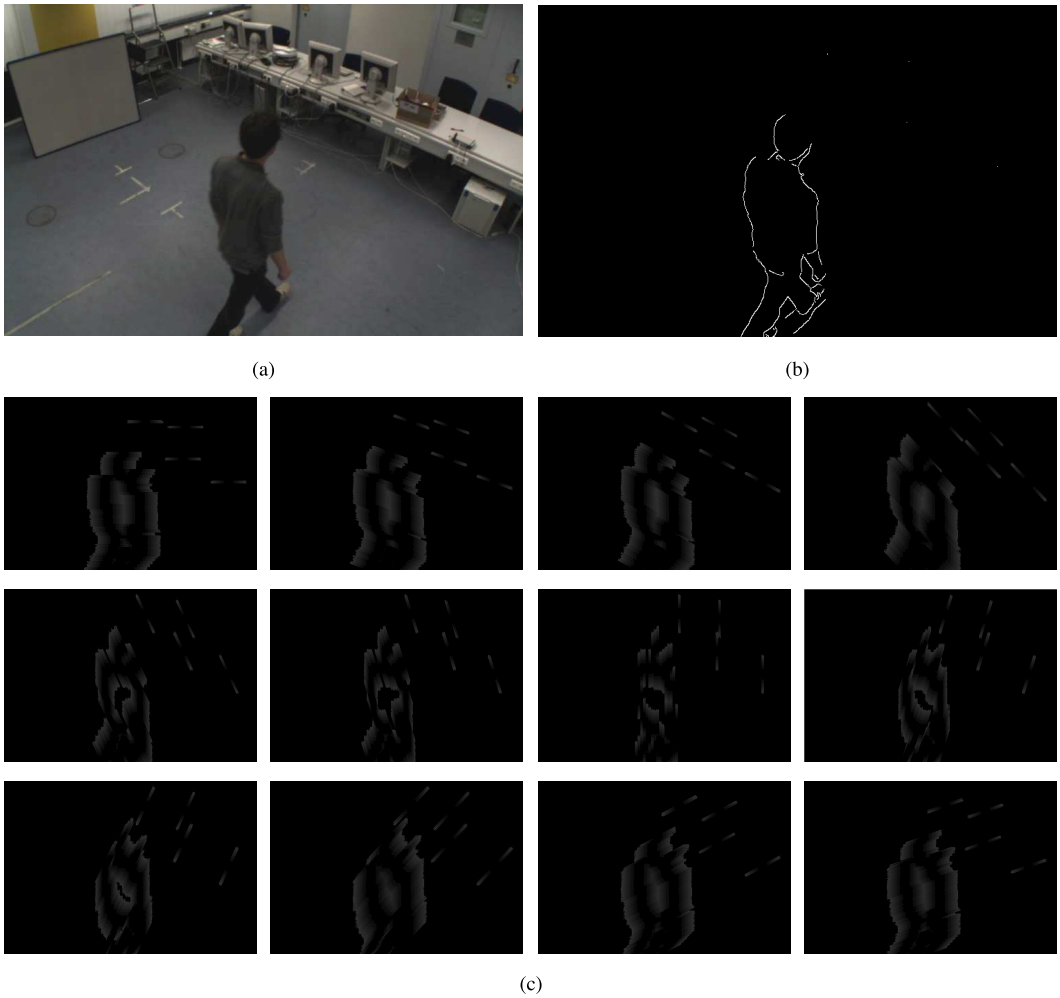
A single iteration consists of: one read operation (to check the current pixel DT), one write operation, and one move (increment of pixel position and DT value). The read operation ensures that, if a pixel has been already visited (i.e. its DT value is not empty), the unit is stopped and removed from the list. By performing one move per unit each time, we make sure that two units coming from two different edges but traveling along the same scan line in opposite directions, will meet exactly at the mid-point, so the DT values will be correctly assigned and their search will be stopped.

When the list is empty, the algorithm terminates. Overall, this strategy reduces the number of operations (read/write/iterate) to a minimum, since only valid pixels are visited, while keeping an exact computation of distance. A pseudo-code of the fast oriented DT is shown in Algorithm 1, while Fig. 4.7 shows an example of results on the optimized fast oriented DT.

Algorithm 1 Fast oriented distance transform

Initialization :
Fill the DT image with ∞, *apart from* 0 *at the foreground edges.*
Create a double − linked, circular list of "exploration units", two for each foreground edge pixel, going in opposite directions (= *line iterators*).
Each unit consist of :
- *A distance counter*(*initialized with* 0);
- *A line iterator*(*initialized with edge pixel position*), *with a given direction*;

Main loop :
while *list is not empty* **do**
  Take current element of the list;
  Read DT value at (x, y);
  if $0 < DT(x, y) < \infty$ **then**
    Remove unit from the list;
  else
    Write the counter value into the DT image;
    Increment counter;
    if *counter* > *validation gate* **then**
      Remove unit from the list;
    else
      Increment line iterator;
    end if
  end if
  Move to the next unit in the list;
end while



(a) (b)

(c)

Figure 4.8: Results of fast oriented distance transform. (a) Original image. (b) Foreground edge map. (c) Fast oriented DT results (at 12 discrete orientations).

4.3.5 Hierarchical Template based Matching

Once DT images are available, template matching can be simply amounted to compute the likelihood, by summing up all values over the silhouette pixels, using the closest orientation to its normal. More formally, a projected template s is represented by a set of N pixel positions and normal orientation angles $\{x_i, y_i, g_i\}_{i=1}^{N}$, and the orientation is used to select the nearest $\gamma \in \Gamma$, say $\gamma(g_i)$ (up to a 180 degrees ambiguity, since the direction of the normal does not matter), from which the DT value will be taken. Therefore, the likelihood for state hypothesis s is given by:

$$P(z|s)_{\text{pos}} \propto \exp\left(-\frac{1}{2NR^2}\sum_{i=1}^{N}\min\left(DT_{\gamma(g_i)}(x_i,y_i)^2, D_{max}^2\right)\right), \tag{4.3}$$

where $\gamma(g_i)$ denotes the closest available direction to the normal, and the sum is performed over all values $\{x_i, y_i, g_i\}_{i=1}^{N}$. R is the measurement standard deviation, and an outlier threshold is usually fixed at $D_{max} = 3R$, which is our validation gate for a more robust matching. We also notice that, in order to avoid problems with different scales, the sum is further normalized by N.

During the computation of likelihood, a coarse-to-fine search strategy is applied by evaluating it, at each level, only for locations where the parent cell likelihood is higher than a given threshold, which is usually obtained as the average likelihood [44]. For those cells where the parent likelihood is under the threshold, its value is simply inherited, thus saving a large amount of computation.

4.3.6 Likelihood Grid Clustering and 3D Localization

In order to obtain the object-level measurements, or target hypotheses, after likelihood computation we employ a isotropic Gaussian filtering procedure on the high-resolution grid, where each cluster is a local maximum, potentially corresponding to a person.

This approach is similar to mean-shift [208], but explicitly done on discrete states. First of all, a Gaussian filtering is applied to the grid, where the isotropic Gaussian corresponds to the filtering kernel. For each cell s_i within the grid, we take the nearest neighbor s_j by looking at the connected states with distance $d_{i,j} = \|s_i - s_j\|$ up to a validation gate $D_{max} = 3\sigma_s^2$, where σ_s is the measurement covariance in *state-space*, these neighbors are pre-computed in the off-line phase. For each neighbor, the Gaussian weight is also pre-computed by

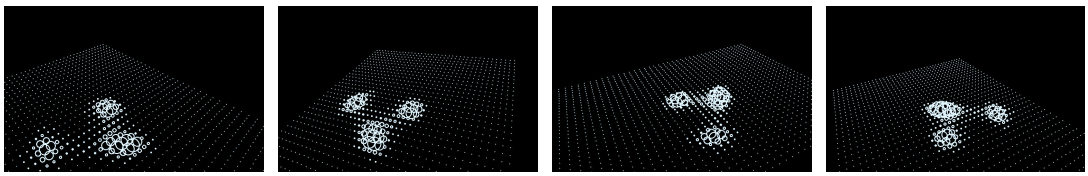
$$W_{i,j} \propto exp(-\frac{d_{i,j}^2}{2\sigma_s^2}), \tag{4.4}$$

the computed weights are also normalized to 1, so that the smoothed likelihood for state cell s_i is given by

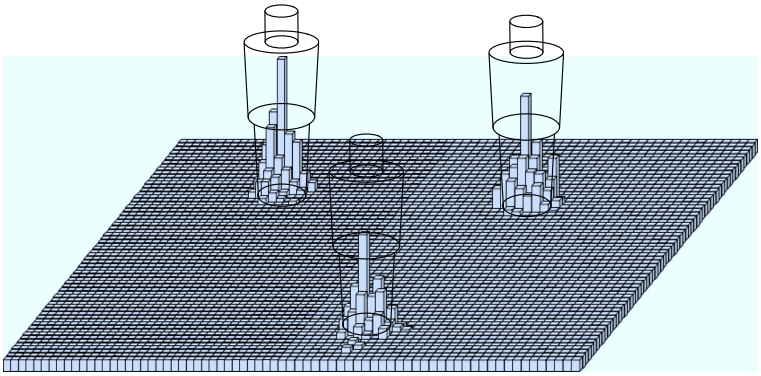
$$P(z|s)_{\text{smooth}}(i) = \sum_{i,j} W_{i,j} \cdot P(z|s)_{\text{pos}}(j). \tag{4.5}$$

Subsequently, local maxima are detected (within the same neighborhood), to obtain the hypotheses, or measurements for targets.

Fig. 4.9 intuitively illustrates the clustering and localization process. In Fig. 4.9 (a), the clustered likelihood grid maps are projected each camera view. The larger circle means higher likelihood. And in Fig. 4.9 (b), the likelihood map is visualized in 3D space, the local maxima can be easily detected so that to obtain potential targets' 3D location.



a) Clustered likelihood grid map projected on each camera view



b) Potential targets are localized in 3D space by detecting the local maxima

Figure 4.9: An illustration of likelihood clustering and 3D localization of local maxima.

4.4 Experimental Results

In this section, we aim to demonstrate both qualitatively and quantitatively the performance of hierarchical grid-based detector, through various video sequences from real world. Firstly, the important implementation details are presented. Secondly, the experiments with both qualitative and quantitative evaluation are comprehensively performed. Thirdly, we study the influence of several parameters on the detector's accuracy.

4.4.1 Implementation Details

Here we provide some important details of the practical implementation. Firstly, in our experiments the state grids are discretized respectively as $N_g \times N_g$, $2N_g \times 2N_g$ and $4N_g \times 4N_g$ from the coarsest to the finest, resulting in a total of $21N_g^2$ grid cells. The dimension of the observing area and grid is varied depend on different scenarios, as shown in Table 4.1.

Table 4.1: Dimension of the observing area and grid.

Scenario	Area	Finest grid size	Total cells
Indoor lab	$4m \times 6m$	40×40	2,100
Campus	$13m \times 12m$	60×60	4,725
Terrace	$7.5m \times 12m$	60×60	4,725
PETS	$20m \times 18m$	80×80	8,400

During edge-based background subtraction, the number of frames used for background learning is varied up to different video sequences, the threshold θ mentioned in Eq. (4.1) is consistently set to 0.9.

Our implementation of the oriented distance transform uses 12 discrete orientations, ranging from 0 to π, with the validation gate of 50 pixels. As it computes each orientation separately, the optimized fast oriented DT requires about 0.12 sec/frame for four images, whereas the original oriented DT is computed in 0.25 sec/frame and a single standard distance transform is computed in 0.1 sec/frame. Therefore, the proposed optimized fast search strategy is effective and competitive.

The cylinder model size in our experiments are set to $175cm$, in which the head size is $30cm$, upper body size is $60cm$ and leg size is $85cm$, those size are chosen according to the average people height in current societies.

4.4.2 Detection Performance

Figs. 4.10 to 4.15 showcase the hierarchical grid-based detector on various video sequences. Within each sample frame, the likelihood maps on the finest level which intuitively illustrate the likelihood for each grid, have been superimposed onto the detection results. As it can be seen from the frames, the challenges are gradually increased. In the following we will discuss more details on the detection performance.

Indoor Scenario

Within our indoor laboratory sequences, there are respectively 3, 4, 6 targets involved. With the increasing number of targets, the scenario is getting much crowded and occlusions between targets from one or more views are getting more serious. For instance, in Fig. 4.11, each two targets are occluded from some views in frame 480, however, since for the same pairs there are no occlusions from another camera view, all targets are successfully detected, and the similar case happens more frequently when 6 targets are involved, as can be seen from Fig. 4.12, almost

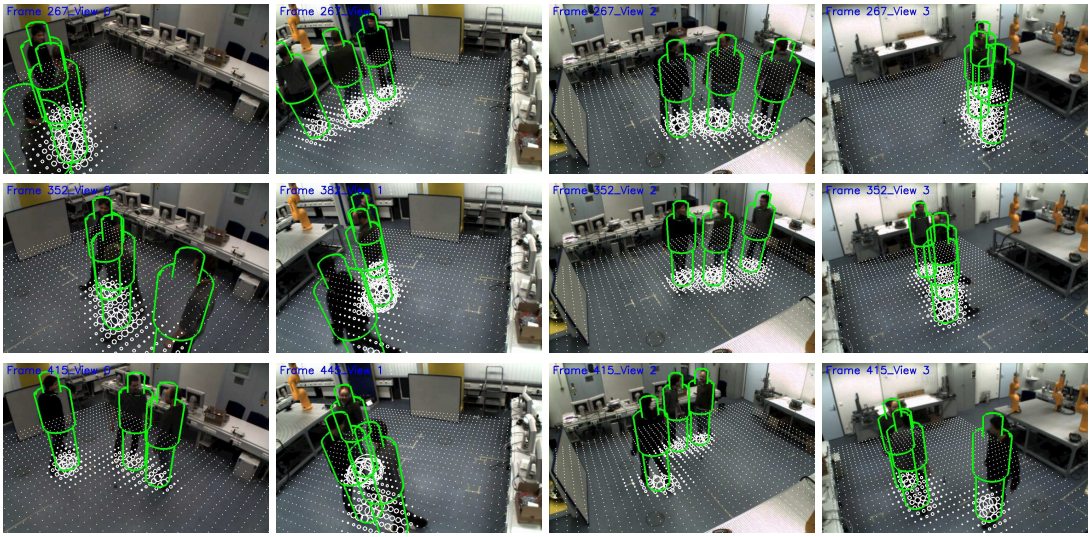


Figure 4.10: Detection results on the sequence of Laboratory 3 Targets. Every row shows a diffent frame, while every column displays diffent camera view. The likelihood grid maps of the finest level are superimposed onto the detection results, illustrating the likelihood for each detected person in corresponding frame.

in each sample frame, the mutual occlusions between targets are existing. In spite of this, the targets can be correctly located, thanks to the robustness of multi-camera fusion and oriented DT matching.

Outdoor Scenario

Now we turn to more difficult outdoor scenarios, with much more challenges involved, including daylight changing, dynamic background, large variability of clothing, fast motion of people, mutual occlusions or occluded by obstacles, extreme close proximity, and so forth. For example, within Fig. 4.13, we can see that in frame 1706, both targets are running with significant speed, and our algorithm can locate them quite well. Fig. 4.14 illustrates a scenario with plenty of challenges, we can notice that the scenario is extremely crowded, due to the reason that there is up to 9 targets involved, thus leading to a significant number of occlusions and interactions. In addition, the daylight conditions are varying during the process of detection. One more difficulty is that the clothing of the people are almost all dark and extremely similar except the white one. Under these challenges, the performance of our detector is satisfying, this can be largely attributed to the edge-based background subtraction, that overcoming the disadvantage from appearance-based approaches. A more challenging outdoor scenario is shown in Fig. 4.15, here we use only one view out of the sequence PETS2009 - S2L1, from the VS-PETS

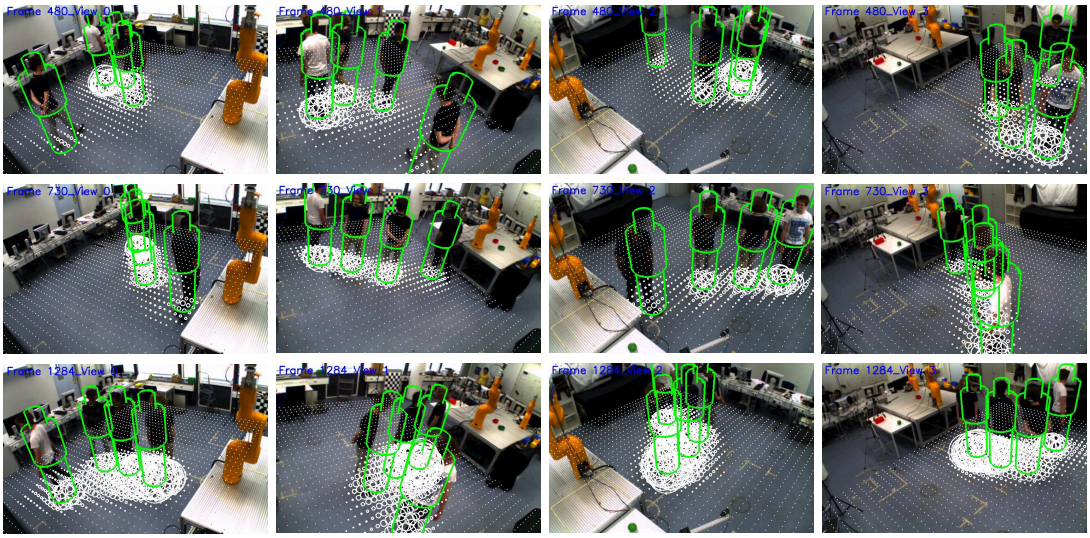


Figure 4.11: Detection results on the sequence of Laboratory 4 Targets. Every row shows a diffent frame, while every column displays diffent camera view. The likelihood grid maps of the finest level are superimposed onto the detection results, illustrating the likelihood for each detected person in corresponding frame.

2009 benchmark dataset. Due to the monocular view, much more additional challenges appear, for example, it is very hard to distinguish targets when they are partly or fully occluded by each other or by obstacles, as in frame 145 and 477. And as the targets are very far from the camera viewpoint, if the targets are getting too close with each other, then it would be difficult for the detector to correctly locate, such as in frame 145 and 262. However, in most cases, our detector can handle quite well. Note that, in frame 525, 614, 662 there are people that not being detected, this is because that they are not in the observing area, which is marked by white dot.

4.4.3 Quantitative Evaluation

In order to better evaluate the performance of our proposed hierarchical grid-based detector, it is instructive to have a quantitative performance evaluation with groundtruth data, both in terms of position accuracy and robustness of detection. For the details on groundtruth annotation and evaluation metrics, please refer to chaper 3.3.

Fig. 4.16 shows evaluation results on all the labelled sequences based on the previous introduced metrics MODA and MODP. Note that a detection is only counted as correct if the overlap ratio between the annotated box and the detection result is greater than a given thresh-

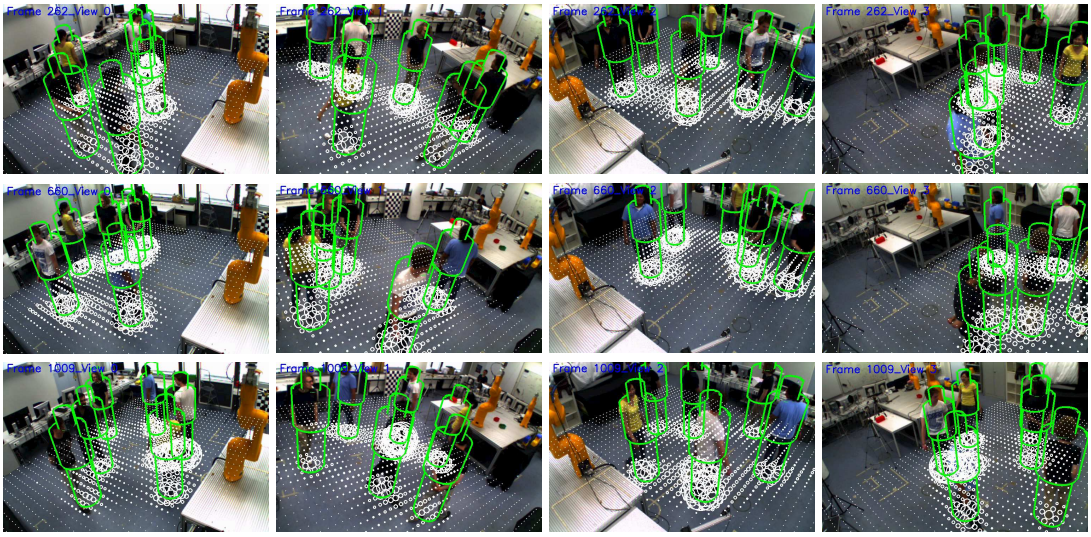


Figure 4.12: Detection results on the sequence of Laboratory 6 Targets. Every row shows a diffent frame, while every column displays different camera view. The likelihood grid maps of the finest level are superimposed onto the detection results, illustrating the likelihood for each detected person in corresponding frame.

old $\tau_{overlap}$. The results on MODA and MODP should be varying according to different given threshold, we further vary the threshold systematically and compute the evaluation metrics at each threshold during our experiments, as illustrated in Fig. 4.17. We can see from the results that, as the overlap threshold decreases, MODA increases and MODP decreases, because more misaligned detections become classified as correct ones. Nevertheless, our MODP curve has a slow decreasing. Fig. 4.18 shows results through a similar threshold varying but using precision/recall metrics.

In our evaluation results illustrated in Fig. 4.16, $\tau_{overlap}$ is set to 0.3. As illustrated by the evaluation results, the localization precision is generally high. Note that our hierarchical grid-based detector not only detect people in the camera views, but also accurately localize them on the 3D ground plance, which is a major advantage of our work compared to most existing methods. For the sequence PETS-S2L1, however, due to the monocular view, the performance of the detector turns to be less successful compared to other sequences. Overall these results indicate that, despite the cluttered situations, the localization results are basically satisfactory. One is because of the local edge-based matching which, despite the simplicity of the model, is more precise with respect to global statistics such as color histograms, or histograms of oriented gradients.

