Active Visual Scene Exploration



Eric Sommerlade
Brasenose College

D.Phil. Thesis
Supervised by Dr Ian D. Reid

Robotics Research Group
Department of Engineering Science
University of Oxford

Michaelmas Term 2010

This thesis is submitted to the Department of Engineering Science, University of Oxford, for the degree of Doctor of Philosophy. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

Eric C. W. Sommerlade
Brasenose College
University of Oxford

Doctor of Philosophy
Michaelmas Term
2010

Active Visual Scene Exploration

Abstract

This thesis addresses information theoretic methods for control of one or several active cameras in the context of visual surveillance. This approach has two advantages. Firstly, any system dealing with real inputs must take into account noise in the measurements and the underlying system model. Secondly, the control of cameras in surveillance often has different, potentially conflicting objectives.

Information theoretic metrics not only yield a way to assess the uncertainty in the current state estimate, they also provide means to choose the observation parameters that optimally reduce this uncertainty. The latter property allows comparison of sensing actions with respect to different objectives. This allows specification of a preference for objectives, where the generated control will fulfil these desired objectives accordingly.

The thesis provides arguments for the utility of information theoretic approaches to control visual surveillance systems, by addressing the following objectives in particular:

Firstly, how to choose a zoom setting of a single camera to optimally track a single target with a Kalman filter. Here emphasis is put on an arbitration between loss of track due to noise in the observation process, and information gain due to higher accuracy after successful observation. The resulting method adds a running average of the Kalman filter's innovation to the observation noise, which not only ameliorates tracking performance in the case of unexpected target motions, but also provides a higher maximum zoom setting.

The second major contribution of this thesis is a term that addresses exploration of the supervised area in an information theoretic manner. The reasoning behind this term is to model the appearance of new targets in the supervised environment, and use this as prior uncertainty about the occupancy of areas currently not under observation. Furthermore, this term uses the performance of an object detection method to gauge the information that observations of a single location can yield. Additionally, this thesis shows experimentally that a preference for control objectives can be set using a single scalar value. This linearly combines the objective functions of the two conflicting objectives of detection and exploration, and results in the desired control behaviour.

The third contribution is an objective function that addresses classification methods. The thesis shows in detail how the information can be derived that can be gained from the classification of a single target, under consideration of its gaze direction. Quantitative and qualitative validation show the increase in performance when compared to standard methods.

Acknowledgements

Firstly, I need to thank my supervisor, Dr Ian Reid. I am especially indebted not only for his guidance and eloquence when it came to the production of academic output, but especially for his outstanding support – both with difficult academic problems and in more mundane situations. I would also like to thank him for the experience he provided me with by letting me take part in the EU project HERMES [1].

I am also very thankful for the deep and insightful comments from the reviewers of this thesis, Prof. David Hogg and Prof. David Murray, and the inspiring discussion during the viva.

Not only in relation to the EU project HERMES, I would like to thank my collaborators that were willing to share common academic ground with me: Pau Baiget, Nicola Bellotto, Ben Benfold, Michael Breitenstein, David Ellis, Carles Fernández, Alonso Patron-Perez and Daniel Roth. I am much obliged for all the fruitful discussions not exclusively on academic topics. Extra special thanks go out to Ben Benfold and Victor Prisacariu for sharing code that is legible, and once more to Ben being a very patient model on a very buggy catwalk.

I am also very grateful for the humane environment the Active Vision Lab has provided during the years of my D.Phil., which must be at least partly attributed to the leaders of the group. It was a joy and inspiration to work with so many gifted people. In addition to the lab members already mentioned, I would like to thank Charles Bibby, Alex Flint, Georg Klein, Christopher Mei, Somkiat Wangsiripitak and Brian Williams for both great academic and social interaction.

Furthermore, I want to thank Earl Campbell for proofreading, and my former colleagues at the CVC in Barcelona for their support, advice and endless humour: Ignasi Rius, Marçal Rusiñol and Eduard Vázquez. Thanks go also to my friends Yalçın Incesu, Derik Schroeter and Hendrik Steffen for putting up with my antics so many times, and to the family Munk/Bøcher-Ellern for a great time in CPG.

Lastly, I want to thank my family, in particular my mother, my partner Kärg Kama, and my daughter for their support, love and understanding.

[1] In this regard I would also like to acknowledge the European Union for supporting such academic interactions in the context of the framework programmes, making all participants a bit more European.