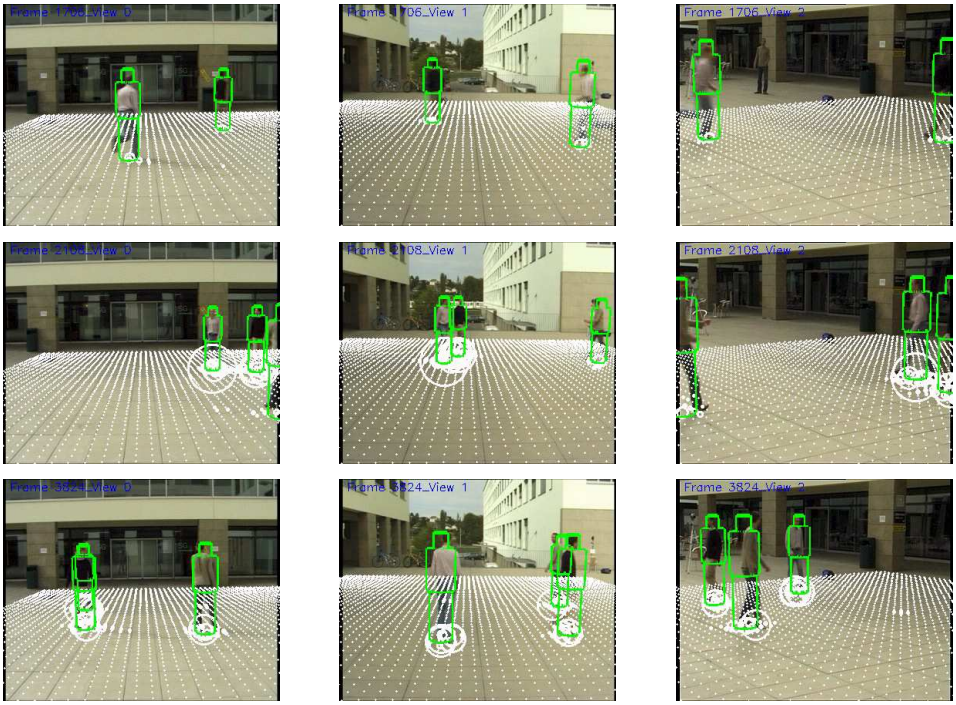


Figure 4.13: Detection results on the sequence of EPFL-Campus. Every row shows a diffent frame, while every column displays diffent camera view. The likelihood grid maps of the finest level are superimposed onto the detection results, illustrating the likelihood for each detected person in corresponding frame.

4.4.4 Discussion

We have demonstrated on a variety of sequences that the detector can robustly handles different challenges. The robustness allows the detector to work both indoor and outdoor without too much parameter tunings.

Note that, the grid resolution is one of the most important factors that influencing the performance of our hierarchical grid-based detector. If the grid resolution is fine enough, the real-world locations can be better explained by the discrete grids, so that for each sampled template silhouette, it is more possible for a target matching with it at the corresponding grid, thus leading to better performance for the detector. However, a finer grid resolution would lead to a higher computational cost. Conversely, if the resolution is coarse, then it is ambiguous to locate a target at the correct grid, especially in the case of crowded scene.

Another important factor that altering the quality of detector's output, is the output of the background subtraction. Current edge-based background subtraction can handle the case

Figure 4.14: Detection results on the sequence of EPFL-Terrace. Every row shows a diffent frame, while every column displays diffent camera view. The likelihood grid maps of the finest level are superimposed onto the detection results, illustrating the likelihood for each detected person in corresponding frame.

of natural illumination changes in both indoor and outdoors, however, the subtraction result could be largely affected if the illumination drastically changes.

The third factor is the proximity of involved targets within observing area. If there are too many people in the scene, then it is too ambiguous to matching templates with foreground DT maps, thus being not able to successfully detect and locate all the targets simultaneously. Currently it is hard to quantify the maximum number of targets that we can deal with, since it is largely up to the hardware configuration including the number of cameras, the viewing angle of the camera, and also the space between the targets.

4.4.5 Runtime Performance

The detection algorithm has been implemented in C++, and runs on a desktop PC with Intel Core 2 Duo CPU (1.86 GHz), 3GB RAM and an Nvidia GeForce 8600 GT graphic card. The computational cost can be affected by two aspects: one is the grid size, the computational complexity is linear with respect to the size of the grid. And the other is the image resolution, downsampling the input images can result in a significant speedup. The execution time is about

Figure 4.15: Detection results on the sequence of PETS-S2L1. Sample frames on single view are shown. The likelihood grid maps of the finest level are superimposed onto the detection results, illustrating the likelihood for each detected person in corresponding frame.

4 FPS for the indoor laboratory sequence with the grid size of 40×40 and image resolution of 752×480.

4.5 Conclusion

In this chapter, we have presented a novel approach for multiple people detection in various unconstrained environments, using a hierarchical grid-based methodology. A template hierarchy is constructed off-line, by partitioning the state space. And frame-by-frame detection is performed by means of hierarchical likelihood grids and clustered on the finest level. Moreover, edge-based background subtraction has been proposed for foreground segmentation, which is quite robust to illumination changes, together with an oriented distance transform, matching the silhouette templates by taking gradient orientations into account, thus significantly reducing the rate of false alarms. Experimental results over both indoor and outdoor video sequences show that our proposed approach deals fairly well with the challenges including complex interactions, mutual occlusions, cluttered environment, illumination changes, and so on.

In the remainder of our work, we rely on the hierarchical grid-based detector to provide the frame-by-frame detection for following parts, including hybrid human body orientation

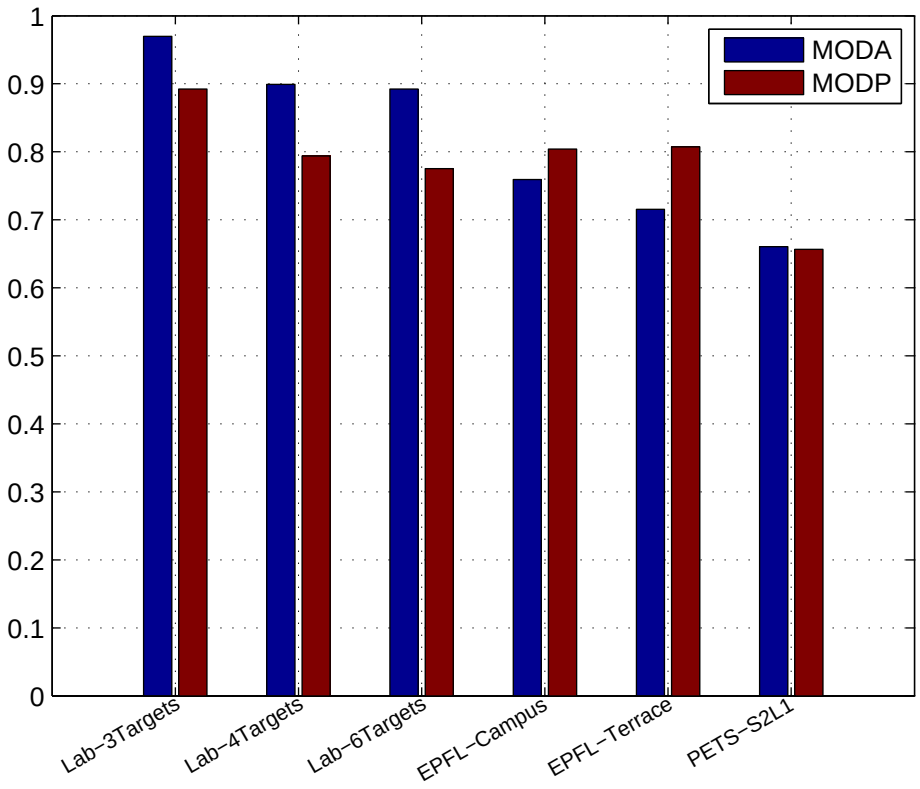


Figure 4.16: Evaluation results on the sequence of Laboratory-3Targets, Laboratory-4Targets, Laboratory-6Targets, EPFL-Campus, EPFL-Terrace, PETS-S2L1 respectively, using Multiple Object Detection Accuracy (MODA) and Multiple Object Detection Precision (MODP) metrics.

estimation and global optimal data association. The next chapter will introduce the novel approach of estimating the 3d body orientation of multiple human, which might interact with each other for long time, slowly move or even stand still. It will provide a detailed discussion of how we dynamically combine the merits of a motion-based and a 3D appearance model-based methods, so that combining the advantages from both mechanisms.

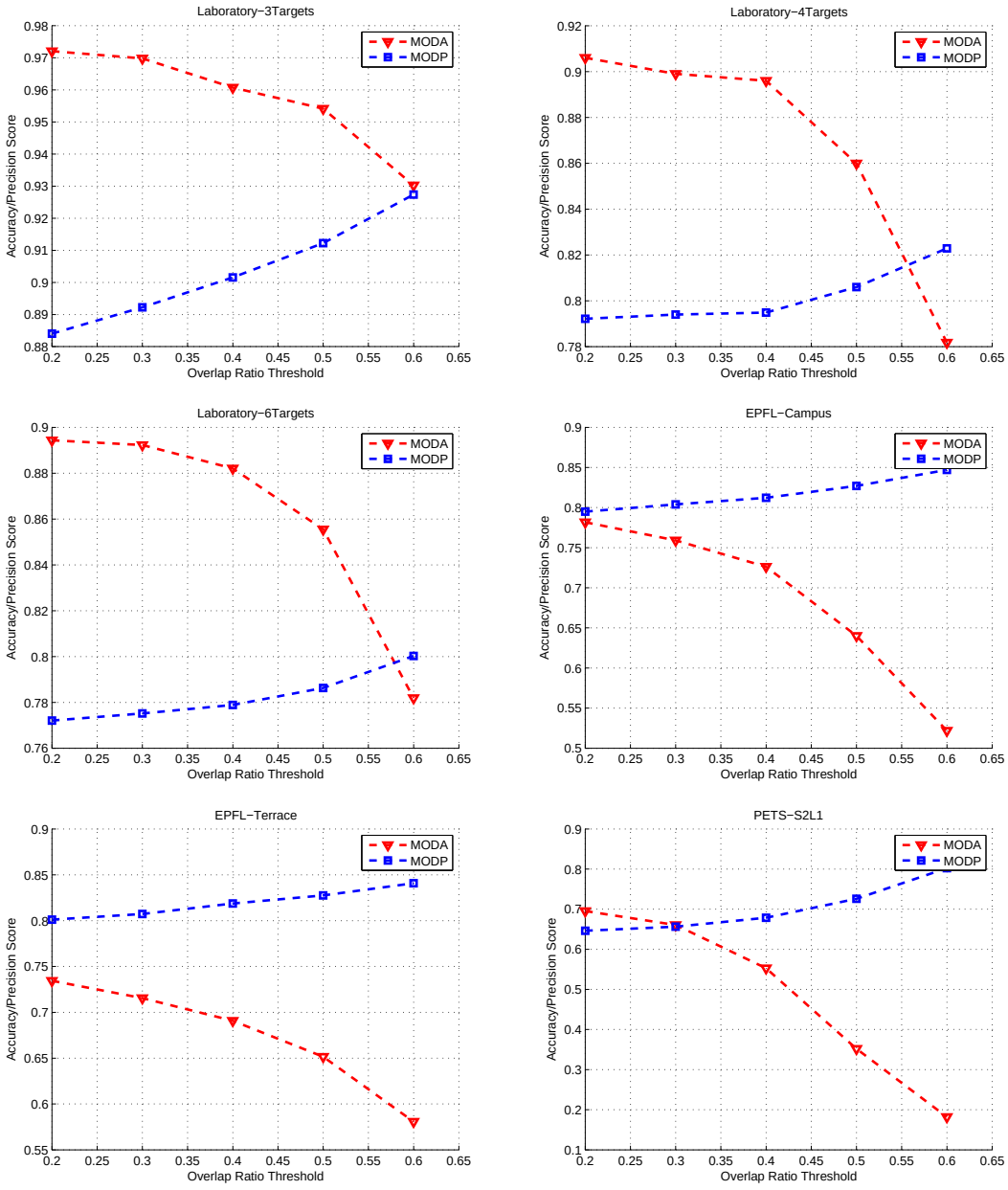


Figure 4.17: Influence of overlap threshold level on the evaluation results (MODA and MODP) for all of the sequences.

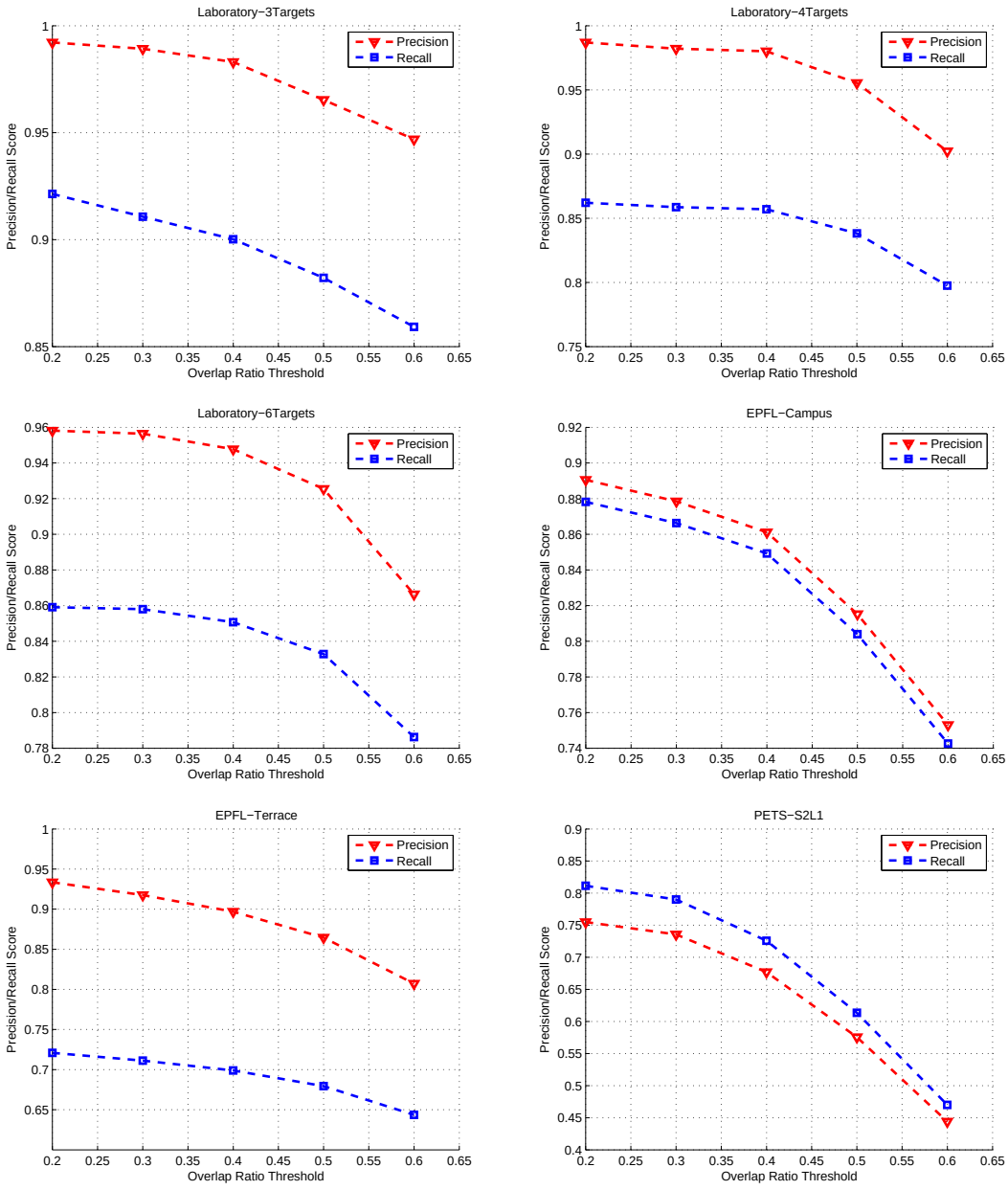


Figure 4.18: Influence of overlap threshold level on the evaluation results (Precision and Recall) for all of the sequences.

Chapter 5

Hybrid Human Body Orientation Estimation

5.1 Introduction

To better tracking people robustly and efficiently in case of complex interaction and significant mutual occlusions, there are some cues that are necessary and important for improved track disambiguation through performing a finer analysis of individual or group human behavior. Among those cues, a person's body orientation (i.e. rotation around 3D torso major axis) conveys much valuable information about the person's current activity, indicating the person's direction of focus, direction of movement or social interaction with other people within the entire scene, thus efficiently contributing to understand people's potential behavior, and therefore can be used to discover interaction between multiple people and disambiguating more complex scenarios with mutual occlusions. Fig. 5.1 gives an example of such ambiguous situations, with extremely close people occluding each other, furthermore with a similar appearance, and remaining static over a few frames. Here, body orientation provides a very discriminative feature.

The aim of this chapter is to robustly and accurately determine the body orientation of a variable number of people in 3D space, through multiple calibrated views mounted on the ceiling from different viewing angles. We propose a hybrid methodology, combining the merits of a motion-based and 3D appearance-based approach, being fast and capable of automatic initialization of 3D appearance models, and working both in case of moving, slowly moving or even still-standing subjects.

Figure 5.1: Sample frames of close interaction and strong mutual occlusion.

Our proposed approach makes following main contributions:

- a motion-based estimation mechanism which gives out the body orientation by considering the person dynamics;
- a 3D appearance model based estimation approach by combining a 3D human body/appearance model with 2D template-based matching, being able to deal with still-standing or slowly moving people, while covering a full 360 degrees range;
- a dynamic hybrid strategy that efficiently combing the advantages from both mechanisms, by taking care that the appearance model is constructed automatically, using the first estimate of the orientation, as well as achieving a speed as close as possible to real-time performance.

The remainder of this chapter is organized as follows: Chapter 5.2 gives out the general overview of proposed approach, followed by the motion-based orientation estimation scheme in Chapter 5.3 and 3D appearance model based orientation estimation approach in Chapter 5.4. The proposed dynamic hybrid strategy is described in Chapter 5.5. Chapter 6.5 describes and discusses the experimental results, and Chapter 6.6 concludes the chapter.

5.2 Overview of the Approach

In this section, an overview of our approach for hybrid human body orientation estimation is described, with the details on the specific components that are involved. The flow chart is

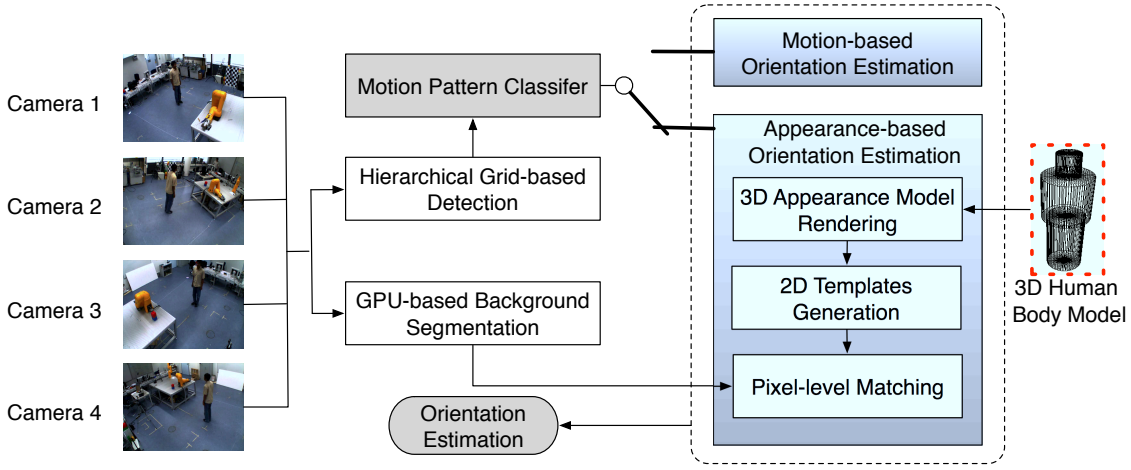


Figure 5.2: Flow chart of the proposed approach on hybrid human body orientation estimation. According to the motion pattern of targets, if its motion has significant speed, motion-based orientation estimation works. Instead, the appearance-based orientation estimation would be automatically launched when the target moves extremely slowly or stops.

outlined in Fig. 5.2. The motion pattern classifier decides which mechanism will be adopted, if the motion of targets has significant speed, the mechanism of motion-based orientation estimation starts, insteadly, the appearance-based orientation estimation mechanism will be launched when the target moves slowly down or stops.

Now we go into full details: after acquiring a frame from all the cameras, we rely on the output of hierarchical grid-based detection which has been explained in Chapter 4, to obtain the location of each person in the scene, which is going to be used as a position reference.

Within motion-based estimation, the orientation of human body is given by target dynamics, under motion with significant speed. The appearance-based estimator, instead, is automatically launched whenever the target moves extremely slowly or stops. The latter, combines a 3D human body/appearance model with 2D template matching: the 3D model is represented by a colored point cloud, which is obtained by back-projecting foreground image pixels onto the surface of the predefined body geometry at the given pose, makes our approach independent of the camera viewpoint. Notice that the body geometry is composed of three cylinders with elliptical section. And in order to avoid collecting background pixels onto the appearance model, a fast GPU-based foreground segmentation is firstly utilized on each view.

Subsequently, the appearance model is rendered onto each view at multiple poses and multiple 2D templates are generated, in order to match the reconstructed appearance to new images

under arbitrary orientations, through a robust similarity function, thus leading to the orientation estimation result.

5.3 Motion-based Orientation Estimation

If a person is walking with significant velocity, the body orientation of this person is usually expected to be aligned with the direction of motion. Given a motion vector $M_{i,j} = (p_x(t) - p_x(t-1), p_y(t) - p_y(t-1))^{\mathrm{T}}$ between two corresponding poses at time $t-1$ and t, orientation estimate could be inferred by means of computing the direction of this motion vector. Notice that this estimate requires a data association to be available (at least a sub-optimal one, based on two adjacent frames), and for this purpose we adopt the Hungarian algorithm for global nearest neighbor assignment, based only on adjacent target locations.

As we estimate the orientation in world coordinate, a reference vector $M_{i,j}^{ref} = (0, c)^{\mathrm{T}}$ is assumed with the start point of $(p_x(t-1), p_y(t-1))$, where c could be an arbitrary constant. Thus the direction of motion is,

$$\theta_r^{motion}(t) = \arccos \frac{M_{i,j} \cdot M_{i,j}^{ref}}{|M_{i,j}| \cdot \left|M_{i,j}^{ref}\right|} \ . \tag{5.1}$$

If the target slows down significantly we can rely on the previous orientation estimate, and therefore we filter the orientation estimates by a first order autoregressive filter, with adaptive coefficients regarding to the target's velocity. The final orientation estimate at timestep t is given by,

$$\theta_r(t) = \alpha\theta_r(t-1) + (1-\alpha)\theta_r^{motion}(t) \ , \tag{5.2}$$

in which α is defined according to a lower velocity threshold Th_{still}

$$\alpha = \begin{cases} 0.1, & if|M_{i,j}| > Th_{\text{still}} \\ 0.9, & else \end{cases} \ . \tag{5.3}$$

However, if the variation of a target's motion vector $M_{i,j}$ is continuously under the threshold Th_{still} for consecutive timesteps N_{still}, then it is accounted as still or extreme slowly moving. As discussed before, in this case the above motion-based estimation scheme can not provide reliable estimate for body orientation, thus some methodology else being able to efficiently deal with this intractable case is imperatively in need. Consequently, a 3D appearance model-based orientation estimation method, being capable of give out the orientation of targets whose

velocity are even close to zero, is launched and integrated with former motion-based estimation scheme. Further details will be described in Chapter 5.4.

We notice that the motion-based scheme automatically provides a reliable initial orientation for appearance model initialization, thus the target is not restrictive to be more or less at the center of the observation space and face a certain direction at the beginning.

5.4 3D Appearance-based Orientation Estimation

To compensate the shortcoming of motion-based method, we propose a 3D appearance-based estimation method, which could be able to deal with still-standing or slowly moving people, therefore combining all the advantages from the components of motion, 3D shape, appearance in a unified framework.

5.4.1 3D Appearance Model Construction

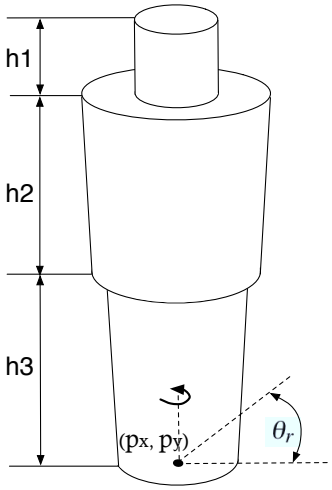


Figure 5.3: 3D geometrical body model and pose parameters ((p_x, p_y, θ_r), where (p_x, p_y) indicates the body location and θ_r indicates the possible orientation.

If we take into account the flattened shape of a human body along the depth dimension, we can approximate the overall geometry by three cylinders with elliptical section, as illustrated in Fig. 5.3, enclosing the head, torso and legs.

Given an estimated body location (p_x, p_y) and orientation θ_r, respectively provided by the detector and motion-based orientation estimator, a 3D appearance model is automatically

reconstructed by back-projecting pixels from each camera view (Fig.5.4(c)) onto the respective surface point (Fig.5.4(d)) using the 3D geometry model shown in Fig.5.3.

Moreover, in order to avoid erroneously collecting background pixels we first perform a background subtraction, with a GPU-based foreground-background segmentation method based on an extended colinearity criterion, which is firstly proposed by Griesser et al. [209].

We here briefly summarize this method: it compares the color values at pixels in a reference (background) image, and a given image. In particular, all color values within a small window (3×3 neighborhood) around a pixel are used for comparison. And change detection is based on a colinearity criterion, if they are colinear, no change is judged to be present and the background is still visible at the pixel in given image. Conversely, if not, the pixels are considered to have different colors, and a foreground pixel then is found. This colinearity criterion is further integrated into a MRF framework, solved in an iterative manner. A priori knowledge is integrated through the change mask of the prior frame as well as the compact property of connected foreground regions. Additionally, it adds an additional component for darkness compensation, which helps the algorithm to correctly detect foreground regions even if the object's color is nearly black [209].

Note that, within this GPU-based foreground-background segmentation, the background model is also learned from a set of frames without people, as for the edge-based background subtraction of Chapter 4.3.2. Fig.5.4(b) shows an example result of this algorithm. The resulting segmentations are smoothly shaped and provide reliable input for 3D appearance model construction.

The result is a large set of N model points x_n (Fig.5.4), including position, color values, and local surface normals from the underlying 3D surface. This constitutes a sparse appearance model, that we denote by

$$M \equiv \{(x_1, v_1, n_1), \cdots, (x_N, v_N, n_N)\}. \tag{5.4}$$

5.4.2 Matching through Planar Reprojection

Once the sparse point cloud is available, pixel-level measurements can be obtained by reprojecting the visible part of the cloud (appearance model) onto the respective image planes, by using the (3×4) linear camera projection matrix given by,

$$y_k = K_k T_{kw} T_{wo} x_o;\ k = 1, \ldots, N_{cam}, \tag{5.5}$$

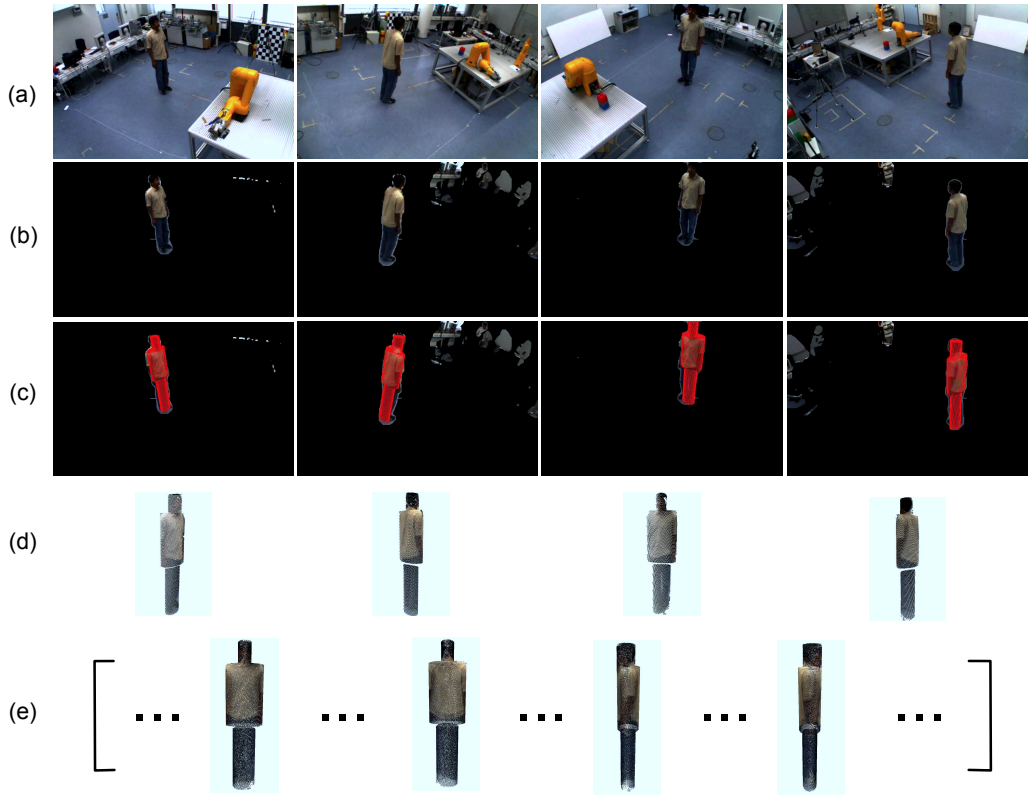


Figure 5.4: Appearance model reconstruction. (a) Original input frames from 4 views. (b) Corresponding foreground images. (c) Detected target, with geometry model superimposed onto foreground images. (d) Back-projected partial 3D cloud, at each view. (e) Final 3D appearance model, covering 360°, with some key-poses shown.

where x_o is a local model point in homogeneous coordinates, y_k is the corresponding image pixel, K is the intrinsic camera projection, known from off-line calibration

$$K = \begin{bmatrix} f_x & 0 & p_{c,x} & 0 \\ 0 & f_y & p_{c,y} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \tag{5.6}$$

with f_x, f_y the focal lengths, $p_{c,x}$, $p_{c,y}$ the principal point, and T_{kw}, is the camera-to-world constant transform, also known by calibration. Finally, T_{wo} is a homogeneous (4×4) transformation matrix that represents the target pose

$$T_{wo} = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}, \tag{5.7}$$

where $t = [p_x, p_y, Z = 0]^T$ is the location on the floor, and the (3×3) rotation matrix R is expressed in terms of XYZ Euler angles (of which only γ is updated by the orientation

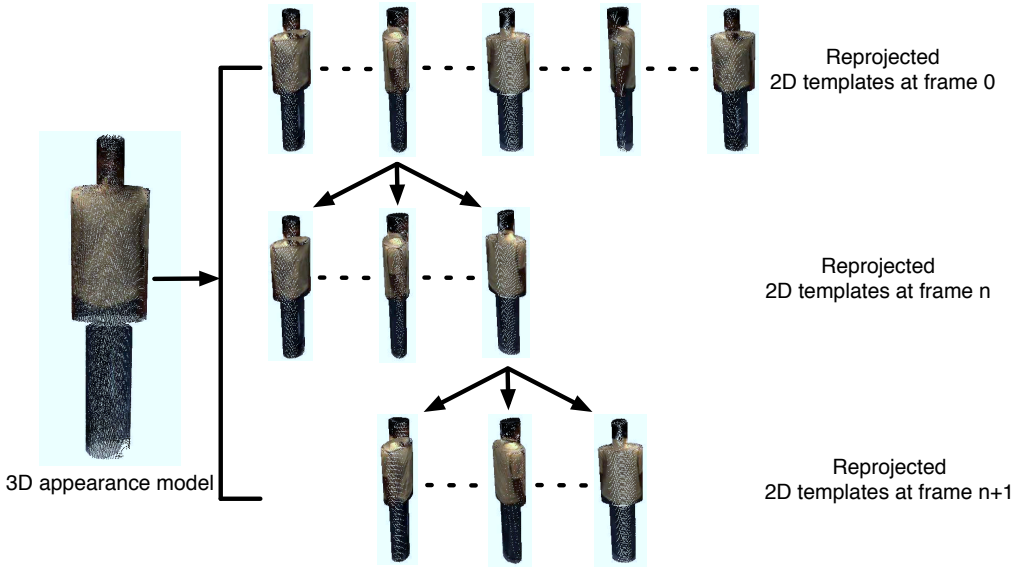


Figure 5.5: Planar templates are obtained by reprojecting the 3D appearance model in different poses, from different camera views. The 2D templates are generated according to all possible poses at the first frame, and afterwards, a prediction mechanism is employed for computational efficiency, that is, the templates are only generated on the poses which are in a fixed range around the former estimation.

estimation):

$$\begin{aligned} \theta_r &= [\alpha, \beta, \gamma]^T \\ R(\theta_r) &= R_x(\alpha)R_y(\beta)R_z(\gamma). \end{aligned} \tag{5.8}$$

Given a model point (x_o, y_o, z_o) in world coordinates, the corresponding location (x_s, y_s) in camera coordinate can easily be obtained via (5.5). During the reprojection, it is worth noting that, due to the rotations of the body, the visible part of the sparse 3D point cloud should change, only a roughly 180° slice of the point cloud is visible in any particular frame, that corresponds with the visible portion of the human body in each video frame.

Thus, it is necessary to test visibility of each point: since the shape is almost everywhere convex, a point p is visible from camera c if the angle between its normal and the camera projection ray through p is less than 90°, that is

$$V_c \cdot n_c < 0 \tag{5.9}$$

where V_c is the viewing vector (i.e. the position of p in camera coordinates) and n_c is the respective normal direction.

The point projection (5.5) and visibility test (5.9) are done at each pose hypothesis θ_r. Fig. 5.5 provides an example of re-projected templates on one camera view.

Subsequently, template matching amounts to evaluate a likelihood function that robustly compares template colors with the underlying foreground pixels. For a predicted pose $\hat{T}_t$ at time t, our similarity measure between the N_k color pixels u from the reprojected template $h_k(\hat{T})$ onto camera k, and the corresponding pixels v of the foreground image, is defined as

$$D = \sum_{k=1}^{N_{cam}} \frac{1}{N_k} \sum_{u \in h_k} \sqrt{\sum_{ch \in (r,g,b)} (u_{ch} - v_{ch})^2} \tag{5.10}$$

which is the sum of absolute pixel-wise difference over (r, g, b) channels.

In this formula we adopt an isotropic L_1-norm that, compared to classical L_2-norm, it is more robust to outliers, such as erroneous colors sampled from the background, as well as non-Gaussian noise statistics. The corresponding likelihood is a Laplacian distribution

$$P(z|s)_{ori} = \frac{1}{2\sigma} \exp\left(-\frac{D}{\sigma}\right) \tag{5.11}$$

where the pose is described by $s = (p_x, p_y, \theta_r)$, and σ is the precision parameter of this distribution.

During orientation estimation, we employ a simple but effective prediction mechanism for computational effiency. In fact, during normal walking it is unlikely for a person to turn more than 90 degrees over one frame of the sequence; therefore, after the initial pose estimation, reprojection and matching are performed only on the orientations which are in a fixed range around the former estimation, thus saving computation while reducing estimation error. Fig. 5.5 illustrates this strategy across frames.

Similarly to the motion-based estimator, this prediction requires a coarse data association between two adjacent frames, also given by the aforementioned position-based Hungarian algorithm.

5.5 Dynamic Hybrid Strategy

To intuitively and comprehensively describe our proposed dynamic hybrid strategy on orientation estimation, we explain it by illustrating one of those typical cases in Fig. 5.14. As it can be seen, during the short period (frame 2377 - frame 2434), the person with beige color rotates itself at his place, thus the 3D localization result remains almost unchange throughout this period. As motion-based estimation approaches rely on motion vector between two

corresponding poses at adjacent time steps, and in such case, the motion vector $M_{i,j}$ remains zero thus it is not able to provide reliable output for orientation estimation. Nevertheless, if this situation continuously remains for a consecutive timestep N_{still}, a 3D appearance model is automatically constructed for this person and corresponding 3D appearance-based orientation estimation method is launched. As we can see from Fig. 5.14, 3D appearance-based estimator can quite robustly give out the orientation of this person even though he is standing still during this period. Therefore, with the hybrid combination of motion-based approach and 3D appearance-model one, we take full advantage of low computational cost of the motion-based orientation estimator and high-precision of 3D appearance-based estimator, consequently making them mutually beneficial to keep both their advantages while compensating for the limits of each.

5.6 Experimental Results

In this section, we demonstrate our approach on the problem of estimating human body orientation in 3D space. As we focus on demonstrating the ability of the approach on dealing with very ambiguous situations, such as people interact with each other for long time with similar appearance, standing extremely close to others and being fully occluded on some camera views, or they walk very slow or even remain static over a long period. Therefore we record our own dataset which exhibit a diverse set of aforementioned features, to test the effectiveness of our proposed hybrid human body orientation estimation approach. Qualitative results as well as quantitative evaluation are both presented in this section.

5.6.1 Qualitative Results

We evaluate our orientation estimation algorithm through four video sequences, showing multiple people that move and turn freely, as well as interacting and occasionally occluding each other in some views. The sequences have been simultaneously recorded from all cameras, as described in Chapter 3.2, with a resolution of (752×480) and a frame rate of 25 fps.

Furthermore, during the initialization phase of our system, the 3D appearance model of each target is automatically reconstructed according to the detected 2D location using the technique described in Chapter 4 and known orientation provided by motion-based method (Chapter 5.3),

Note that the body orientation θ_r is continuously estimated by the motion-based method, while in the appearance-based method, for computational effiency, 12 discrete orientations

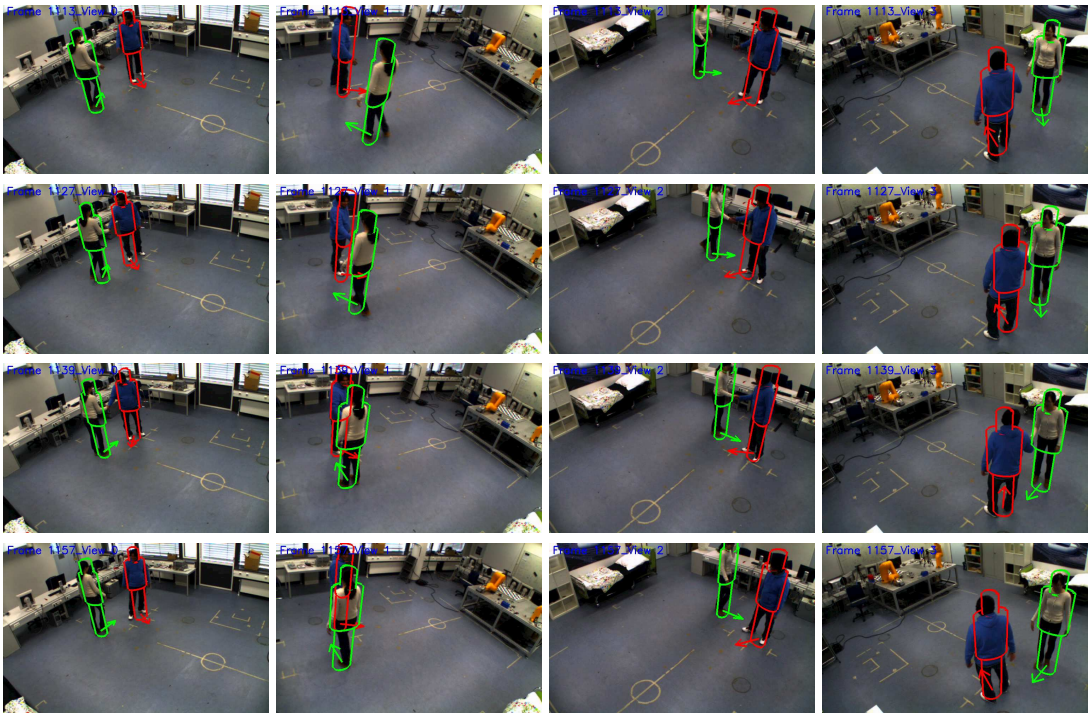


Figure 5.6: Human body orientation estimation results on the sequence of Laboratory 2 Targets with longterm handshake. Every row consists of views from four different camera views at the same time frame. The arrow indicates the orientation of the human body in 3D world space.

covering 360° are utilized during the reprojection of the appearance model onto the image planes, and at subsequent frames the model is reprojected only within 90° around the previous estimate.

Figs. 5.6 to 5.9 show the example orientation estimation results on each of the evaluated sequence. In each image, the orientation of each target is indicated by a colored line, with arrow pointing to the estimated orientation. And the silhouette of the geometry model is also superimposed onto each target, to illustrate the tracked location that is used as reference.

Following we will go into more discussions on each corresponding sequence.

Laboratory 2 Targets Within this sequence (Fig. 5.6), there are two targets involved, who are shaking hands with each other. During sample frame 1113 and frame 1157, we can see that the two persons walk, meet, shake hands and then separate away. In frame 1127 and 1139, the person's speed magnitude is very small, in this case, the appearance-based method is launched and works efficiently, so that the orientations of both persons are correctly estimated.

Laboratory 3 Targets This sequence (Fig. 5.7) involves three targets, during this se-

Figure 5.7: Human body orientation estimation results on the sequence of Laboratory 3 Targets. Every row consists of views from four different camera views at the same time frame. The arrow indicates the orientation of the human body in 3D world space.

quence, people are walking with notable speed in most case, and according to our model, the motion direction provides a good prior for the body orientation. In this sequence we also emphasize the challenges due to mutual occlusions, from one or more views. At frame 2582, although people keep very close to each other, in addition, the person superimposed with green cylinder is occluded by others simultaneously on view 0 and 2, our estimation results are still satisfactory, thanks to the multi-camera environments that resolving the ambiguities. Additionally, we notice that during frame 2525 and 2582, the target superimposed with red cylinder remains almost static, however its body orientation is correctly estimated.

Laboratory 4 Targets - Interaction In this sequence (Fig. 5.8), there are more targets involved, and it features a more challenging scenario consists of long-term interaction and still-standing case. As we can see from frame 2456 until frame 2500, the two persons correspondingly with green and blue cylinder get extremely close with body oriented towards each other and have close interaction across several frames, while their speed magnitude is particularly small. Our approach is still able to successfully estimate correct body orientation from noisy observations.

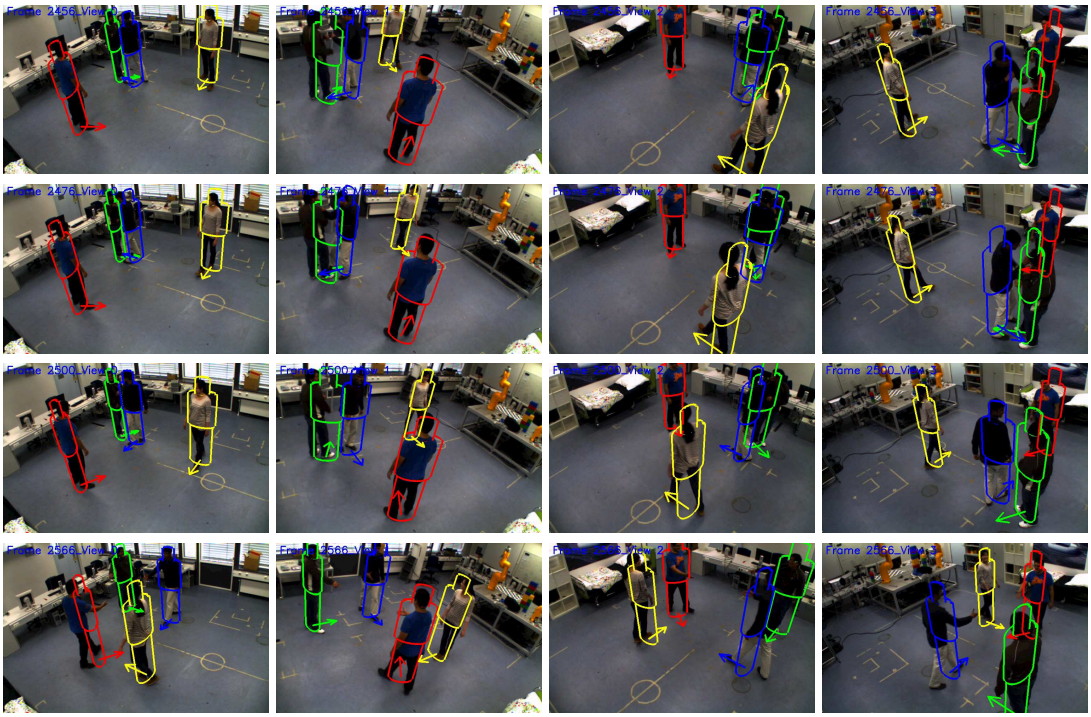


Figure 5.8: Human body orientation estimation results on the sequence of Laboratory 4 Targets, aim to evaluate the performance with longterm interaction and still standing case. Every row consists of views from four different camera views at the same time frame. The arrow indicates the orientation of the human body in 3D world space.

The success on one hand demonstrates the effectiveness of our approach, and on the other hand we should thank to our hierarchical grid-based detector that provides reliable position reference even in case of close proximity. Again, as can be seen from the four sample frames, the target with red cylinder remains static for a long time, and our approach successfully gives out correct estimation.

Laboratory 4 Targets - Crossing This sequence (Fig. 5.9) also involves four targets, however, comparing to previous sequence we have a different emphasis on evaluating the performance. Firstly, we can see the challenges as targets are strongly occluded by each other from one or two views almost in each sample frame. Secondly, between sample frame 912 and 985, it clearly illustrates a process that the person with blue cylinder is passing between the other two people respectively in red and green cylinder, while their spatial locations are very close to each other. Thirdly, in frame 1061, the two persons in red and green cylinder are not only with very close proximity but also standing almost still over few frames. Nevertheless, our approach

Figure 5.9: Human body orientation estimation results on the sequence of Laboratory 4 Targets, aim to evaluate the performance when target walking across others with very close proximity. Every row consists of views from four different camera views at the same time frame. The arrow indicates the orientation of the human body in 3D world space.

performs quite well under above various challenges.

In a nutshell, the motion direction provides a good prediction of body pose, is reliably exploited while the targets have signifant speed. And when people are static (i.e. not moving forward), e.g. while waiting of during interaction, our appearance-based method demonstrates its advantage and gives a very reliable compensation. Therefore, despite various challenges of this task as mentioned above, our proposed approach successfully estimates the body orientation in most cases.