Contents

List of figures **ix**

List of tables **x**

1 Introduction **1**
1.1 Motivation and Objective . . . 1
1.2 Approach . . . 3
1.3 Methodology . . . 7
1.4 Thesis Outline . . . 7
1.5 Contributions . . . 9
1.6 Related Publications . . . 10

2 Related Work **12**
2.1 Introduction . . . 12
2.2 Surveillance Systems . . . 12
2.3 Tracking . . . 14
2.4 Detection and Classification . . . 15
2.5 Planning . . . 17

3 Background **21**
3.1 Introduction . . . 21
3.2 Notation . . . 21
3.3 Information theoretic Metrics . . . 22
3.3.1 Entropy . . . 22
3.3.2 Joint and Conditional Entropy . . . 23
3.3.3 Entropy in Continuous State Spaces . . . 24
3.3.4 Entropy of Multivariate, Normal-distributed Random Variables . . . 25
3.3.5 Entropy in Mixed State Spaces . . . 25
3.3.6 Mutual Information . . . 26
3.4 System Definition . . . 27
3.4.1 Expected Utility for Greedy Planning . . . 30
3.4.2 Utility over Multiple Steps . . . 31

3.5 Utility of Information . . . 33
3.5.1 Maximisation of Mutual Information . . . 34
3.6 Combination of objectives . . . 35
3.7 Measuring information . . . 36
3.8 Other Performance Metrics . . . 40
3.9 System Architecture . . . 43
3.9.1 Passive Visual Sensing . . . 46
3.9.2 Active Visual Tracking . . . 46
3.10 Conclusion . . . 49

4 Tracking **50**
4.1 Introduction . . . 50
4.2 Related Work . . . 51
4.3 Robust Entropy-based zoom selection . . . 52
4.3.1 Entropy-based zoom selection . . . 52
4.3.2 Innovation based zoom selection . . . 59
4.4 Analysis . . . 61
4.4.1 Combination of Previous Methods . . . 66
4.4.2 Estimation of the innovation sequence . . . 68
4.5 Discussion . . . 73
4.5.1 Live experiments . . . 75
4.5.2 Application in Visual Surveillance . . . 78
4.6 Conclusion . . . 79

5 Multiple Targets and Scene Exploration **80**
5.1 Introduction . . . 80
5.2 Related Work . . . 82
5.3 Multiple Targets . . . 86
5.3.1 Prioritisation of Targets . . . 88
5.4 Appearance/Disappearance of Targets . . . 90
5.4.1 Scene Exploration with a Single Camera . . . 91
5.4.2 Poisson Process . . . 91
5.4.3 Learning scene activity . . . 95
5.4.4 Evaluation . . . 97
5.4.5 Metrics . . . 97
5.4.6 Experiments . . . 98
5.5 Conclusion and Future Work . . . 102
5.5.1 Future Work . . . 102

6 Cooperative Scene Exploration **104**

6.1 Introduction . . . 104
6.2 Tracking with multiple cameras . . . 105
6.2.1 Merging of Kalman Filters . . . 106
6.2.2 Examples . . . 109
6.2.3 Prior Knowledge . . . 111
6.2.4 Live system . . . 113
6.3 Scene exploration using Multiple Cameras . . . 115
6.3.1 Prior: Birth-and-death process . . . 116
6.3.2 Observations: Detector performance . . . 118
6.3.3 Multiple observations . . . 119
6.3.4 Implementation and behaviours . . . 121
6.4 Combining objectives . . . 121
6.5 Experiments . . . 123
6.6 Conclusion . . . 125
6.6.1 Future Work . . . 126

7 Target Identification **128**
7.1 Introduction . . . 128
7.2 Related Work . . . 129
7.3 Camera Control . . . 131
7.3.1 Formal Problem Statement . . . 132
7.3.2 Mutual Information from Classification . . . 135
7.4 Performance Comparison . . . 137
7.4.1 Evaluation . . . 138
7.4.2 Discussion . . . 143
7.5 Simplified approach . . . 143
7.6 Incorporating Gaze . . . 145
7.7 Results . . . 151
7.7.1 Quantitative Evaluation . . . 152
7.7.2 Live System . . . 157
7.8 Conclusion . . . 158
7.8.1 Future work . . . 160

8 Conclusion **161**
8.1 Summary . . . 161
8.1.1 Conclusion . . . 163
8.2 Contributions . . . 165
8.3 Future Work . . . 166
8.3.1 Data association . . . 166
8.3.2 Target reacquisition . . . 167

8.3.3 Live System . . . 168
8.3.4 Inclusion of scene knowledge . . . 168
8.3.5 Planning Horizon . . . 169
8.3.6 Reinforcement Learning . . . 170
8.3.7 Higher Level Control . . . 170

A Information-theoretic Identities **171**
A.1 Derivation of Mutual Information . . . 171
A.2 Mutual Information always increases with more data . . . 171
A.3 Mutual Information yields diminishing returns . . . 172
A.4 Difference between Differential and Discrete Entropy . . . 173
A.5 Entropy of Multivariate, Normal-distributed Random Variables . . . 174

B Birth and Death Process **176**

C Kalman Filter **177**
C.1 Kalman filter . . . 177
C.2 Jacobians for Various Observation Functions . . . 178
C.2.1 Pin-hole Projection . . . 178
C.2.2 Spherical Projection . . . 179
C.2.3 Pin-hole Projection in Spherical Coordinates . . . 179

D Classification Systems **181**
D.1 Biometric performance indicators . . . 181
D.2 Worked example of classification . . . 182