5.6.2 Quantitative Evaluation

In order to evaluate more precisely the performances of our approach, we compare our method with the ground truth data on a per-frame basis. To this aim, we firstly manually label the position of each target for each frame of the sequences with the annotation method introduced in Chapter 3.3.1, meanwhile, its respective body orientation is annotated by rendering the 3D

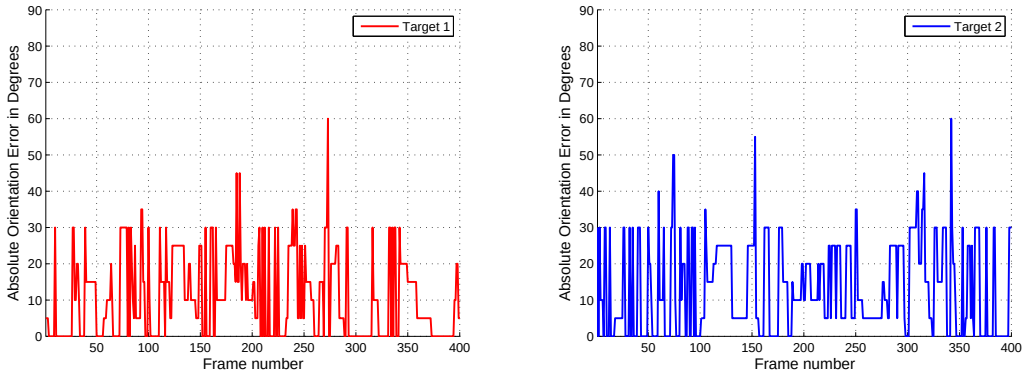


Figure 5.10: Ground truth evaluation on sequence Laboratory 2 Targets. The errors are in degree.

elliptical body model and visually matching it with the body of the target in all views where the person can be seen. The annotated body orientation is taken to be the person body facing direction and labeled with an interval of 5 degrees.

For sequence Laboratory 2 Targets and Laboratory 3 Targets, the most challenging clip, covering 400 frames from both sequences, was selected for ground truth evaluation. The results are shown in Fig. 5.10 and 5.11 respectively, which shows the absolute error between estimated orientation and ground truth for each person. As it can be seen from the results, the orientation errors are most of the time below 30 degrees.

For sequence Laboratory 4 Targets - Interaction and Laboratory 4 Targets - Crossing, we correspondingly select the most challenging clip with 500 frames for quantitative evaluation, the absolute errors are illustrated in Fig. 5.12 and 5.13 respectively. These two sequences are much more challenging compared to previous twos, one is because of the increased number of targets, leading to more crowded scene, thus it will challenge the output of our detector, consequently influencing the performance of orientation estimation; on the other hand, within these two sequences, long-term interaction, mutual occlusion and still-standing cases are frequently presented. Despite the increased challenges, the absolute orientation errors are still below 30-40 degrees mostly. In addition, we notice that in Fig. 5.12, the absolute orientation error of target 1 and 2 remain almost constant for quite a long period, as we mentioned in previous section, target 1 and 2 get very close and interact with each other for long time during this period, their speed is extremely slow even gets to be static. In this case, motion-based approach can not provide reliable output, nevertheless in our estimation scheme, it will automatically turn to

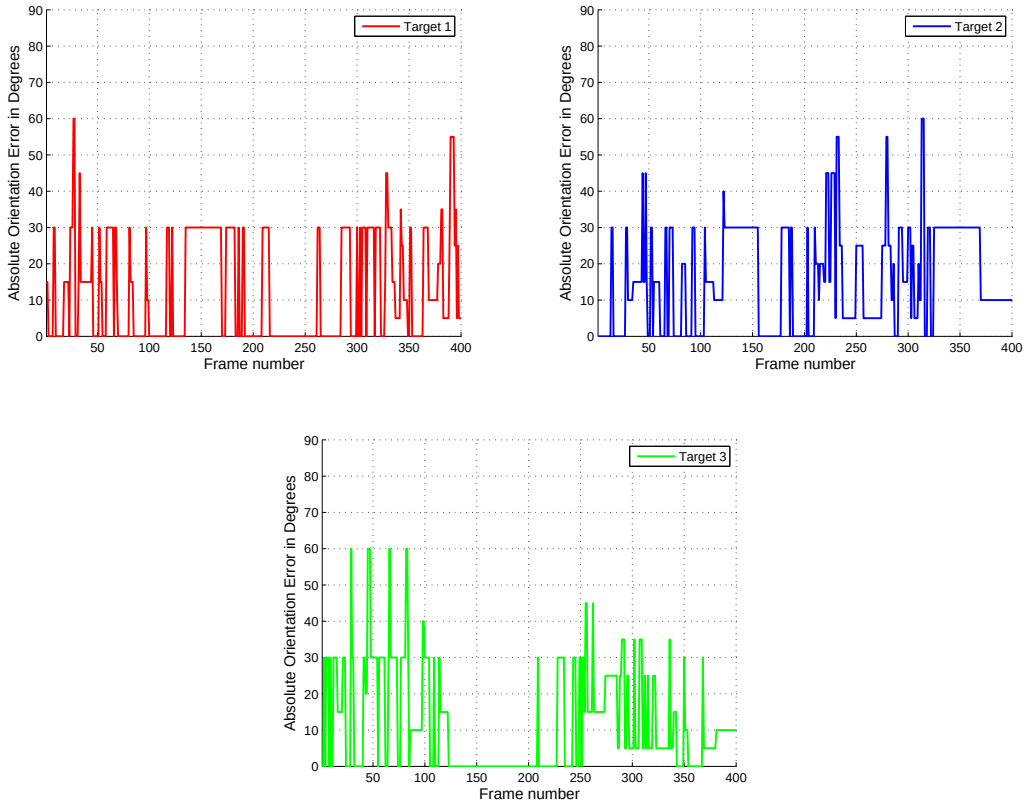


Figure 5.11: Ground truth evaluation on sequence Laboratory 3 Targets. The errors are in degree.

the 3D appearance-based approach, and it can be seen from Fig. 5.12, very robust orientation output is provided through this methodology with very low absolute orientation error, even though the targets stand very close to each other which is hard for localization.

Our approach has demonstrated a good robustness in various cases through evaluation. By resuming, the effectiveness of the approach is due to four main aspects: one is the reconstruction of a detailed 3D appearance, that makes a pixel-level matching more precise with respect to global or local statistics, such as color histograms. The second is the dynamic hybrid strategy of combining motion-based and appearance-based method, so that efficiently handling various cases. The third is the use of a robust multi-target detector, since a good location reference is necessary for an accurate orientation estimation. And the fourth is the integration of calibrated multi-camera views, both for position and orientation estimation, so that the ambiguities can be efficiently resolved.

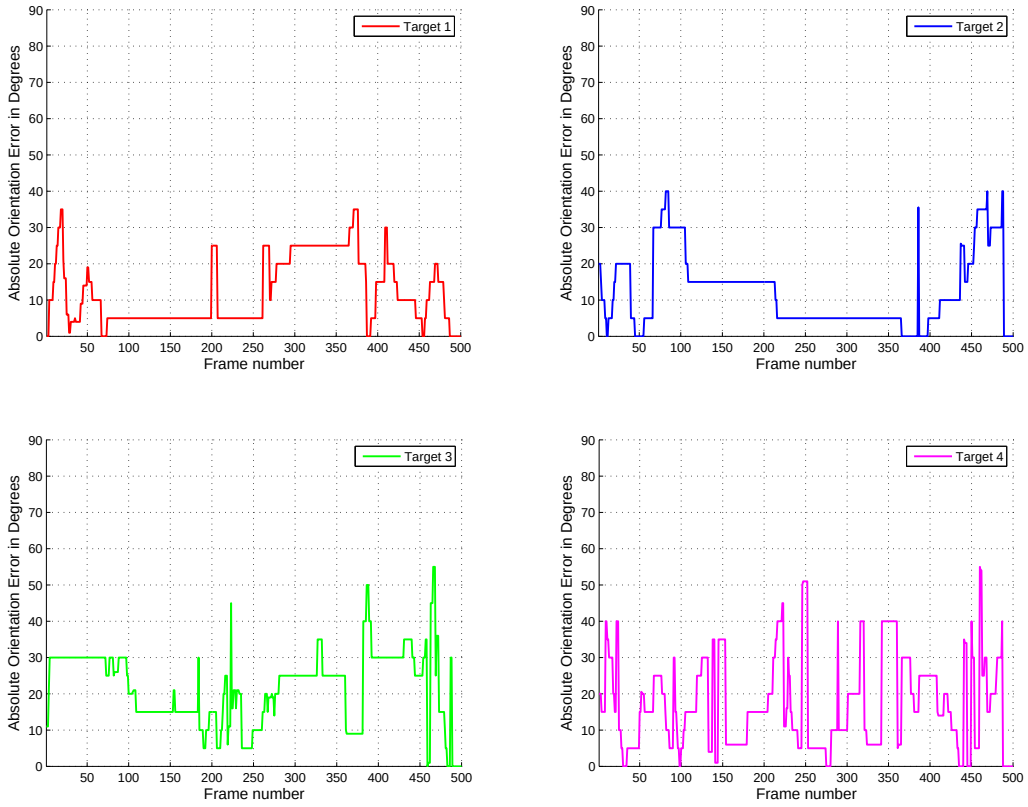


Figure 5.12: Ground truth evaluation on sequence Laboratory 4 Targets − Interaction. The errors are in degree.

5.7 Conclusion

In this chapter we presented a robust algorithm for estimating the body orientation of multiple people simultaneously in a calibrated multi-camera environment. The merits of motion-based and 3D appearance-based orientation estimation method is dynamically combined, providing a solution to the crucial requirements on not only speed but also automatic initialization of 3D appearance model, as well as being capable to deal with still-standing or slowly moving people. Experiments over several real-world sequences have been performed and also evaluated against ground-truth data, thus the validity of our method is demonstrated. The experimental results evince that our approach could reliably estimate the body orientation in 360° scope.

Apart from facilitating an accurate orientation estimate, the proposed method has further potential to offer disambiguation for improving tracking performance in multiple people tracking

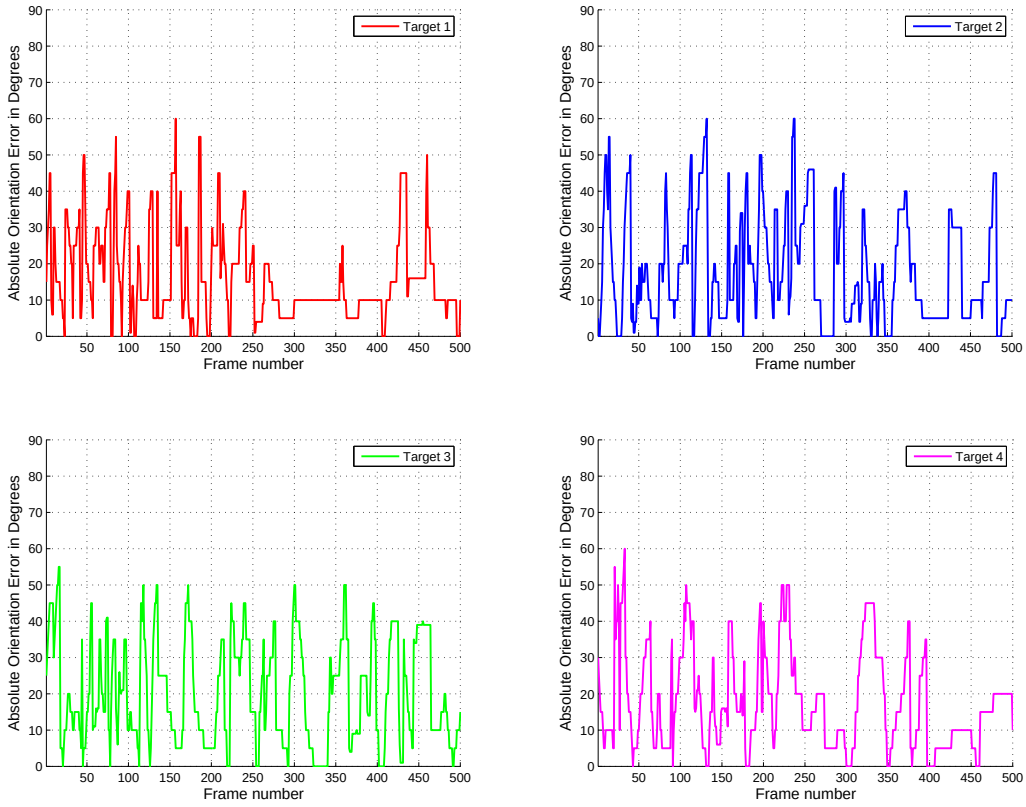


Figure 5.13: Ground truth evaluation on sequence Laboratory 4 Targets − Crossing. The errors are in degree.

scenario. In the next chapter, we show how the body orientation can be efficiently integrated into the data association problem, and how it provides ability on revolving ambiguities between crossing trajectories.

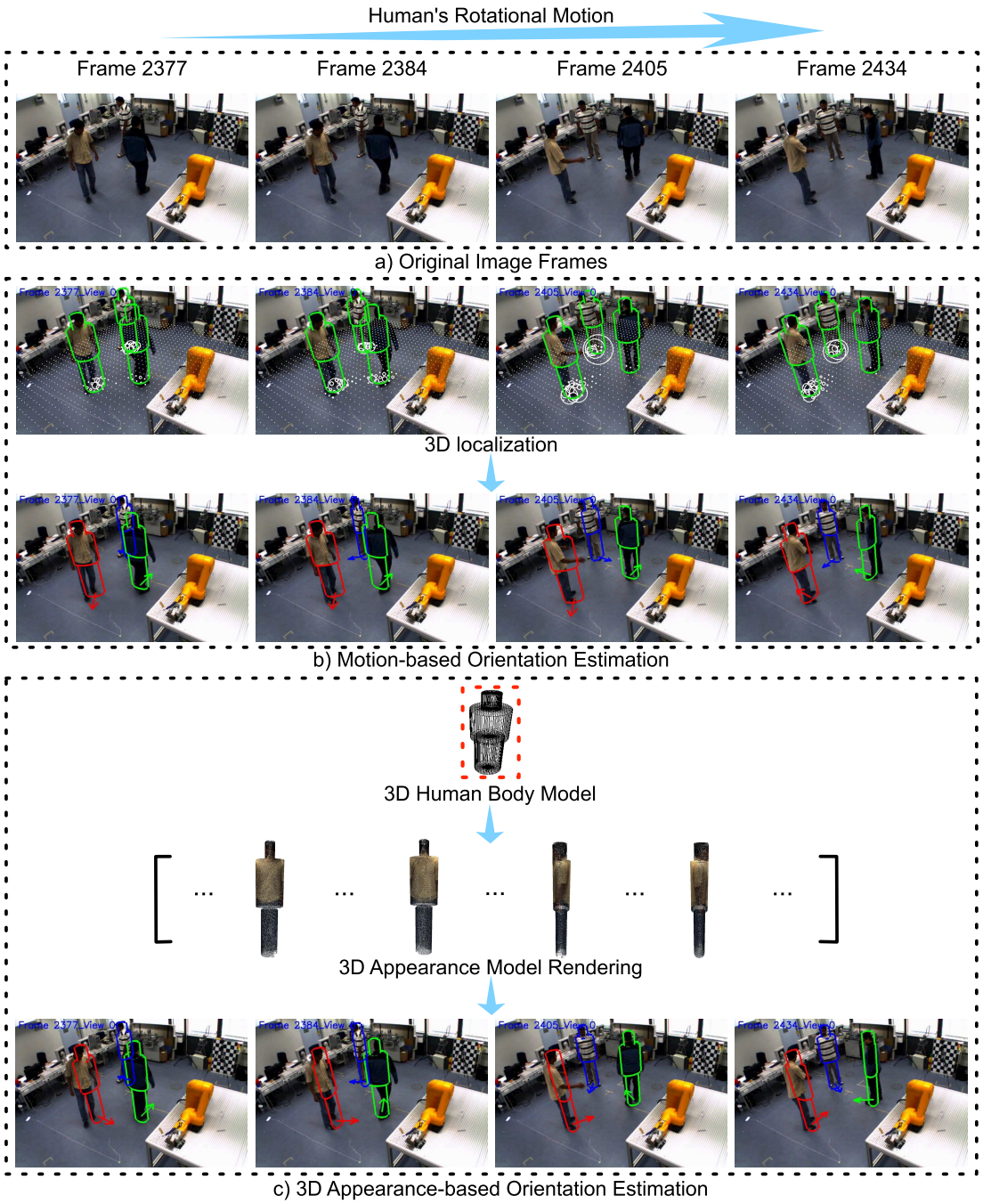


Figure 5.14: An illustration of the proposed hybrid strategy on estimating body orientation. It shows a typical case that if the person does not walk with significant velocity, the motion-based orientation estimation method becomes ambiguous, however the 3D appearance-based method can provide very robust estimation result in such case. Our hybrid strategy can keep both their advantages while compensating for the limits of each.

Chapter 6

Global Optimal Data Association for Multiple People Tracking

6.1 Introduction

Multiple people tracking is an intensively studied area in computer vision. Its primary goal is to retrieve the trajectories of targets by localizing the targets individually at each frame, and maintain their identities throughout the video sequence. Tracking multiple people accurately in cluttered and crowded scenes is a highly challenging task due to frequent occlusions between people, low resolution, abrupt motion, illumination and appearance changes, in particular, similar appearance and complicated interactions between different targets often result in tracking fails such as track fragmentation and identity switches.

Tracking-by-detection approaches, with the advantage of being resistant to divergence, have demonstrated impressive results in addressing these challenges. Such approaches involve two steps, namely independent detection in individual frames, and association of observations across frames. With the output from our hierarchical grid-based people detector (Chapter 4) as evidence for tracking, the main aim of this chapter is to link the detections across frames, allowing to track and identify targets even though complex interactions and significant mutual occlusions with automatic initialization and termination.

As discussed in Chapter 2.3, classic data association approaches such as JPDAFs or MHT has the drawback of an exponentially growing computational complexity with the increasing number of targets and measurements. Additionally, they can not guarantee a global optimal solution in sub exponential time even though they attempt to model joint trajectories of all

objects. Thus, the data association approaches based on global optimization by using Linear Programming, being able to optimize all trajectories simultaneously and optimize the whole video sequence, have appeared to be recent popular.

Some algorithms have been proposed for the global optimal association framework by integrating several cues such as appearance, size, location, and motion into an affinity model to measure similarity between detections or tracklets [171, 174, 177, 210–212]. However, the main evidence used to link detections of tracklets is still visual cues, there has been seldom effort addressed to explore the benefit from human behavior, which is a higher-level reasoning evidence. As already illustrated in Fig. 5.1 in Chapter 5.1, showing an example that targets are interacted extremely close or highly occluded by each other, with similar position even appearance, and remains almost static over a few frames, in this case, the behavior cue - body orientation provides valuable insight into the dynamics of a social interaction.

In this chapter, we propose a global optimization approach for long-term tracking of an *a priori* unknown number of targets, randomly walking in the environment with variable number of cameras. And the problem of complex interaction and mutual occlusion is addressed by exploiting a consistency scheme on behavior cue, as well as compensating measurements of location and appearance, in which the related features are respectively provided by the hierarchical grid-based people detector (Chapter 4) and hybrid human body orientation estimation method (Chapter 5).

More precisely, the multiple target tracking problem in this work is formulated in terms of finding the global maximum of a convex objective function, then solved efficiently through a linear programming relaxation. By taking the full advantage of our hierarchical grid-based detector, the regular discretization scheme is further adapted here. With this particular scheme, a grid-based network flow model is constructed, in which the nodes and edges encoded correspondingly, as inspired by the work of [184]. This scheme allows to effectively avoid intermediate hard decisions and simply model mutual occlusion because of the specific graph structure. To enable the tracker recovering from mis-detections, we carry out non-maxima suppression during tracking rather than during detection, with the contrast to previous approaches that the state-space only consisting of observations, which are not able to interpolate trajectories smoothly in case of false negatives. Moreover, the measurements of body orientation, target location and appearance are incorporated in a global manner. The explicit use of behavior cue can disambiguate the situation such as in Fig. 5.1. This is distinctive compared to many state-of-the-art approaches that only using purely visual cues (like appearance and motion information).

The remainder of the chapter is organized as follows. Chapter 6.2 describes the general system overview with hardware setup and algorithmic flow of software. A classic data association method without global optimization is given in Chapter 6.3, in order to serve as one of the baselines while comparing quantitatively with the global optimal solution. And a global optimal data association approach is described in detail in Chapter 6.4 including problem formulation and optimization framework. Chapter 6.5 presents and discusses experimental results through different indoor and outdoor scenarios. At last, concluding remarks are drawn in Chapter 6.6.

6.2 Overview of the Approach

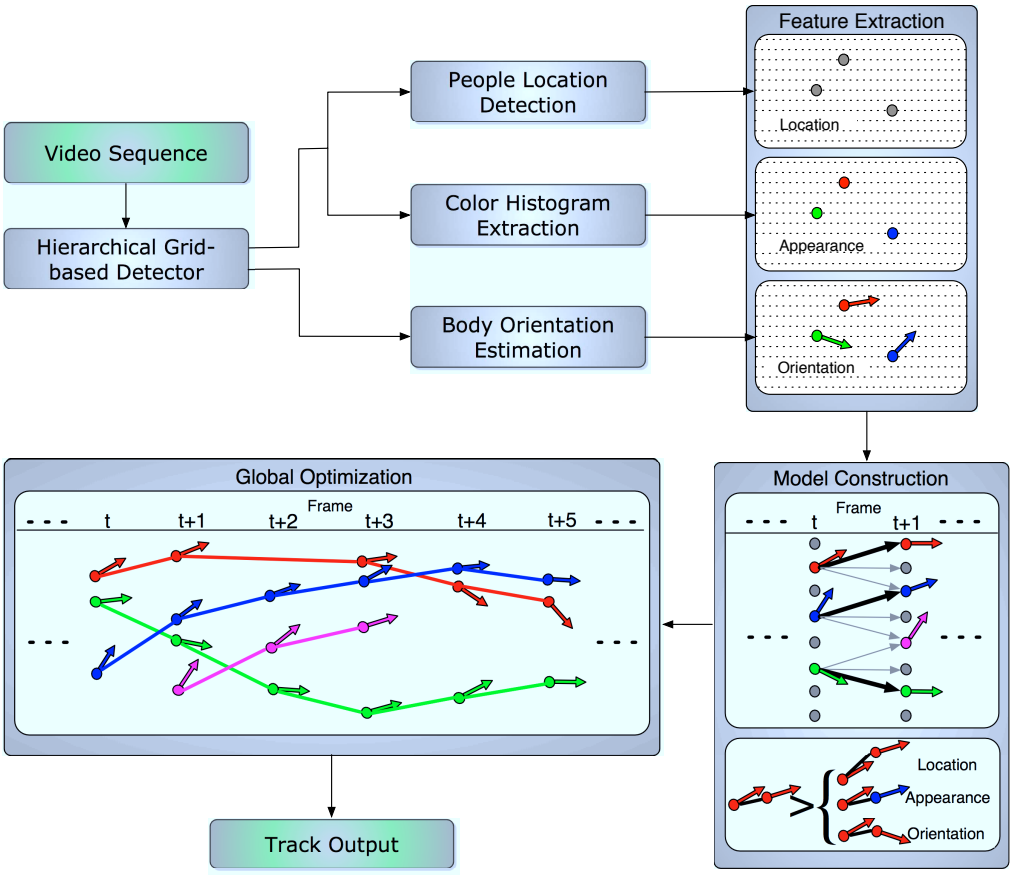


Figure 6.1: Overview of the proposed approach on global optimal data association for multiple people tracking.

The flow chart of our proposed approach is outlined in Fig. 6.1. After acquisition of original frames from all the cameras and hierarchical grid-based detection, the potential observations

are obtained. Note that in order to allow tracker to recover the most probable locations in accordance with all evidences, we eschew non-maxima suppression during detection here. With the output from detector, each observation is characterized by a descriptor that records the features including location and appearance. However, it is not sufficient for a people tracking approach to determine data association only according to the location reference and appearance model, e.g. tracking may fail if two targets get very close or wear similar clothing. To overcome this limitation, we incorporate a discriminative cue on body orientation, which is estimated by utilizing the technique proposed in previous chapter. As proposed in our detector, the state space is partitioned into integral grids with a coarse-to-fine strategy. We follow the discretization structure in data association part, with the per-frame measurements sampled on regular grids. A grid-based network model can be constructed afterwards as concisely illustrated in model construction part, while the corresponding detailed model will be shown in Chapter 6.4.1. A consistency scheme on behavior cue, as well as measurements of location and appearance, is modeled as transitional cost between nodes at two consecutive time steps. As illustrated in model construction part, the affinity measure can achieve highest only if the nodes have simultaneous similarity on all cues of location, appearance and orientation. Next follows the global optimization part, consists of formulating the data association problem as finding the global maximum of a convex objective function, which in our work is solved by a linear programming relaxation, and at last leading to track output with identity associated to each target.

6.3 Classic Data Association without Global Optimization

Before we go to full details on the global optimization approach for dealing with data association problem, in this section we first describe a classic data association method, in order to serve as one of the baselines while comparing quantitatively with the global optimal solution.

The data association problem consists in deciding which measurement should correspond to which track. Although our detection algorithm is fairly robust, it is also not person-specific, and therefore in a small observing environment there are always ambiguities, arising from neighboring targets, as well as from missing detections and false alarms caused by background clutter. To this respect we employ the Global Nearest Neighbor (GNN) approach, that gives a good solution for this problem [213], while requiring relative low computational cost.

The first step of the GNN is to set up a distance (or cost) matrix: assuming that, at time t, there are M existing tracks and N measurements, the cost matrix is given by

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M1} & d_{M2} & \cdots & d_{MN} \end{pmatrix}, \tag{6.1}$$

where d_{ij} is the Euclidean distance between track i and measurement j, and $i = 1, 2, \ldots, M; j = 1, 2, \ldots, N$. In particular, d_{ij} is set to ∞ if it exceeds the validation gate, which is a circle with fixed radius around the predicted position, eliminating unlikely observation-to-track pairs. Moreover, it is commonly required that a target can be associated with at most one measurement (none, in case of misdetection), and a measurement can be associated to at most one target (none, in case of false alarms).

The GNN solution to this problem is the one that maximizes the number of valid assignments, while minimizing the sum of distances of the assigned pairs. To this aim, we adopt the extended Munkres' algorithm [180], where the input is the cost matrix D, and output are the indices (row, col) of assigned track-measurement pairs.

In particular, the track management follows a strategy indicated in [214]:

- *Track initiation* In case of new targets entering into the scene, they will generate measurements that are too far from the existing targets, and therefore can be used to start new tracks. In this case, they are labeled with a unique ID, and a counter for the number of consecutive, successful detections for this target is also initialized to 1.

- *Track maintainance* During tracking, a target is successfully detected whenever the data association algorithm provides one valid measurement for it, so its counter is increased up to a maximum value (which can be taken as a confirmation time), while in case of misdetection it will be decreased. Those targets which are successfully detected over the confirmation time, can be considered as stable targets and maintained by the algorithm. In this way, if a target is misdetected for a few frames in case of occlusion, it can still be recovered until the counter goes to 0.

- *Track termination* When a target exits the scene, or after occlusion for a too long time, its misdetection counter goes to 0, and its track is terminated.

A pseudo-code of the whole procedure is shown in Algorithm 2, a track is initialized when new measurements come and no corresponding target. And a track is only maintained under

Algorithm 2 Track management with GNN

```
1: if nMeasurements = 0 then
2:   for i = 0 to nTargets do
3:     DecreaseCounter(target[i]);
4:     if Counter(target[i]) > 0 then
5:       MaintainTarget(target[i]);
6:     else
7:       TerminateTarget(target[i]);
8:     end if
9:   end for
10: else
11:   if nTargets = 0 then
12:     for j = 0 to nMeasurements do
13:       newTarget = CreateTarget(meas[j]);
14:       ResetCounter(newTarget);
15:     end for
16:   else
17:     for i = 0 to nTargets do
18:       for j = 0 to nMeasurements do
19:         D(i, j) = Distance(target[i], meas[j]);
20:         if D(i, j) > ValidGate then
21:           D(i, j) = ∞;
22:         end if
23:       end for
24:     end for
25:     (i ↔ j) = GNN(D);
26:     for i = 0 to nAssocTargets do
27:       if D(i, j(i)) ≤ ValidGate then
28:         MoveTarget(target[i], meas[j]);
29:         IncreaseCounter(target[i]);
30:         if Counter(target[i]) > MaxC then
31:           Counter(target[i]) = MaxC;
32:         end if
33:       else
34:         DecreaseCounter(target[i]);
35:         if Counter(target[i]) = 0 then
36:           TerminateTarget(target[i]);
37:         end if
38:       end if
39:     end for
40:     for j = 0 to nUnassocMeas do
41:       newTarget = CreateTarget(meas[j]);
42:       ResetCounter(newTarget);
43:     end for
44:   end if
45: end if
```

the case that $Counter(target[i])$ reaches $MaxC$. If the distance between measurement and target is lower than a given validation gate, then $Counter(target[i])$ is decreased and if it goes to 0, then this target is removed so that the track is terminated.

6.4 Global Optimal Data Association

In this section, more details are provided about the proposed global optimal data association approach, which finds global optimal solution for multi-target tracking. We start with the formulation of a grid-based network flow model, with the nodes and edges encoded. Then we transform the maximum a-posteriori trajectory estimation into an Integer Linear Programming (ILP) problem, solved through relaxation. Followed by the association affinity model, in which a consistency scheme is exploited on behavior cue, as well as the compensation with measurements of location and appearance.

6.4.1 Grid-based Network Model

We recall that the state space has been partitioned into discrete coarse-to-fine regions during detection phase (Chapter 4). Each discrete region $\left\{R^{i,l}\right\}_{i=1}^{N_l}$ is sampled at its center, where $1 \leq l \leq L$, L is total levels of state space hierarchy, N_l is the number of grids at level l. With the refinement through detection, a set of observations with world-space position then would be on the leaf level L. Assume there are N_o^t observations at time instant t, $1 \leq t \leq T$, the observation set then be $R(t) = \left\{(R_t^{1,L}, R_t^{2,L}, \cdots, R_t^{N_o^t,L}\right\}$, while the full set of observations is $\Re = \{R(t)\}$. As we avoid non-maxima suppression during detection phase, these observations may contains many false positives.

By assuming that a pre-defined maximum number of people n_{max} can appear throughout the video sequence, our goal is to find a unique track for target $n = 1, \ldots, n_{max}$, by eliminating false positives and recovering from false negatives, given by an ordered sequence of observations at all times $T_n = \left\{R_1^{i_n^1,L}, R_2^{i_n^2,L}, \cdots, R_T^{i_n^T,L}\right\}$ (some of which may be empty), where $R_t^{i_n^t,L} \in \Re$, i_n^t is the region assigned to target n at time t, and the set of all tracks is given by $\Im = \{T_n\}$.

In order to simplify the formulation, we construct a grid-based flow model inspired by the work of [184], extended to a new consistency scheme within time intervals. For N_L discrete grids and T consecutive time steps, a directed acyclic graph (DAG) with $N_L T$ nodes is introduced as shown in Fig. 6.2, in which every node represents a discrete grid at a given time step. For a simpler flow-based analysis, the nodes are represented in the form of pairs within our

model, allowing to explicitly model object dynamics through transition costs by considering the relationship of observations between two consecutive time steps. By contrast, the transition cost in the model of [184], is assigned only with the occupancy probability of corresponding grid. For any location $R^{i,L}$, that an object located at $R^{i,L}$ (which will be encoded as node i in following text) at time t can reach its neighbors $\mathcal{N}(i)$ including itself at time $t+1$. Therefore,

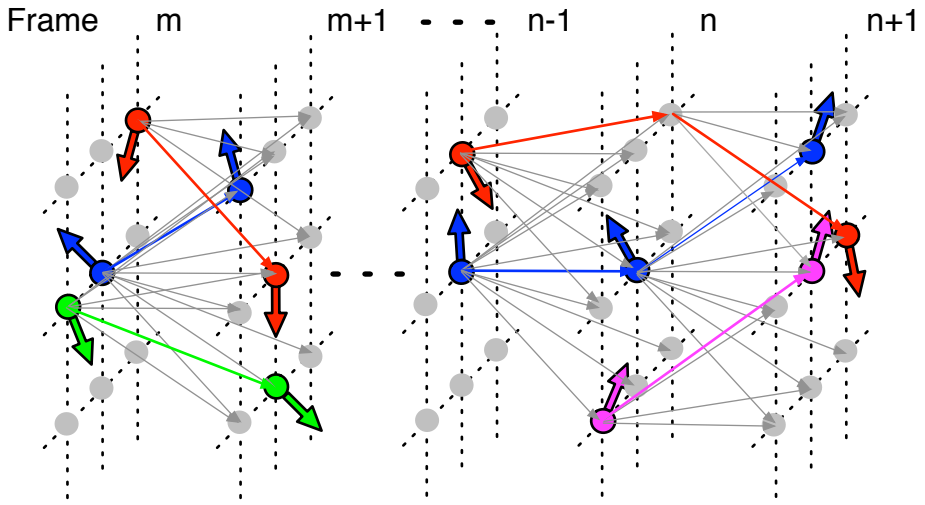


Figure 6.2: The grid-based network flow model for multiple object tracking. The nodes are represented in the form of pairs, in which the gray nodes encode the possible location of detection, the colored nodes encode the detected location of measurements while the color encodes the appearance information, and the arrow encodes the orientation information.

a path for the object starting from node i to node j is represented as $p_t^{i,j}$, valued $p_t^{i,j} \in \{0,1\}$, encoding that if the path is within part of some trajectory, that is, $p_t^{i,j} = 1$ means that the path is on the trajectory, and $p_t^{i,j} = 0$ means not. The cost $c(i,j)$ of each $p_t^{i,j}$ between node i and node j is assigned in the light of an association affinity model, which will be further described in Chapter 6.4.3.

By taking advantage of the grid-based network flow model, we define a list of constraints to guarantee that each edge(path) through the DAG is practically possible:

Continuity Potential As illustrated in Fig. 6.3, in order to enforce continuous trajectories for tracks, that for any node j, paths arriving at j at time t should be equal to the sum of paths leaving from j at time $t+1$,

$$\forall t, j, \sum_{i:j\in\mathcal{N}(i)} p_t^{i,j} = \sum_{k\in\mathcal{N}(j)} p_{t+1}^{j,k} \,. \tag{6.2}$$

Intersection Avoidance With the sampled grid resolution is sufficiently fine, no two objects should occupy the same grid at one time, thus, for any node j, the sum of paths from j should be no more than 1,

$$\forall t, j, \sum_{k \in \mathcal{N}(j)} p_t^{j,k} \leq 1 \ . \tag{6.3}$$

Initialization and Termination Scheme For automatically initializing and terminating a track, two special nodes, source and sink nodes – v_{source} and v_{sink}, are introduced into the proposed network flow model, as shown in Fig. 6.3. Both nodes are connected to *any* node in the network, and represent unknown locations for targets entering or exiting the observation area. In Fig. 6.3, to simplify we only show nodes at the first frame connected to the source, and nodes at the last frame connected to the sink.

The source and sink nodes are subject to the constraint, that all paths should start from v_{source} and end at v_{sink},

$$\sum_{j \in \mathcal{N}(v_{source})} p^{v_{source},j} = \sum_{k:v_{sink} \in \mathcal{N}(k)} p^{k,v_{sink}} \ . \tag{6.4}$$

as we already mentioned, here the neighborhood of v_{source}, v_{sink} coincides with the entire network.

6.4.2 Linear Programming Formulation

The objective of global optimal tracking is to link all the detections together over the whole sequence, choosing links so that the total probability is maximized, that is, maximizing the posteriori probability of $\mathcal{T}$ with given observation set $\Re$,

$$\begin{aligned} \mathcal{T}^* &= \arg\max_{\mathcal{T}} P(\mathcal{T}|\Re) \\ &= \arg\max_{\mathcal{T}} P(\Re|\mathcal{T})P(\mathcal{T}) \\ &= \arg\max_{\mathcal{T}} \prod_t P(R(t)|\mathcal{T})P(\mathcal{T}), \end{aligned} \tag{6.5}$$

where the last row assumes that measurements are conditionally independent between times, given the trajectories $\mathcal{T}$. If further assume that measurements are independent between targets, i.e. by ignoring the effect of close interactions, then (6.5) can be decomposed as:

$$\mathcal{T}^* = \arg\max_{\mathcal{T}} \prod_t P(R(t)|\mathcal{T}) \prod_{T_n \in \mathcal{T}} P(T_n) \ . \tag{6.6}$$

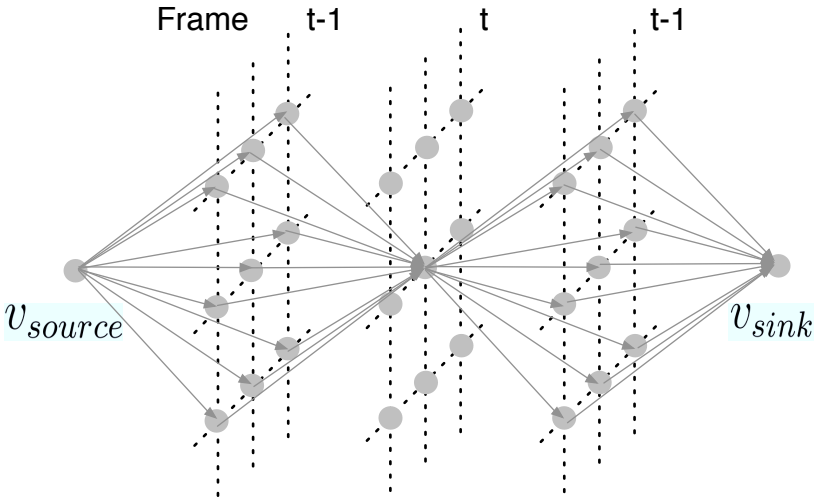


Figure 6.3: Illustration of constraints. They enforce continuous trajectories for each track by constraining each node can be passed by any path, and intersection is avoided by constraining each node can be only occupied by one object at the same time, the track can also be initialized and terminated automatically introducing source and sink nodes.

Using the network flow formalism we can easily cast it to an Integer Linear Programming(ILP) problem, with an objective function that is linearized with respect to a set of binary path (or flow) variables $p_t^{i,j} \in \{0, 1\}$, which indicate if a path is within part of some trajectory or not, as mentioned earlier. Then the proposed grid-based network flow model can be expressed as an ILP with the following objective function, by minimizing the total cost,

$$\begin{aligned} X^* &= \arg\min_X C^T p \\ &= \arg\min_X \sum_i c(v_{source}, i) p^{v_{source},i} + \sum_{i,j,t} c(i,j) p_t^{i,j} \\ &\quad + \sum_i c(i, v_{sink}) p^{i,v_{sink}} \ , \end{aligned} \tag{6.7}$$

in which the corresponding cost function C will be described in more details in Chapter 6.4.3.

Minimizing the criterion of (6.7) under the constraints of (6.2) to (6.4) can be rewrote as follows,

$$
\begin{aligned}
\text{minimize} \quad & C^T p \\
\text{subject to} \quad & \forall t, j, \sum_{i:j\in\mathcal{N}(i)} p_t^{i,j} = \sum_{k\in\mathcal{N}(j)} p_{t+1}^{j,k} \\
& \forall t, j, \sum_{k\in\mathcal{N}(j)} p_t^{j,k} \leq 1 \\
& \sum_{j\in\mathcal{N}(v_{source})} p^{v_{source},j} = \sum_{k:v_{sink}\in\mathcal{N}(k)} p^{k,v_{sink}} \\
& \forall t, i, j,\ p_t^{i,j} \in \{0,1\}\ .
\end{aligned}
\tag{6.8}
$$

Since Integer Linear Programming is NP-complete, we relax the condition $p_t^{i,j} \in \{0,1\}$ to $0 \leq p_t^{i,j} \leq 1$, resulting in a significant complexity reduction, and the relaxed formulation can be sufficiently solved with the simplex or interior-point method. The LP results then, are no longer guaranteed to be integer. However, we find in the experiments that the results are in most cases round integral, therefore giving a globally optimized solution.

6.4.3 Association Affinity Model

The details on the association affinity model are provided, which incorporate all meaningful features, including the measurements on behavior cue, as well as location and appearance in a global manner. With the set of observations $\Re$, we extract the features respect to location, color appearance, human body orientation, as illustrated in feature extraction module in Fig. 6.2. Location feature of each observation is represented as a grey node, its corresponding color measurement is represented as a colored one, while the arrow indicates the orientation cue.

Therefore, the transition probability term for each path $p^{i,j}$ leaving from node i to node j, is according to,

$$
A(i,j) = \begin{cases} A_{pos}(i,j)\cdot A_{col}(i,j)\cdot A_{ori}(i,j), & if\ t_j - t_i = 1, j \in \mathcal{N}(i) \\ 0, & otherwise \end{cases}, \tag{6.9}
$$

which is a product of these three affinities $A_{pos}(i,j)$, $A_{col}(i,j)$, $A_{ori}(i,j)$, respectively are location, appearance, orientation affinity between nodes i and j. To minimize the total cost according to (6.8), the corresponding cost is given by $c(i,j) = -log(A(i,j))$. The higher the affinity is, the more negative the cost of the edge, subsequently, confident tracks are likely to be in the path of the flow in order to minimize the total cost. And after taking *log* of the product of three affinities, the cost $c(i,j)$ becomes a weighted sum of them, thus these three affinities can be compensated with each other if any of these three features becomes ambiguous.

To be more detailed on each affinity term, the location affinity term $A_{pos}(i,j)$ concerns the spatial distances between two detection responses within two consecutive time steps,

$$A_{pos}(i,j) = exp(-\frac{||l_i - l_j||}{\sigma_l^2}) \ . \tag{6.10}$$

Note that the detection responses are on the 3D ground plane, not in 2D image plane. And the absolute spatial location difference is a L_1 norm.

For the color appearance term, histograms in CIE-Lab color space are employed for better characterizing the color content, which has the advantage of being perceptually uniform. In particular, $64 \times 64 \times 64$ color histograms are extracted from foreground images according to the detection responses. It is worth noting that the foreground images are obtained through utilizing a GPU based foreground/background segmentation approach proposed by Griesser et al. [215].

To compare the color feature similarity, Bhattacharyya distance measure is utilized because of its good classification property, allowing the combination of different features in a straightforward way. The similarity is multiplied through all views and assigned to corresponding path between nodes i and j.

$$A_{col}(i,j) = \prod_{n_v} exp(-\frac{d_B(a_i,a_j)}{\sigma_a^2}) \ , \tag{6.11}$$

where d_B is the Bhattacharyya distance between color feature a_i and a_j.

Finally, we integrate the crucial body orientation term, and the affinity is simply defined by difference between the two consecutive orientations,

$$A_{ori}(i,j) = exp(-\frac{0.5 * (1 - \cos(|\theta_i - \theta_j|)}{\sigma_\theta^2}) \ . \tag{6.12}$$

Note that the orientation θ_i and θ_j are computed in 3D space, being defined as the rotation with the axis perpendicular to the ground plane. The form of $0.5 * (1 - \cos(|\theta_i - \theta_j|))$ makes the orientation difference lie in the interval of $[0, 1]$.

As already emphasized before, the body orientation cue provides hints for resolving ambiguities between crossing trajectories, being discriminative enough even if crossed targets have similar appearance or move very slowly. The significant advantage of adding the orientation affinity term, is that more accurate trajectories can be estimated in case of close interaction or mutual occlusion, which will be demonstrated in our experiments in Chapter 6.5.

6.5 Experimental Results

This section aims to show the demonstrative results of our proposed approach. We evaluate the algorithm through a large variety of pre-recorded video sequences from our own dataset and public dataset as mentioned in Chapter 3.2, that involving both indoor and outdoor scenarios, with multiple people entering and leaving the scene, as well as closely interacting with each other for long time, or be seriously occluded by others.

6.5.1 Implementation Details

In order to keep a tractable complexity we separate long sequences into several batches, each one including 50 frames, processed one after the other. The ILP problem is solved by using the IBM ILOG CPLEX Optimizer [1]. This optimizer is able to solve very large linear programming problems using either primal or dual variants of the simplex method of the barrier interior point method.

At the same time, it is important to ensure consistency across batches, therefore we include the last frame of the previous batch into the current one, not only considering the detected locations but also the flows. This can be implemented as an additional constraint into the ILP problem (6.8),

$$\sum_{j \in \mathcal{N}(i)} p_{-1}^{i,j} = m_i, \ \forall (i,j). \tag{6.13}$$

As described in Chapter 4.4.1, the space has been discretized into $4N_g \times 4N_g$ grids on the finest level, with each node of the grid at time t connecting to its 9-neighborhood at time $t+1$ (8 neighbors plus the central location itself), resulting in $144N_g^2$ flows between consecutive frames. We define the transition cost $c(i,j)$ to be 0 if there is no observation on node i, that significantly reduces the size of graph and decreases the computational cost. Note, however, that there are approximately 100 observations in each frame, since we avoid non-maxima suppression during detection: the redundant observations greatly help for preserving the tracks during heavy occlusion and long-term interaction.

We also pay special attention to check the resulting, continuous values of the variables after linear programming relaxation. By processing a batch of 50 frames with a fine-grid discretization of 40×40, which results in 720,000 variables, we noticed that $719,800$ of them were in the range $[0, 0.01]$, and 200 in the range $[0.99, 1]$. Therefore, the relaxed linear programming is able to give an almost globally optimal solution to the original, integer-valued problem.

[1] `http://www.ibm.com/software/integration/optimization/cplex-optimizer`

6.5.2 Tracking Performance

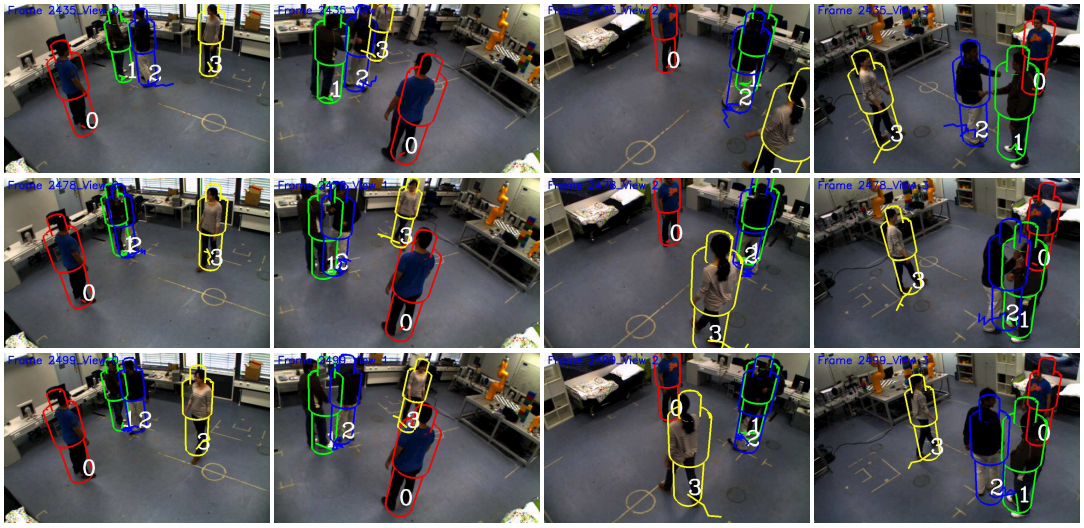


Figure 6.4: Tracking performance of our proposed approach on our own laboratory dataset with 4 targets involved, aiming to test the performance under the case of long term interaction. Every row shows a diffent frame, while every colomn displays diffent camera view.

We provide first a qualitative evaluation of results for each test sequence. Quantitative results will be shown and discussed in the following section.

Laboratory 4 Targets - Interaction This sequence consists of 3160 frames, in which the objects have interaction for long time. This scenario is aiming at evaluating the ability of our approach for dealing with long-term interaction and still-standing case, especially to verify the validity of the affinity term on behavior cue. Fig. 6.4 illustrates some sample frames between frame 2435 and frame 2499, target 1 and target 2 get extremely close and interact with each other across several frames, even that they almost stand still over time and their clothes are quite similar. Nevertheless, our hybrid orientation estimation scheme efficiently gives out the orientation of each target, consequently their corresponding opposite body orientations provide a powerful compensation in this ambiguous case, target 1 and 2 successfully maintain their own identity throughout the interaction.

Laboratory 4 Targets - Crossing The second sequence consists of 1800 frames. Within this sequence, most of the targets are wearing very dark clothing, with ambiguous appearance compared to each other. The objective of this case is to evaluate the capability of how the three affinity terms compensate with each other if any of the three features becomes ambiguous. As shown in Fig. 6.5, we can see the challenges due to targets that are occluded by each other

Figure 6.5: Tracking performance of our proposed approach on our own laboratory dataset with 4 targets involved, aiming to test the performance when target walking across others with very close proximity. Every row shows a diffent frame, while every colomn displays diffent camera view.

from one or two views. Sample frame 923 and frame 970 clearly illustrates a process that target 1 is trying to pass between target 0 and 3, their spatial locations are very close, however the distinguishing appearance of target 1 helps itself successfully maintain its identity during this crossing, as well as owing to its different orientation compared to other two targets. Conversely at frame 1055, target 0 and target 3 have close interaction while wearing extreme similar dark clothing, however their distinctive orientation provides efficient hint to solve the ambiguity despite of the similar appearance.

Laboratory 6 Targets An even more challenging sequence is recorded from our laboratory, which consists of 1520 frames, involving 6 targets in small observing area, therefore being with high density and heavy occlusion, so that to test the ability of our approach to cope with crowded environment. Example snapshots from the resulting tracks are shown in Fig. 6.6, clearly, the tracks for all targets are successfully obtained, even though they are densely located and some of them are occluded by others. More importantly, as illustrated from the tracking results, although the high density and severe occlusion, the identities of corresponding tracks are maintained very well over significant time intervals.

EPFL - Campus Sequence EPFL-Campus consists of around 5800 frames, which is recorded with outdoor scenario, within which the sunlight condition dynamically changes and there are shadows casted by moving people, furthermore, some targets run with fast speed within the

Figure 6.6: Tracking performance of our proposed approach on our own laboratory dataset with 6 targets involved, aiming to test the performance under crowded environment. Every row shows a diffent frame, while every colomn displays diffent camera view.

scene. Towards this sequence, we aim to evaluate whether our approach is able to handle the case of illumination changes, shadows and fast motion. Fig. 6.7 illustrates sample results of the tracking results. Despite the illumination changes and shadows, our approach correctly tracking targets throughout the sequence. In particular, as we can see from Frame 1706 apparently, target 7 runs along with target 6 with a very high speed, nevertheless it is successfully handled while their corresponding identity is well maintained. Note that within this sequence, the targets frequently leave and reenter the observing area, then they would be assigned a new unique ID everytime, therefore the identity changes for the same target across frames as we see from the sample frames.

EPFL - Terrace This sequence contains 5000 frames, up to 9 targets walking freely within a small outdoor area, which features a much more challenging outdoor scenario including a large number of occlusions, interactions, significant scale changes as well as illumination changes. This scenario makes the tracking performance evaluation of our proposed approach more convincing and significative. As can be seen from Fig. 6.8, plenty of challenges are obviously existing, targets get very close to each other and most of them have similar dark clothing. There are respectively 6, 7 and 8 targets appeared in the sample frame 1634, 2290 and 2540, all the targets are successfully tracked in spite of the various challenges. Especially in frame 2540, there are 8 targets in such a small observing area, thus the scene gets densely crowed, such as the cluster of

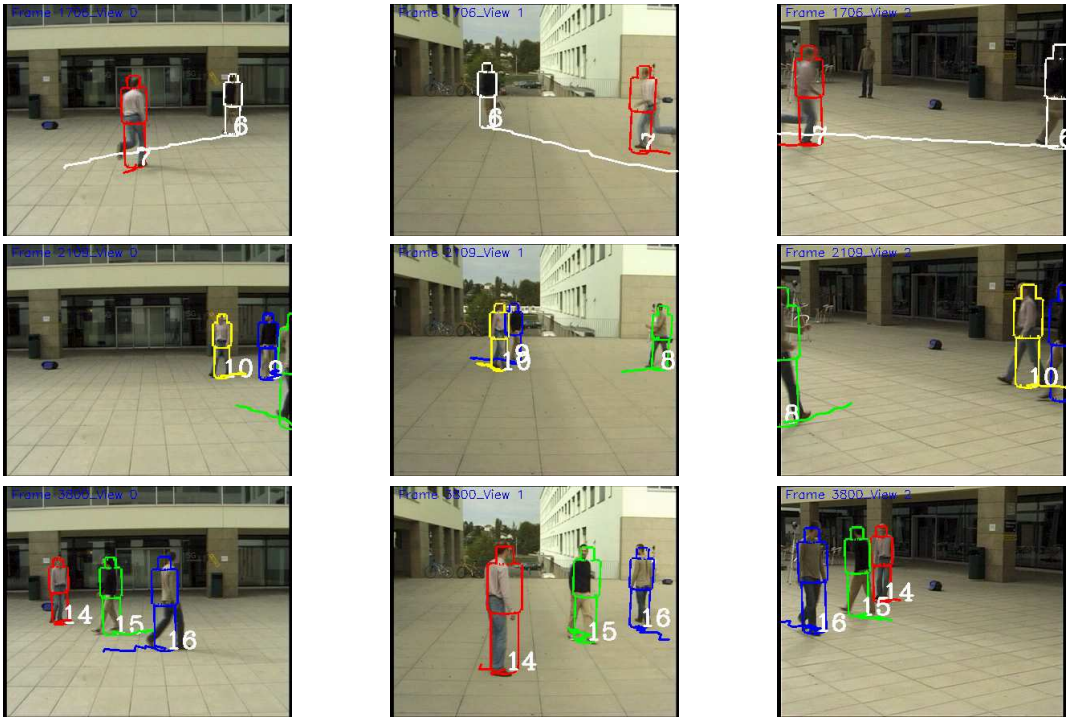


Figure 6.7: Tracking performance of our proposed approach on public dataset EPFL-Campus with outdoor scenario, aiming to evaluate the ability of our approach to handle the case of illumination changes, shadows and fast motion. For the first two datasets, every row shows a diffent frame, while every colomn displays diffent camera view.

target 10, 15, 17 and target 7, 16, they locate extremely close to each other. Nevertheless, they can be tracked very well. However we notice that the same target is assigned with different identity across the frames, in this sequence it is because of the frequent interobject occlusions, track-loss occurs sometimes throughout the long sequence, but it recovers very soon afterwards and assigns a new identity to the recovered track.

PETS2009 - S2L1 To further challenge the performance of our approach, monocular tracking is performed by using only one view out of the sequence PETS2009 - S2L1, from the VS-PETS 2009 benchmark dataset. This sequence shows an outdoor scene with up to 8 people walking freely and being occluded by each other for numerous times. Due to the single view this sequence poses additional challenges compared to previous ones. Firstly, targets frequently form together and split away. Secondly, it is totally ambiguous to distinguish targets when they are occluded by other tracking targets or traffic sign. Thirdly, the motion of some targets is very dynamic, since they are suddenly stopping, moving backward, or in circles. Fig. 6.9

Figure 6.8: Tracking performance of our proposed approach on public dataset EPFL-Terrace, which features a highly challenging outdoor scenario. We aim to evaluate the performance under large number of occlusions, interactions, significant scale changes and illumination changes. For the first two datasets, every row shows a diffent frame, while every colomn displays diffent camera view.

depicted 12 sample frames of the monocular tracking results. From the results we can explicitly see the challenges, in frame 71 target 1 is occluded by the traffic sign while the four targets in the middle are in close proximity; similarly, at frame 355 target 9 is almost fully occluded by target 10 and by the traffic sign, while at frame 481 target 9 is again occluded by target 13. Especially at frames 696 and 711, the scene is highly crowded. Our approach can handle most of the above cases, although a few mis-detections happen, e.g. at frames 71 and 123, due to a long-term, close proximity. Note that during frames from 514 to 696 the target walking on the lawn is not tracked, because it is out of the observation area.

6.5.3 Quantitative Evaluation

For a more fine-grained analysis, it is instructive to have a quantitative performance evaluation of our approach under our aforementioned unified performance evaluation framework, which include annotated ground-truth data and stand evaluation metrics.

Table 6.1 gives the quantitative results for all of the tracking metrics, which have been

Table 6.1: Quantitative performance evaluation results. This table shows the comparison results between our approach with different state-of-the-art approaches. The results are evaluated with various metrics as described in Chapter 3.3.2.

Dataset	Method	MOTA	MOTP	FN	FP	IDS	GT	MT	ML	FM
Laboratory 4 Targets - Interaction	Ours	**98.3**	77.5	**16**	**16**	**2**	4	**4**	**0**	**2**
	Chen 2012 [216]	87.3	**79.2**	26	215	12	4	4	0	5
	Berclaz 2009 [184]	95.5	74.6	22	64	4	4	4	0	3
Laboratory 4 Targets - Crossing	Ours	**98.3**	**79.7**	**9**	**9**	**16**	4	**4**	**0**	**2**
	Chen 2012 [216]	81.9	77.8	113	210	37	4	4	0	10
	Berclaz 2009 [184]	95.6	76.3	25	38	22	4	4	0	8
Laboratory 6 Targets	Ours	**91.2**	**82.7**	**35**	**24**	**27**	6	**6**	**0**	**3**
	Chen 2012 [216]	82.9	78.9	37	86	44	6	6	0	32
	Berclaz 2009 [184]	86.6	80.1	40	51	40	6	6	0	28
EPFL - Campus	Ours	**73.4**	**79.8**	**275**	**430**	**6**	24	**24**	**0**	**0**
	Chen 2012 [216]	60.5	74.4	316	782	53	24	22	2	35
	Berclaz 2009 [184]	66.2	77.6	290	593	21	24	24	0	10
EPFL - Terrace	Ours	70.9	**80.7**	364	670	**15**	9	7	1	**7**
	Berclaz 2011 [192]	**94.0**	77.1	-	-	-	9	-	-	-
	Andr. 2010 [186]	86.2	39.3	-	-	18	9	-	-	8
	Andr. no OM [217]	84.9	79.6	-	-	19	9	7	1	10
PETS - S2L1	Ours	71.1	66.9	566	255	53	23	20	2	27
	Berclaz 2011 [192]	82.3	52.1	641	126	**13**	23	17	2	22
	Andr. no OM [217]	81.4	76.1	262	**53**	16	23	**21**	**0**	**11**
	Andr. OM [165]	**88.6**	**76.9**	**171**	259	19	23	**21**	**0**	12

Figure 6.9: Tracking performance of our proposed approach on public dataset PETS-S2L1 with outdoor scenario. Due to the monocular view, frequent occlusion, grouping together and splitting away of the targets poses much more challenges. Sample frames on single view are shown.

detailedly explained in Chapter 3.3.2. We further compare our approach to different state-of-the-art approaches. In particular, on the first three our own datasets and EPFL-Campus dataset, we respectively compare with two baseline methods: Chen 2012 [216] and Berclaz 2009 [184]. On one hand, we mainly focus on demonstrating the improvement achieved by global optimal data association over classic data association, while using the same hierarchical grid-based detector input. Unsurprisingly, results on overall tracking accuracy (MOTA), precision (MOTP) as well as misdetections (FN), false alarms (FP), identity switches (IDS), mostly tracked (MT), mostly lost (ML), fragment (FM) significantly outperform the classic data association ones except the precision metric for dataset Laboratory 4 Targets - Interaction. The global optimal one efficiently reduces misdetections, false alarms and identity switches, therefore producing much smoother tracks and more consistent trajectories.

On the other hand, we compare with the baseline of standard Linear Programming approach in the work of [184], in order to show the effectiveness of our association affinity model which has incorporated the cues on location, appearance as well as body orientation. As can be seen from the evaluation results, our multi-feature based affinity model shows significant improvements over standard LP approach on all metrics. Especially thanks to the factor of explicitly taking

behavior cue - body orientation into account, so that efficiently decrease the ambiguities in crowded cases, therefore reduce the number of FN, FP and IDS, which in turn lead to higher accuracy score.

For dataset EPFL-Terrace, we instead compare to the tracking results of Berclaz 2011 [192], Anton 2010 [186] and Anton 2011 [217], the latter without occlusion modeling (denoted as Andr. no OM). Even though the MOTA score of our method does not achieve the highest value, the lowest IDS metric clearly indicates that our approach preserves identity much better than others due to including behavior cue. Note that a low number of ID switches is one of the most significant properties of a good tracking method. On the other hand, we notice that a relative higher FP value may lead to the lower MOTA score, this is due to the fact that we eschew non-maxima suppression during people detection, in oder to allow recovery from misdetections during global optimization for data association, false negatives are efficiently reduced by this procedure, nevertheless to some extent it leads to a slightly higher false alarm score.

Finally, for the widely used dataset PETS-S2L1, we compare ours with Berclaz 2011 [192], Anton no OM [217] and Anton OM [165], the latter again with occlusion modeling (denoted as Andr. OM). As it can be noticed, on this monocular dataset we have slight decreases of the performance. It can be explained by the fact that, comparing to multiple views, the performance of our hierarchical grid-based detector is influenced by only one view, being not able to provide the most reliable output in case of mutual occlusions (that can become almost total occlusions), thus producing more false negatives and positives that influence the final global optimized track output.

In addition, for the MOTA and MOTP value computed from our own approach, we again systematically vary the overlap ratio between the annotated box and tracking result, and obtain the corresponding varied MOTA and MOTP value, as shown in Fig. 6.10. Not surprisingly, we can see from the results that, as the overlap threshold decreases, MOTA increases and MOTP decreases, because more misaligned track results become classified as correct ones.

6.6 Conclusion

We have presented a global optimization framework for tracking a varying number of targets, the problem of multiple target tracking has been formulated with a grid-based network flow

model, then casted into Integer Linear Programming and solved through relaxation, achieving global optimality in most cases.

Our approach initiates, maintains and terminates tracks in a fully automatic way. Experiments over several benchmark sequences and quantitative comparisons with existing state-of-the-art approaches have been performed, demonstrating that our approach deals fairly well with mutual occlusions and long-term interactions that involve moving as well as still-standing people, while also revealing the importance and effectiveness of explicitly integrating 3D body orientation cue, as compared to previous approaches based on planar appearance or motion models.

The proposed methodology can be easily applied to different camera setups and different scenarios, and additional features can be easily included. In addition, we will consider to extend our framework to unstable illumination conditions, and apply the system to high-level scenarios, such as analysis of the trajectories, human-robot interaction, as well as autonomous robot navigation.

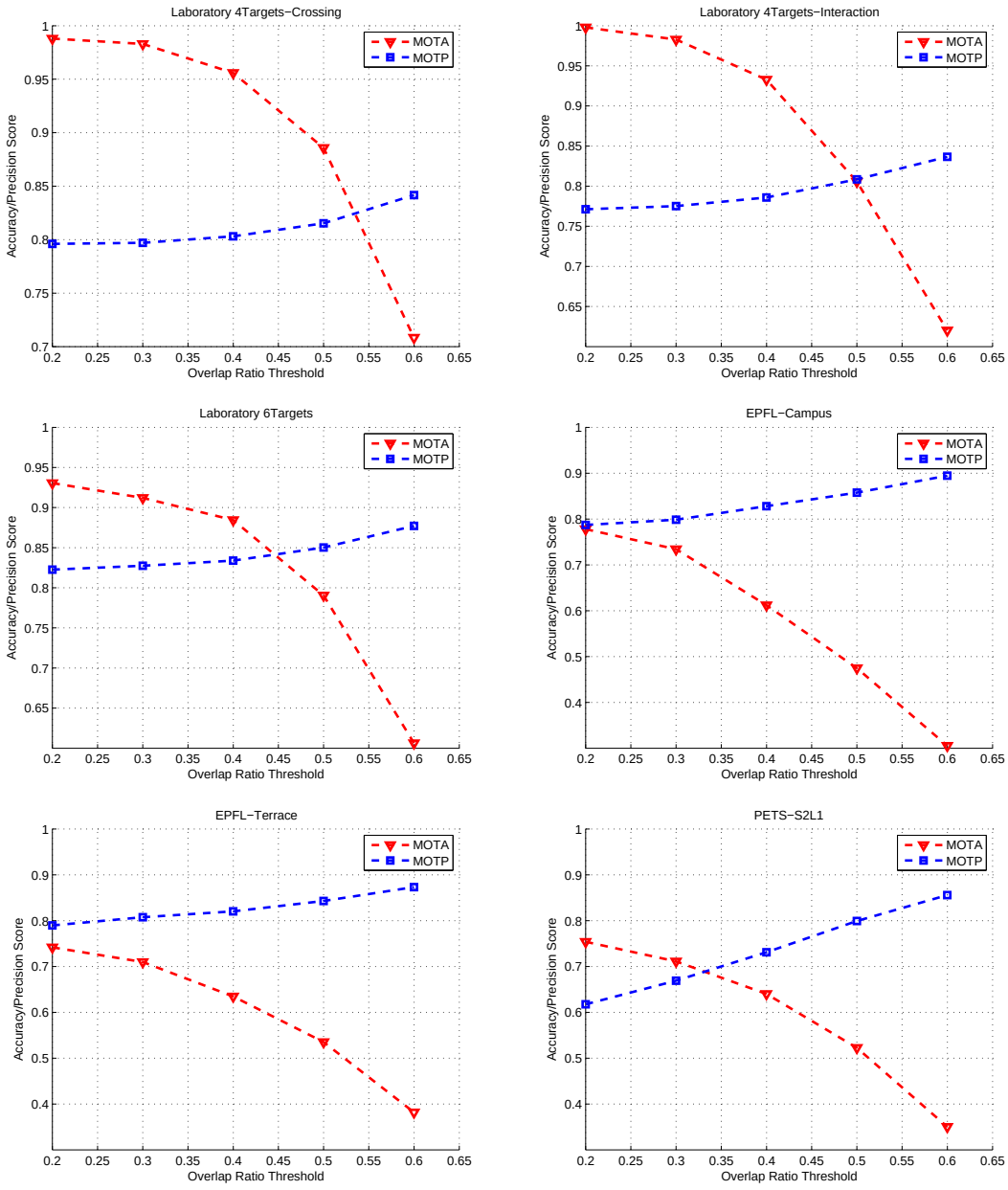


Figure 6.10: Influence of overlap threshold level on the evaluation results (MOTA and MOTP) for all of the sequences.

Chapter 7

Conclusion and Future Work

In this final chapter, we will summarize the main contributions of this thesis and thereby concluded the work. According to these observations, possible improvements and some interesting directions for future research are pointed out.

7.1 Summary

In this thesis, we have presented an unified hierarchical tracking-by-detection framework, for long-term detection and tracking of an a-priori unknown number of people with global optimization in complex and crowded environment, particularly the problem of complex interactions and heavy mutual occlusions for long period are well addressed. We further consider the human behavior as a higher-level reasoning evidence to improve the performance of data association based tracking. This work presented in this thesis contributes to the research and application areas of multiple people detection, human body orientation estimation as well as multiple people tracking. More specifically,

People Detection We first built a robust hierarchical grid-based people detector, which works on a frame-by-frame basis and merges information from multiple views, to produce an accurate localization on the ground plane. In particular, we relied on a *segment-then-locate* scheme, that detecting people by firstly obtaining foreground masks computed in multiple views, unlike mostly appearance-based background subtraction methods, we proposed an edge-based background subtraction for foreground segmentation, which has been demonstrated quite robust to environmental illumination changes. We also have constructed a template hierarchy, that matching with the foreground edge masks with a new oriented distance transform, by taking not only the location of edge points into account but also the gradient orientation. We showed

that this technique can significantly reduce the rate of false alarms in cluttered environment. Our proposed hierarchical grid-based detection methodology has been demonstrated that being able to yield nice results for detecting and localizing targets with no prior knowledge and deal fairly well with the various challenges such as mutual occlusions, illumination changes, fast motion, cluttered environment, and so on.

Human Body Orientation Estimation We studied part of the behavior analysis problem, by estimating the most representative cue - body orientation, which provides the direct evidence of person's potential behavior, that is, what the person is probably going to do and where the person is looking at. We explored this specific cue which is seldom considered by others, to improve tracking performance by disambiguating complex scenarios such as long-term interactions and heavy mutual occlusions when general used features (e.g. appearance, motion) become unreliable. The body orientation has been estimated by a hybrid 3D human body orientation estimation approach, dynamically combining the merits of a motion-based orientation estimator with a 3D appearance model-based orientation estimator. This approach has demonstrated a good robustness in various cases such as long-term interaction, mutual occlusion, slowly moving or even still-standing, etc.

Multiple People Tracking We have presented a global optimization framework for tracking a varying number of targets, by modeling the tracking problem with a grid-based network flow. And we formulate it in terms of finding the global maximum of a convex objective function, then cast into an Integer Linear Programming (ILP), leading to output with identity associated to corresponding target. It is demonstrated that global optimization is well suited for linking detections from a people detector at individual frames. Particularly, we have shown that how the body orientation can be efficiently integrated into the data association problem, and how it provides its capability on revolving ambiguities between crossing trajectories. We have also validated the tracking framework through multiple experiments over various challenging sequences, including both our self-recorded datasets and public benchmark datasets, demonstrating that our approach is able to deal fairly well with mutual occlusions and long-term interactions that involve moving as well as still-standing people, and the importance and effectiveness of explicitly integrating 3D body orientation cue has been revealed. Comparing to previous affinity models which consider appearance or motion cue, this new affinity model efficiently decreases the ambiguities in crowded cases.

Performance Evaluation Framework This work has proposed a novel evaluation framework, to quantitatively evaluate the performance of the entire tracking system. For this, we

have annotated the ground truth data in 3D space for each video sequence, so that analyzing the strengths and weaknesses of the system with groundtruth data. Furthermore, the evaluations are based on standard metrics, such as Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP), Precision, Recall, Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), False Negatives (FN), False Positives (FP), Identity Switches (IDS), giving out an intuitive measure of the detector and tracker's performance at detecting objects, localizing objects, keeping their trajectories, and so on. More importantly, with these standard evaluation metrics, it provides a much easier way to compare our system to state-of-the-art, so that clearly indicating respective strengths and weaknesses.

7.2 Future Work

This thesis has contributed to related areas of people detection, tracking, human body orientation estimation, and there are several future research directions that have spawned, including future improvements and potential extensions.

On one hand, the future improvements aim to improve the performance of the entire system in certain aspects where it is currently lagging. Among the main ones is the improvement of the hierarchical grid-based detector, it could be optimized by utilizing the full power of the latest GPUs in order to obtain performance boost, and in order to better handle the outdoor scenarios, the edge-based background subtraction algorithm, as one key for the final performance of the detector, should also be improved in the future. Possible ideas that include processing binary images with, for instance, shape analysis, so that to remove noises that not belong to peoples. And also a part-based people detector, that does not look for a whole human body but searches for isolated body parts would also be preferred.

Another potential improvement of our work deals with the human body orientation estimation module, the estimation results could be improved by using more advanced inference approaches such as particle filters, that instead of maximum likehood approach for selecting the body orientation in 3D appearance-based estimation method. Additionally, within the motion-based estimation method, the motion vector can rely on not only two consecutive frames, but also a tracklet consists of several frames, which would be helpful to achieve more robust results.

In the context of our approach of people tracking based on global optimization, on one hand, we plan to work on the optimization itself in order to further improve the optimization process that better utilizing current affinity terms. On the other hand, an important direction

of future studies will be to not only utilize the body orientation output from the human body orientation estimation module for the people tracking module, but also incorporate the built 3D appearance model to make full use of this module, so that better disambiguating tracks and keeping identities, further improving the tracking performance. Furthermore, for the computation cost of the optimization part, there is still space to improve its efficiency by exploiting graphics hardware. And our work can be easily applied to include additional features, such as color or motion, also can be scaled to different camera views, as well as being used for tracking different objects, for example 3D tracking of flying *quadrotors*.

Besides these straightforward improvements, we also plan to test and extend our system to more challenging scenarios, as well as people tracking on mobile robots, with a non-static background and viewpoint. To further enhance its applicability, future extension could be aimed at more advanced technologies (i.e. to recognize the actions of person, as they need to be tracked first), and can be used to an aid to enhance the utility of others (i.e. crowd analysis may be able to detect people moving against the flow of traffic, which can then be tracked using people tracking techniques). An interesting direction of research would be to detect and track groups in crowd, the output of our human body orientation estimation module could provide reliable information for group formulation if neighboring people share the same orientation. The tracking results can also be used in the analysis of group or crowd activities. People in crowds tend to show peculiar patterns of collective behavior like, gathering, scattering, clique behavior, marching, milling, and so on. It would be interesting to develop crowd behavior analysis and use the tracking results to detect and recognize normal and abnormal activities. Our complete multi-view people detection and tracking framework lends itself very well to this behavior analysis task. The ground plane localization can be provided as the input data for a system whose task is to automatically analyzing the trajectories and potentially recognizing the behavior patterns. By analyzing the behaviors of people, we could determine whether their behaviors are normal or abnormal, so that provide high-level description of actions and interactions between or among people.

In conclusion, our system could be extended in a variety of ways, making it an exciting and rewarding concept to be involved in.

References

[1] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. **Pfinder, Real Time Tracking of the Human Body**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **19**(7):780–785, 1997. 1

[2] T. Olson and F. Brill. **Moving Object Detection and Event Recognition Algorithms for Smart Cameras**. In *Proceedings of DARPA Image Understanding Workshop*, pages 159–175, 1997. 1

[3] A. J. Lipton, H. Fujiyoshi, and R. S. Patil. **Moving Target Classification and Tracking from Real-Time Video**. In *Proceedings of IEEE Workshop Applications of Computer Vision*, pages 8–14, 1998. 1

[4] I. Haritaoglu, D. Harwood, and L. Davis. **W4: Real Time Surveillance of People and Their Activities**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **22**(8):809–830, 2000. 1

[5] R. T. Collins, A. J. Lipton, T. Kanade, D. Fujiyoshi, Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, and L. Wixson. **A System for Video Surveillance and Monitoring**. Technical report, Carnegie Mellon University, 2000. 1

[6] N. T. Siebel and S. Maybank. **The ADVISOR Visual Surveillance System**. In *Proceedings of the ECCV 2004 Workshop on Applications of Computer Vision*, pages 103–111, 2004. 1

[7] X. Liu, P. H. Tu, J. Rittscher, A. Perera, and N. Krahnstoever. **Detecting and Counting People in Surveillance Applications**. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 306–311, 2005. 1

[8] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. **Integrated Person Tracking Using Stereo, Color and Pattern Detection**. *International Journal of Computer Vision*, **37**(2):175–185, 2000. 1

[9] K. Nickel and R. Stiefelhagen. **Real-time Person Tracking and Pointing Gesture Recognition for Human-Robot Interaction**. In *ECCV 2004 Workshop on Human-Computer Interaction*, pages 28–38, 2004. 1

[10] R. Munoz-Salinas, E. Aguirre, and Garcia-Silvente M. **People Detection and Tracking using Stereo Vision and Color**. *Image and Vision Computing*, **25**(6):997–1007, 2007. 1

[11] S. Nair, G. Panin, T. Roeder, T. Friedelhuber, and A. Knoll. **A Distributed and Scalable Person Tracking System for Robotic Visual Servoing with 8 DOF in Virtual Reality TV Studio Automation**. In *Proceedings of the 6th International Symposium on Mechatronics and its Applications (ISMA09)*, pages 1–7, 2009. 1

[12] Y. Kobayashi and Y. Kuno. **People Tracking using Integrated Sensors for Human Robot Interaction**. In *IEEE International Conference on Industrial Technology*, pages 1617–1622, 2010. 1

[13] A. Ekin, A. M. Tekalp, and R. Mehrotra. **Automatic Soccer Video Analysis and Summarization**. *IEEE Transaction on Image Processing*, **12**(7):796–807, 2003. 1

[14] P. Figueroa, N. Leite, R. M. L. Barros, I. Cohen, and G. Medioni. **Tracking Soccer Players using the Graph Representation**. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 787–790, 2004. 1

[15] M. Beetz, S. Gedikli, J. Bandouch, B. Kirchlechner, N. Hoyningen-Huene, and A. Perzylo. **Visual Tracking Football Games based on TV Broadcasts**. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2066–2071, 2007. 1

[16] S. Gedikli, J. Bandouch, N. Hoyningen-Huene, B. Kirchlechner, and M Beetz. **An Adaptive Vision System for Tracking Soccer Players from Variable Camera Settings**. In *Proceedings of International Conference on Computer Vision Systems*, 2007. 1

[17] N. Gengembre and P. Perez. **Probabilistic Color-based Multi-Object Tracking with Application to Team Sports**. Technical report, INRIA, 2008. 1

[18] D. M. Gavrila. **Pedestrian Detection from a Moving Vehicle**. In *Proc. of European Conference on Computer Vision*, pages 37–49, Dublin, Ireland, 2000. 1, 12, 40, 42, 48

[19] D. M. Gavrila and S. Munder. **Multi-Cue Pedestrian Detection and Tracking from a Moving Vehicle**. *International Journal on Computer Vision*, **73**(1):41–59, 2007. 1, 12

[20] M. Bertozzi, L. Bombini, A. Broggi, P. Cerri, P. Grisleri, and P. Zani. **GOLD: A Framework for Developing Intelligent-Vehicle Vision Applications**. *IEEE Intelligent Systems*, **23**(1):69–71, 2008. 1

[21] D. Geronimo, A. Lopez, A. Sappa, and T. Graf. **Survey of Pedestrian Detection for Advanced Driver Assistance Systems**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **32**(7):1239–1258, 2010. 1

[22] Y. Xu, S. Xu, S. Lin, and T. X. Han. **Detection of Sudden Pedestrian Crossings for Driving Assistance Systems**. *IEEE Transactions on Systems, Man, and Cybernetics*, **42**(3):729–739, 2000. 1

[23] A. Prioletti, A. Mogelmose, P. Grisleri, and M. M. Trivedi. **Part-Based Pedestrian Detection and Feature-Based Tracking for Driver Assistance: Real-Time, Robust Algorithms, and Evaluation**. *IEEE Transactions on Intelligent Transportation Systems*, **14**(3):1346–1359, 2013. 1

[24] X. Wang, X. Han, and S. Yan. **An HOG-LBP Human Detector with Partial Occlusion Handling**. In *IEEE International conference on Computer Vision*, pages 32–39, 2009. 11

[25] S. Walk, N. Majer, K. Schindler, and B. Schiele. **New Features and Insights for Pedestrian Detection**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1030–1037, 2010. 11

[26] J. Marin, D. Vazquez, D. Geronimo, and A. M. Lepoz. **Learning Appearance in Virtual Scenarios for Pedestrian Detection**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 137–144, 2010. 11

[27] G. Duan, H. Ai, and S. Lao. **A Structural Filter Approach to Human Detection**. In *11th European Conference on Computer Vision*, pages 238–251, 2010. 11

[28] Y. Liu, S. Shan, X. Chen, J. Heikkila, W. Gao, and M. Pietikainen. **Spatial-Temporal Granularity-Tunable Gradients Partition (STGGP) Descriptors for Human Detection**. In *11th European Conference on Computer Vision*, pages 327–340, 2010. 11

[29] Y. Zheng, C. Shen, R. Hartley, and X. Huang. **Pyramid Center-Symmetric Local Binary/Trinary Patterns for Effective Pedestrian Detection**. In *10th Asian Conference on Computer Vision*, pages 281–292, 2011. 11

[30] W. Ouyang and X. Wang. **A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3258–3265, 2012. 11

[31] C. Papageorgiou and T. Poggio. **A Trainable System for Object Detection**. *International Journal of Computer Vision*, **38**(1):15–33, 2000. 11, 13, 14

[32] P. Viola, M. Jones, and D. Snow. **Detecting Pedestrians using Patterns of Motion and Appearance**. In *International Conference on Computer Vision*, pages 734–741, 2003. 11, 13, 14

[33] N. Dalal and B. Triggs. **Histograms of Oriented Gradients for Human Detection**. In *IEEE conference on Computer Vision and Pattern Recognition(CVPR)*, pages 886–893, 2005. 11, 13, 14

[34] B. Leibe, E. Seemann, and B. Schiele. **Pedestrian Detection in Crowded Scenes**. In *IEEE conference on Computer Vision and Pattern Recognition(CVPR)*, pages 878–885, 2005. 11, 13

[35] M. Andriluka, S. Roth, and B. Schiele. **People-Tracking-by-Ddetection and People-Detection-by-Tracking**. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 11, 20

[36] C. Wojek, G. Dorkó, A. Shulz, and B. Schiele. **Sliding-Windows for Rapid Object Class Localization: A Parallel Technique**. In *DAGM-Symposium*, pages 71–81, 2008. 11

[37] C. Wojek, S. Walk, and B. Schiele. **Multi-Cue Onboard Pedestrian Detection**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 794–801, 2009. 11

[38] J. Berclaz, F. Fleuret, and P. Fua. **Robust People Tracking with Global Trajectory Optimization**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 744–750, 2006. 11, 15, 20, 21

[39] A. Yuille. **Deformable Templates for Face Recognition**. *Journal of Cognitive Neuroscience*, **3**(1):59–70, 1991. 12

[40] A. Mohan, C. Papageorgiou, and T. Poggio. **Example-Based Object Detection in Images by Components**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **23**(4):349–361, 2001. 12, 13, 14

[41] Y. Yang, G. Shu, and M. Shah. **Semi-supervised Learning of Feature Hierarchies for Object Detection in a Video**. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1650–1657, 2013. 12

[42] M. Enzweiler and D. M. Gavrila. **Monocular Pedestrian Detection: Survey and Experiments**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **31**(12):2179–2195, 2009. 12, 13, 14

[43] D. M. Gavrila. **A Bayesian, Exemplar-based Approach to Hierarchical Shape Matching**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **29**(8):1408–1421, 2007. 12

[44] B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla. **Model-based Hand Tracking using a Hierarchical Bayesian Filter**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**:1372–1384, 2006. 12, 42, 52

[45] K. Toyama and A. Blake. **Probabilistic Tracking with Exemplars in a Metric Space**. *International Journal on Computer Vision*, **48**(1):9–19, 2002. 12

[46] D. M. Gavrila. **Real-time Object Detection for Smart Vehicles**. In *IEEE International Conference on Computer Vision*, pages 87–93, 1999. 12

[47] T. Heap and D. Hogg. **Improving Specificity in PDMs using a Hierarchical Approach**. In *Proceedings of British Machine Vision Conference*, pages 80–89, 1997. 13

[48] T. Heap and D. Hogg. **Wormholes in Shape Space: Tracking through Discontinuous Changes in Shape**. In *Proceedings of International Conference on Computer Vision*, pages 344–349, 1998. 13

[49] A. Baumberg. **Hierarchical Shape Fitting Using and Iterated Linear Filter**. In *Proceedings of British Machine Vision Conference*, pages 313–323, 1996. 13

[50] M. J. Jones and T. Poggio. **Multidimensional Morphable Models**. In *Proceedings of International Conference on Computer Vision*, pages 683–688, 1998. 13

[51] M. Enzweiler and D. M. Gavrila. **A Mixed Generative-Discriminative Framework for Pedestrian Classification**. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 13

[52] T. F. Cootes and C. J. Taylor. **Statistical Models of Appearance for Computer Vision**. Technical report, University of Manchester, 2004. 13

[53] S. Munder, C. Schnoerr, and D. M. Gavrila. **Pedestrian Detection and Tracking Using a Mixture of View-Based Shape-Texture Models**. *IEEE Transactions Intelligent Transportation Systems*, **9**(2):333–343, 2008. 13

[54] A. Yilmaz, O. Javed, and M. Shah. **Object Tracking: A Survey**. *ACM Computing Surveys*, **38**(4):1–45, 2006. 13

[55] B. Schiele, M. Andriluka, N. Majer, S. Roth, and C. Wojek. **Visual People Detection: Different Models, Comparison and Discussion**. In *Proceedings of the IEEE ICRA 2009 Workshop on People Detection and Tracking*, pages 1–8, 2009. 13

[56] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. **Pedestrian Detection Using Wavelet Templates**. In *International Conference on Computer Vision and Pattern Recognition*, pages 193–199, 1997. 13

[57] P. Dollar, C. Wojek, B. Schiele, and P. Perona. **Pedestrian Detection: A Benchmark**. In *International Conference on Computer Vision and Pattern Recognition*, pages 304–311, 2009. 13

[58] S. Belongie, J. Malik, and J. Puzicha. **Shape Matching and Object Recognition using Shape Contexts**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **24**(4):509–522, 2002. 13

[59] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla. **Shape Context and Chamfer Matching in Cluttered Scenes**. In *International Conference on Computer Vision and Pattern Recognition*, pages 127–133, 2003. 13

[60] G. Mori, S. Belongie, and J. Malik. **Efficient Shape Matching using Shape Contexts**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **27**(11):1832–1837, 2005. 13

[61] B. Leibe, A. Leonardis, and B. Schiele. **Combined Object Categorization and Segmentation with an Implicit Shape Model**. In *ECCV Workshop on Statistical Learning in Computer Vision*, pages 17–32, 2004. 13

[62] L. Spinello, R. Triebel, and R. Siegwart. **Multimodal People Detection and Tracking in Crowded Scenes**. In *Proceedings of The AAAI Conference on Artificial Intelligence*, 2008. 13

[63] K. Jungling and A. Michael. **Feature Based Person Detection Beyond the Visible Spectrum**. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 30–37, 2009. 13

[64] J. Gall, N. Razavi, and L. V. Gool. **On-line Adaption of Class-specific Codebooks for Instance Tracking**. In *Proceeding of British Machine Vision Conference*, 2010. 13

[65] Q. Zhu, M. C. Yeh, K. T. Cheng, and S. Avidan. **Fast Human Detection Using a Cascade of Histograms of Oriented Gradients**. In *International Conference on Computer Vision and Pattern Recognition*, pages 1491–1498, 2006. 13, 14

[66] F. Han, Y. Shan, R. Cekander, H. S. Sawhney, and R. Kumar. **A Two-Stage Approach to People and Vehicle Detection with HOG-Based SVM**. In *Performance Metrics for Intelligent Systems 2006 Workshop*, pages 133–140, 2006. 13

[67] X. Wang, T. X. Han, and S. Yan. **An HOG-LBP Human Detector with Partial Occlusion Handling**. In *IEEE International Conference on Computer Vision (ICCV)*, pages 32–39, 2009. 13

[68] L. Spinello and K. O. Arras. **People Detection in RGB-D Data**. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 3838–3843, 2011. 13

[69] B. Wu and R. Nevatia. **Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors**. *International Journal of Computer Vision*, **75**(2):247–266, 2007. 13, 14, 20

[70] P. Sabzmeydani and G. Mori. **Detecting Pedestrians by Learning Shapelet Features**. In *IEEE conference on Computer Vision and Pattern Recognition(CVPR)*, pages 1–8, 2007. 13

[71] M. Bertozzi, A. Broggi, M. D. Rose, M. Felisa, A. Rakotomamonjy, and F. Surad. **A Pedestrian Detector using Histograms of Oriented Gradients and a Support Vector Machine Classifer**. In *Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference*, pages 143–148, 2007. 13

[72] L. Ladicky, P. H. S. Torr, and A. Zisserman. **Latent SVMs for Human Detection with a Locally Affine Deformation Field**. In *Proceedings of the British Machine Vision Conference*, pages 10.1–10.11, 2012. 13

[73] P. Viola and M. Jones. **Rapid Object Detection Using a Boosted Cascade of Simple Features**. In *International Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001. 13

[74] K. Mikolajczyk, C. Schmid, and A. Zisserman. **Human Detection based on a Probabilistic Assembly of Robust Part Detectors**. In *European Conference on Computer Vision*, pages 69–82, 2004. 13

[75] M. Szarvas, A. Yoshizawa, M. Yamamoto, and J. Ogata. **Pedestrian Detection with Convolutional Neural Networks**. In *Intelligent Vehicles Symposium*, pages 224–229, 2005. 13

[76] J. Fan, W. Xu, Y. Wu, and Y. Gong. **Human Tracking Using Convolutional Neural Networks**. *IEEE Transaction on Neural Networks*, **21**(10):1610–1623, 2010. 13

[77] Z. Kira, R. Hadsell, G. Salgian, and S. Samarasekera. **Long-Range Pedestrian Detection using Stereo and a Cascade of Convolutional Network Classifiers**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2396–2403, 2012. 13

[78] A. Opelt, A. Pinz, and A. Zisserman. **A Boundary-Fragment-Model for Object Detection**. In *European Conference on Computer Vision*, pages 575–588, 2006. 13

[79] J. Shotton, J. Winn, C. Rother, and A. Criminisi. **TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Recognition and Segmentation**. In *European Conference on Computer Vision*, pages 1–15, 2006. 13

[80] T. P. Tian and S. Sclaroff. **Fast Globally Optimal 2D Human Detection with Loopy Graph Models**. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 81–88, 2010. 13

[81] S. Agarwal, A. Awan, and D. Roth. **Learning to Detect Objects in Images via a Sparse, Part-Based Representation**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **26**(11):1475–1490, 2004. 13

[82] E. Seemann, M. Fritz, and B. Schiele. **Towards Robust Pedestrian Detection in Crowded Image Sequences**. In *International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 13

[83] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. **Full-Body Recognition System**. *Pattern Recognition*, **36**:1997–2006, 2003. 14

[84] H. Shimizu and T. Poggio. **Direction Estimation of Pedestrian from Multiple Still Images**. In *IEEE Intelligent Vehicles Symposium*, pages 596–600, 2004. 14, 17

[85] N. Dalal, B. Triggs, and C. Schmid. **Human Detection Using Oriented Histograms of Flow and Appearance**. In *European Conference on Computer Vision*, pages 428–441, 2006. 14

[86] V. D. Shet, J. Neumann, V. Ramesh, and L. S. Davis. **Bilattice-Based Logical Reasoning for Human Detection**. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 14

[87] S. Kang, H. Byun, and S. W. Lee. **Real-Time Pedestrian Detection Using Support Vector Machines**. In *Proceedings of the First International Workshop on Pattern Recognition with Support Vector Machines*, pages 268–277, 2002. 14

[88] S. Munder and D. M. Gavrila. **An Experimental Study on Pedestrian Classification**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **28**(11):1863–1868, 2006. 14

[89] I. P. Alonso, D. F. Llorca, M. A. Sotelo, L. M. Bergasa, P. Revenga de Toro, J. Nuevo, M. Ocana, and M. A. G. Garrido. **Combination of Feature Extraction Methods for SVM Pedestrian Detection**. *IEEE Transactions on Intelligent Transportation Systems*, **8**(2):292–307, 2007. 14

[90] D. G. Lowe. **Distinctive Image Features from Scale Invariant Keypoints**. *International Journal on Computer Vision*, **60**(2):91–110, 2004. 14

[91] K. Okuma, A. Taleghani, N. De Freitas, J. Little, and D. Lowe. **A Boosted Particle Filter: Multitarget Detection and Tracking**. In *European Conference on Computer Vision*, 2004. 14, 20

[92] O. Tuzel, F. Porikli, and P. Meer. **Human Detection via Classification on Riemannian Manifolds**. In *International Conference on Computer Vision and Pattern Recognition*, pages 258–265, 2005. 14

[93] L. Zhang, B. Wu, and R. Nevatia. **Detection and Tracking of Multiple Humans with Extensive Pose Articulation**. In *International Conference on Computer Vision*, pages 1–8, 2007. 14

[94] K. O. Arras, O. M. Mozos, and W. Burgard. **Using Boosted Features for the Detection of People in 2D Range Data**. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 3402–3407, 2007. 14

[95] C. C. R. Wang and J. J. J. Lien. **AdaBoost Learning for Human Detection Based on Histograms of Oriented Gradients**. In *Proceedings of 8th Asian Conference on Computer Vision*, pages 885–895, 2007. 14

[96] L. Pishchulin, A. Jain, C. Wojek, T. Thormaehlen, and B. Schiele. **In Good Shape: Robust People Detection based on Appearance and Shape**. In *Proceedings of the British Machine Vision Conference*, pages 5.1–5.12, 2011. 14

[97] I. Mikic, S. Santini, and R. Jain. **Video Processing and Integration from Multiple Cameras**. In *Proceedings of the Image Understanding Workshop*, pages 183–187, 1998. 15

[98] D. Focken and R. Stiefelhagen. **Towards Vision-Based 3D People Tracking in a Smart Room**. In *IEEE International Conference on Multimodal Interfaces*, pages 400–405, 2002. 15

[99] K. Otsuka and N. Mukawa. **Multi-View Occlusion Analysis for Tracking Densely Populated Objects based on 2-D Visual Angles**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 90–97, 2004. 15

[100] S. M. Khan and M. Shah. **A Multiview Approach to Tracking People in Crowded Scenes Using a Planar Homograhy Constraint**. In *European Conference on Computer Vision*, pages 133–146, 2006. 15

[101] O. Lanz. **Approximate Bayesian Multibody Tracking**. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **28**(9):1436–1449, 2006. 15

[102] X. Song, J. Cui, H. Zha, and H. Zhao. **Vision-based Multiple Interacting Targets Tracking via On-Line Supervised Learning**. In *European Conference on Computer Vision*, pages 642–655, 2008. 15

[103] T. Zhao and R. Nevatia. **Tracking Multiple Humans in Complex Situations**. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **26**(9):1208–1221, 2004. 15

[104] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. **Multi-Camera Multi-Person Tracking for Easy Living**. In *IEEE International Workshop on Visual Surveillance*, pages 3–10, 2000. 15

[105] D. Baltieri, R. Vezzani, R. Cucchiara, and A. Utasi. **Multi-View People Surveillance using 3D Information**. In *Proceedings of 2011 IEEE International Conference on Computer Vision Workshops*, pages 1817–1824, 2011. 15

[106] M. C. Liem and D. M. Gavrila. **A Comparative Study on Multi-Person Tracking Using Overlapping Cameras**. In *The 9th International Conference on Computer Vision Systems*, pages 203–212, 2013. 15

[107] S. M. Khan and M. Shah. **A Multiview Approach to Tracking People in Crowded Scenes Using a Planar Homograhy Constraint**. In *European Conference on Computer Vision*, pages 133–146, 2006. 15

[108] S. M. Khan and M. Shah. **Tracking Multiple Occluding People by Localizing on Multiple Scene Planes**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **31**(3):505–519, 2009. 15

[109] D. Delannay, N. Danhier, and C. D. Vleeschouwer. **Detection and Recognition of Sports (Wo)Men from Multiple Views**. In *Proceedings of ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–7, 2009. 15

[110] R. Eshel and Y. Moses. **Homography Based Multiple Camera Detection and Tracking of People in a Dense Crowd**. In *International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 15

[111] B. Kwolek. **Multi Camera-Based Person Tracking Using Region Covariance and Homography Constraint**. In *2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 294–299, 2010. 15

[112] K. Kim and L. Davis. **Multi-Camera Tracking and Segmentation of Occluded People on Ground Plane using Search-Guided Particle Filtering**. In *European Conference on Computer Vision*, pages 98–109, 2006. 15

[113] A. Laurentini. **The Visual Hull Concept for Silhouette Based Image Understanding**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **16**(2):150–162, 1994. 15

[114] K. M. Cheung, T. Kanade, J. Y. Bouguet, and M. Holler. **A Real Time System for Robust 3D Voxel Reconstruction of Human Motions**. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 714–720, 2000. 15

[115] J. Franco and E. Boyer. **Fusion of Multi-View Silhoette Cues Using a Space Occupancy Grid**. In *Proceedings of International Conference on Computer Vision*, pages 1747–1753, 2005. 15

[116] F. Feuret, J. Berclaz, R. Lengagne, and P. Fua. **Multicamera people tracking with a probabistic occupancy map**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(2):267–282, February 2008. 15, 31

[117] A. Alahi, Y. Boursier, L. Jacques, and P. Vandergheynst. **A Sparsity Constrained Inverse Problem to Locate People in a Network of Cameras**. In *Proceedings of the 16th International Conference on Digital Signal Processing*, pages 22–28, 2009. 16

[118] R. Mandeljc, S. Kovačič, M. Kristan, and J. Perš. **Non-sequential Multi-view Detection, Localization and Identification of People Using Multi-modal Feature Maps**. In *Asian Conference on Computer Vision*, pages 691–704, 2013. 16

[119] D. B. Yang, H. H. González-Baños, and L. J. Guibas. **Counting People in Crowds with a Real-Time Network of Simple Image Sensors**. In *International Conference on Computer Vision*, pages 122–129, 2003. 16

[120] Á. Utasi and C. Benedek. **A 3-D Marked Point Process Model for Multi-View People Detection**. In *International Conference on Computer Vision and Pattern Recognition*, pages 3385–3392, 2011. 16

[121] Á. Utasi and C. Benedek. **A Bayesian Approach on People Localization in Multi-Camera Systems**. *IEEE Transactions on Circuits and Systems for Video Technology*, **23**(1):105–115, 2012. 16

[122] T. T. Santos and C. H. Morimoto. **Multiple Camera People Detection and Tracking using Support Integration**. *Pattern Recognition Letters*, **32**(1):47–55, 2011. 16

[123] T. Gandhi and M. M. Trivedi. **Image based Estimation of Pedestrian Orientation for Improving Path Prediction**. In *IEEE IV Symposium*, pages 506–511, 2008. 17

[124] K. Panachit, O. S. Guat, and E. How-Lung. **Estimation of Human Body Orientation using Histogram of Oriented Gradients**. In *IAPR Conference on Machine Vision Applications*, pages 459–462, 2011. 17

[125] C. Weinrich, C. Vollmer, and H. M. Gross. **Estimation of Human Upper Body Orientation for Mobile Robots using an SVM Decision Tree on Monocular Images**. In *IEEE Internation Conference on Intelligent Robots and Systems (IROS)*, pages 2147–2152, 2012. 17

[126] C. Chen, A. Heili, and J. Odobez. **Combined Estimation of Location and Body Pose in Surveillance Video**. In *IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 5–10, 2011. 17

[127] C. Chen and J. Odobez. **We are not Contortionists: Coupled Adaptive Learning for Head and Body Orientation Estimation in Surveillance Video**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1544–1551, 2012. 17

[128] M. Enzweiler and D. Gavrila. **Integrated Pedestrian Classification and Orientation Estimation**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 982–989, 2010. 17

[129] N. Robertson and I. Reid. **Estimating Gaze Direction from Low-Resolution Faces in Video**. In *Proceedings of 9th European Conference on Computer Vision*, pages 402–415, 2006. 17

[130] N. Krahnstoever, M. C. Chang, and W. Ge. **Gaze and Body Pose Estimation from a Distance**. In *IEEE International Conference on Advanced Video Signal based Surveillance*, pages 11–16, 2011. 17

[131] M. W. Lee and R. Nevatia. **Body Part Detection for Human Pose Estimation and Tracking**. In *IEEE Workshop on Motion and Video Computing*, page 23, 2007. 17

[132] O. Ozturk, T. Yamasaki, and K. Aizawa. **Tracking of Humans and Estimation of Body/Head Orientation from Top-view Single Camera for Visual Focus of Attention Analysis**. In *IEEE Conference on Computer Vision Workshops*, pages 1020–1027, 2009. 18

[133] S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, K. Ebihara, and S. Morishima. **Human Body Postures from Trinocular Camera Images**. In *Internation Conference on Automatic Face and Gesture Recognition*, pages 326–331, 2000. 18

[134] W. Zhang, T. Matsumoto, J. Liu, M. Chu, and B. Begole. **An Intelligent Fitting Room using Multi-Camera Perception**. In *International Conference on Intelligent User Interfaces*, pages 60–69, 2008. 18

[135] B. Peng and G. Qian. **Binocular Dance Pose Recognition and Body Orientation Estimation via Multilinear Analysis**. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008. 18

[136] L. Rybok, M. Voit, H. Ekenel, and R. Stiefelhagen. **Multi-View based Estimation of Human Upper-Body Orientation**. In *Proceeding of 20th International Conference on Pattern Recognition*, pages 1558–1561, 2010. 18

[137] J. Yao and J. M. Odobez. **Multi-Camera 3-D Person Tracking with Particle Filter in a Surveillance Environment**. In *Proceeding of 16th European Signal Processing Conference*, 2008. 18

[138] T. Gandhi and M. M. Trivedi. **Person Tracking and Reidentification: Introducing Panoramic Appearance Map (PAM) for Feature Representation**. *Machine Vision and Applications*, **18**(3):207–220, 2007. 19

[139] M. C. Liem and D. M. Gavrila. **Person Appearance Modeling and Orientation Estimation using Spherical Harmonics**. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013. 19

[140] N. Peterfreund. **Robust Tracking of Position and Velocity with Kalman Snakes**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **22**:564–569, 2000. 19

[141] C. J. Needham and R. D. Boyle. **Tracking Multiple Sports Players Through Occlusion, Congestion and Scale**. In *Proceedings of British Machine Vision Conference*, pages 93–102, 2001. 19

[142] J. Black, T. Ellis, and P. Rosin. **Multi-View Image Surveillance and Tracking**. In *IEEE Workshop on Motion and Video Computing*, pages 169–174, 2002. 19

[143] A. Mittal and L. S. Davis. **M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene**. *International Journal of Computer Vision*, **51**(3):189–203, 2003. 19

[144] B. Merven, F. Nicolls, and G. De Jager. **Multi-Camera Person Tracking using an Extended Kalman Filter**. In *Fifteenth Annual Symposium of the Pattern Recognition Association of South Africa*, 2003. 19

[145] S. Iwase and H. Saito. **Parallel Tracking of all Soccer Players by Integrating Detected Positions in Multiple View Images**. In *International Conference on Pattern Recognition*, pages 751–754, 2004. 19

[146] D. R. Magee. **Tracking Multiple Vehicles Using Foreground, Background and Motion Models**. *Image and Vision Computing*, **22**(2):143–155, 2004. 19

[147] M. Xu, J. Orwell, and G. A. Jone. **Tracking Football Players with Multiple Cameras**. In *International Conference on Image Processing*, pages 2909–2912, 2004. 19

[148] J. Kang, I. Cohen, and G. Medioni. **Tracking People in Crowded Scenes Across Multiple Cameras**. In *Proceedings of Asian Conference on Computer Vision*, pages 390–395, 2004. 19

[149] M. Isard and A. Blake. **Contour Tracking by Stochastic Propagation of Conditional Density**. In *Proceedings of the European Conference on Computer Vision*, pages 343–356, 1996. 19

[150] K. Choo and D. J. Fleet. **People Tracking Using Hybrid Monte Carlo Filtering**. In *IEEE International Conference on Computer Vision*, pages 321–328, 2001. 19

[151] J. Vermaak, A. Doucet, and P. Perez. **Maintaining Multi-Modality through Mixture Tracking**. In *International Conference on Computer Vision*, pages 1110–1116, 2003. 19

[152] Z. Khan, T. Balch, and F. Dellaert. **Efficient Particle Filter-based Tracking of Multiple Interacting Targets using an MRF-based Motion Model**. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 254–259, 2003. 19

[153] J. Giebel, D. Gavrila, and C. Schnorr. **A Bayesian Framework for Multi-Cue 3D Object Tracking**. In *European Conference on Computer Vision*, pages 241–252, 2004. 19

[154] Z. Khan, T. Balch, and F. Dellaert. **MCMC-Based Particle Filtering for Tracking a Variable Number of Interacting Targets**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(11):1805–1918, 2005. 19

[155] W. Du and J. Piater. **Multi-Camera People Tracking by Collaborative Particle Filters and Principal Axis-Based Integration**. In *Proceedings of the 8th Asian Conference on Computer Vision*, pages 365–374, 2007. 19

[156] T. Mauthner, M. Donoser, and H. Bischof. **Robust Tracking of Spatial Related Components**. In *International Conference on Pattern Recognition*, pages 1–4, 2008. 19

[157] K. Smith, D. Gatica-Perez, and J. M. Odobez. **Using Particles to Track Varying Numbers of Interacting People**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 962–969, 2005. 19

[158] C. Yang, R. Duraiswami, and L. Davis. **Fast Multiple Object Tracking via a Hierarchical Particle Filter**. In *International Conference on Computer Vision*, pages 212–219, 2005. 19

[159] S. Oh, S. Russell, and S. Sastry. **Markov Chain Monte Carlo Data Association for General Multiple Target Tracking Problems**. In *Proceedings of the 43rd IEEE Conference on Decision and Control*, pages 1–8, 2004. 19

[160] 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 Transactions on Pattern Analysis and Machine Intelligence*, **28**(12):1960–1972, 2006. 19

[161] Q. Yu, G. Medioni, and I. Cohen. **Boosted Markov Chain Monte Carlo Data Association for Multiple Target Detection and Tracking**. In *IEEE International Conference on Pattern Recognition*, pages 675–678, 2006. 19

[162] Q. Yu, G. Medioni, and I. Cohen. **Multiple Target Tracking using Spatio-Temporal Markov Chain Monte Carlo Data Association**. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 19

[163] Q. Yu and G. Medioni. **Integrated Detection and Tracking for Multiple Moving Objects using Data-Driven MCMC Data Association**. In *IEEE Workshop on Motion and Video Computing*, pages 1–8, 2008. 19

[164] W. Ge and R. T. Collins. **Multi-target Data Association by Tracklets with Unsupervised Parameter Estimation**. In *Britisch Machine Vision Conference*, pages 1–10, 2008. 19

[165] A. Andriyenko, S. Roth, and K. Schindler. **An Analytical Formulation of Global Occlusion Reasoning for Multi-target Tracking**. In *IEEE Workshop on Visual Surveillance*, pages 1806–1819, 2011. 20, 103, 105

[166] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. **Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera**. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 1820–1833, 2011. 20

[167] A. G. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu. **Multi-object Tracking through Simultaneous Long Occlusions and Split-Merge Conditions**. In *IEEE conference on Computer Vision and Pattern Recognition(CVPR)*, pages 666–673, 2006. 20

[168] S. Avidan. **Ensemble Tracking**. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **29**:261–271, 2007. 20

[169] H. Grabner and H. Bischof. **On-line boosting and vision**. In *IEEE conference on Computer Vision and Pattern Recognition(CVPR)*, pages 260–267, 2006. 20

[170] B. Leibe, K. Schindler, and L. V. Gool. **Coupled Detection and Trajectory Estimation for Multi-object Tracking**. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007. 20, 22

[171] C. Huang, B. Wu, and R. Nevatia. **Robust Object Tracking by Hierarchical Association of Detection Responses**. In *European Conference on Computer Vision*, pages 788–801, 2008. 20, 86

[172] L. Zhang, Y. Li, and R. Nevatia. **Global Data Association for Multi-Object Tracking Using Network Flows**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 20, 22

[173] Y. Li, C. Huang, and R. Nevatia. **Learning to Associate: HybridBoosted Multi-Target Tracker for Crowded Scene**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2953–2960, 2009. 20, 34

[174] J. Xing, H. Ai, and S. Lao. **Multi-Object Tracking through Occlusions by Local Tracklets Filtering and Global Tracklets Association with Detection**

Responses. In *IEEE conference on Computer Vision and Pattern Recognition(CVPR)*, pages 1200–1207, 2009. 20, 86

[175] 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 *IEEE 12th International Conference on Computer Vision*, pages 1515–1522, 2009. 20

[176] W. Choi and S. Savarese. **Multiple Target Tracking in World Coordinate with Single, Minimally Calibrated Camera**. In *European Conference on Computer Vision*, pages 553–567, 2010. 20

[177] C. H. Kuo, C. Huang, and R. Nevatia. **Multi-Target Tracking by On-Line Learned Discriminative Appearance Models**. In *International Conference on Computer Vision and Pattern Recognition*, pages 685–692, 2010. 20, 86

[178] H. Kuhn. **The Hungarian Method for the Assignment Problem**. *Naval Research Logistics Quaterly*, **2**:83–87, 1955. 20

[179] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1988. 20

[180] F. Burgeois and J. C. Lasalle. **An Extension of the Munkres Algorithm for the Assignment Problem to Rectangular Matrices**. *Communications of the ACM*, pages 802–806, 1971. 20, 89

[181] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe. **Sonar tracking of multiple targets using joint probabilistic data association**. *IEEE Journal of Oceanic Engineering*, **8**(3):173–184, 1983. 20

[182] D. B. Reid. **An algorithm for tracking multiple targets**. *IEEE Transaction on Automatic Control*, **24**(6):843–854, December 1979. 21

[183] H. Jiang, S. Fels, and J. J. Little. **A Linear Programming Approach for Multiple Object Tracking**. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 21

[184] J. Berclaz, F. Fleuret, and P. Fua. **Multiple Object Tracking Using Flow Linear Programming**. In *Winter-PETS*, pages 1–8, 2009. 21, 22, 86, 91, 92, 103, 104

[185] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua. **Tracking Multiple People under Global Appearance Constraints**. In *IEEE International Conference on Computer Vision*, pages 137–144, 2011. 22

[186] A. Andriyenko and K. Schindler. **Globally Optimal Multi-target Tracking on a Hexagonal Lattice**. In *European Conference on Computer Vision*, pages 466–479, 2010. 22, 103, 105

[187] A. Andriyenko, K. Schindler, and S. Roth. **Discrete-Continuous Optimization for Multi-Target Tracking**. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1926–1933, 2012. 22

[188] B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool. **Coupled Detection and Tracking from Static Cameras and Moving Vehicles**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(10):1683–1698, 2008. 22

[189] Z. Wu, T. H. Kunz, and M. Betke. **Efficient Track Linking Methods for Track Graphs using Network-Flow and Set-Cover Techniques**. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1185–1192, 2011. 22

[190] G. Panin, C. Lenz, S. Nair, E. Roth, M. Wojtczyk, T. Friedlhuber, and A. Knoll. **A Unifying Software Architecture for Model-Based Visual Tracking**. In *IS&T/SPIE 20th Annual Symposium of Electronic Imaging*, pages 343–356, San Jose, CA, 2008. 26, 27

[191] G. Panin. *Model-based Visual Tracking: the OpenTL Framework.* Wiley-Blackwell, 2011. 26

[192] J. Berclaz, F. Fleuret, E. Tueretken, and P. Fua. **Multiple Object Tracking Using K-Shortest Paths Optimization**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1806–1819, 2011. 31, 103, 105

[193] J. Ferryman and A. Shahrokni. **PETS2009: dataset and challenge.** In *IEEE Workshop on performance evaluation of tracking and surveillance*, pages 1–6, 2009. 31

[194] J. Ferrymani. In *1st IEEE Workshop on Performance Evaluation of Tracking and Surveillance*, Grenoble, France, 2000. 34

[195] K. Bernardin and R. Stiefelhagen. **Evaluating Multiple Object Tracking Performance: the CLEAR MOT Metrics**. *Journal on Image and Video Processing*, pages 1–10, 2008. 34

[196] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan. **The CLEAR 2006 Evaluation**. *Multimodal Technologies for Perception of Humans*, pages 1–44, 2006. 34

[197] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. **Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**(2):319–336, 2009. 34

[198] C. Stauffer and W.E.L. Grimson. **Learning Patterns of Activity using Real-Time Tracking**. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **22**(8):747–757, 2000. 39

[199] M. Harville, G. Gorden, and J. Woodfill. **Foreground Segmentation using Adaptive Mixture Models in Color and Depth**. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 3–11, 2001. 39

[200] O. Javed, K. Shafique, and M. Shah. **A Hierarchical Approach to Robust Background Subtraction using Color and Gradient Information**. In *Proceedings of the Workshop on Motion and Video Computing*, pages 22–27, 2002. 39

[201] H. Eng, J. Wang, A. Kam, and W. Yau. **A Bayesian Framework for Robust Human Detection and Occlusion Handling using a Human Shape Model**. In *International Conference on Pattern Recognition*, pages 257–260, 2004. 39

[202] T. Yu, C. Zhang, M. Cohen, Y. Rui, and Y. Wu. **Monocular Video Foreground/Background Segmentation by Tracking Spatial-Color Gaussian Mixture Models**. In *Proceedings of the Workshop on Motion and Video Computing*, 2007. 39

[203] J. Canny. **A Computational Approach to Edge Detection**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**:679–698, November 1986. 40

[204] I. Sobel. *Camera Models and Perception*. PhD thesis, Stanford University, 1970. 40

[205] I. Sobel. **An Isotropic 3 X 3 Gradient Operator**. 40

[206] G. Borgefors. **Hierarchical Chamfer Matching: A Parametric Edge Matching algorithm**. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **10**(6):849–865, November 1988. 48

[207] C. F. Olsen and D. P. Huttenlocher. **Automatic Target Recognition by Matching Oriented Edge Pixels**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**:103–113, January 1997. 48

[208] D. Comaniciu and P. Meer. **Mean Shift: A Robust Approach Toward Feature Space Analysis**. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **24**(5):603–619, 20027. 52

[209] A. Griesser, De S. Roeck, A. Neubeck, and L. Van Gool. **GPU-based Foreground-Background Segmentation using an extended colinearity criterion**. In *Proceedings of Vison, Modeling, and Visualization(VMV)*, pages 319–326, 2005. 70

[210] M. Yang, F. Lv, W. Xu, and Y. Gong. **Detection Driven Adaptive Multi-cue Integration for Multiple Human Tracking**. In *IEEE International Conference on Computer Vision*, pages 1554–1561, 2009. 86

[211] B. Yang, C. Huang, and R. Nevatia. **Learning Affinities and Dependencies for Multi-Target Tracking using a CRF Model**. In *International Conference on Computer Vision and Pattern Recognition*, pages 1233–1240, 2011. 86

[212] D. Salvi, J. Waggoner, A. Temlyakov, and S. Wang. **A Graph-based Algorithm for Multi-Target Tracking with Occlusion**. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 489–496, 2013. 86

[213] P. Konstantinova, A. Udvarev, and T. Semerdjiev. **A Study of a Target Tracking Algorithm using Global Nearest Neighbor Approach**. In *International Conference on Computer Systems and Technologies*, pages 290–295, 2003. 88

[214] Y. Bar-Shalom and X. Li. *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS Publishing, 1995. 89

[215] A. Griesser, De S. Roeck, A. Neubeck, and L. Van Gool. **GPU-based Foreground-Background Segmentation using an Extended Colinearity Criterion**. In *Proc. of Vison, Modeling, and Visualization(VMV)*, pages 319–326, 2005. 96

[216] L. Chen, G. Panin, and A. Knoll. **Hierarchical Grid-Based People Tracking with Multi-camera Setup**. In *Computer Vision, Imaging and Computer Graphics. Theory and Applications*, **274** of *Communications in Computer and Information Science*, pages 187–202. Springer Berlin Heidelberg, 2013. 103, 104

[217] A. Andriyenko and K. Schindler. **Multi-target Tracking by Continuous Energy Minimization**. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1265–1272, 2011. 103, 105