References **184**

List of Figures

1.1 Uses for surveillance . . . 2
1.2 Uses for surveillance . . . 3
1.3 Schematics of control approach . . . 4

3.1 Target behaviour influenced by surveillance . . . 28
3.2 Mutual information from varying filters . . . 39
3.3 False positives and negatives in tracking . . . 41
3.4 Temporal and spatial overlap of trajectories . . . 44
3.5 System architecture . . . 45
3.6 Target detection from supervisor camera . . . 47
3.7 Multi-resolution tracking over long distances . . . 48

4.1 Overview of the decision process . . . 53
4.2 Uncertainty reduction over focal length . . . 57
4.3 Zoom setting, visibility and entropy in a surveillance video . . . 58
4.4 Observation noise varying with focal length influences zoom control . . . 61
4.5 Tracking behaviour of Kalman filters with entropy based zoom control . . . 63
4.6 Tracking behaviour of Kalman filters with covariance based zoom control . 65
4.7 Inclusion of the innovation term in entropy based control . . . 67
4.8 Tracking behaviour of entropy based control with focal length adaption . . 70
4.9 Tracking performance for entropy based control with innovation covariance 72
4.10 Mean tracking performance for varying α . . . 73
4.11 Comparison of zoom control methods . . . 74
4.12 Zoom control bounded by overestimated noise . . . 74
4.13 Minimum conditional entropy vs. zoom and covariance . . . 75
4.14 Mean conditional entropy three filter types . . . 75
4.15 Screenshots from sample implementation . . . 77
4.16 Recovered trajectories from different zoom control approaches . . . 79

5.1 Mutual information of three targets . . . 89
5.2 Example for exploration . . . 94
5.3 Comparison of target positions . . . 96
5.4 Appearance maps from CAVIAR data set . . . 97

5.5 Comparison of scheduling methods w.r.t. resolution of targets 101

6.1 Multiple updates 108
6.2 Merged multiple updates 109
6.3 Visibility based target selection 111
6.4 Collaborative tracking of a target 112
6.5 Zoom settings of two cameras with occlusions 114
6.6 Screenshots from live experiment 115
6.7 Birth and death process per scene location 118
6.8 Mutual information over pan, tilt settings 122
6.9 Performance w.r.t. weight of objectives 125

7.1 Graphical model per target 133
7.2 Example run of the simulation 139
7.3 Evaluation w.r.t. residence time 141
7.4 Recall/precision plot for varying parameters. 142
7.5 Sketch of angle of incidence 146
7.6 Sample model of detector performance and resulting information gain . . . 149
7.7 Mutual information over varying angles 150
7.8 Information from different views 151
7.9 Example view of the data set 153
7.10 Mutual information for constant and varying prior 154
7.11 Change in identification likelihood 155
7.12 Performance comparison w.r.t. angle of incidence 156
7.13 Performance comparison w.r.t. camera delay 157
7.14 Sample head images with the corresponding gaze estimations 157
7.15 Screenshots from live system 159

8.1 Preference specification for objectives 163
8.2 Information theoretic control as middle layer 164

D.1 Development of classification likelihood for a target 183

List of Tables

3.1 Notation used throughout this thesis 22
3.2 Overview of vision processes used in the system 49

5.1 False positives and negatives w.r.t. appearance rate 100
5.2 Latency and resolution w.r.t. appearance rate 100

6.1 Performance comparison of mutual information based scheduling 125

Chapter 1

Introduction

1.1 Motivation and Objective

Object detection, tracking, and classification in video data are crucial elements for further reasoning in modern vision-based systems. In various situations, such as sports events, surveillance[21], and healthcare (see figure 1.1(a) and 1.1(b)), this video data stems from a set of active and passive cameras. Whereas passive cameras maintain the pose they have been installed in, active cameras are able to change some of their parameters according to external demands. The potential benefits of the active cameras come at the cost of control.

For example, tasks such as person identification require or profit from a higher resolution [117] that usually cannot be obtained from cameras which serve to overview the scene and have a short focal length. Since the cost of installation and resulting amount of video data to be transferred, stored and observed, prohibits naïve addition of passive cameras, an alternative solution is to use active cameras with a pan/tilt/zoom (PTZ) functionality, which focus on occurrences of interest.

With the pervasive presence of cameras comes a higher demand for automatic control, as manual operation is expensive and straining, and can lead to sub-optimal control decisions (see figure 1.2). However for active cameras to deliver information not achievable by their passive counterparts, this requires the automatic control to be designed carefully.



(a) Multiple cameras for sports coverage.



(b) Surveillance at an airport.

Figure 1.1: Uses for surveillance. Photos courtesy EU-project APIDIS, `http://www.apidis.org`, EU-project AVITRACK

The objective of a surveillance system is set by the maintainer of the system and can vary depending on the time or other circumstances of its use: at times it might be preferable to have an early detection of targets, at other times most accurate detection is optimal, or a combination of the two. Furthermore, these objectives can be conflicting – early detection of targets favours a wide field of view, whereas identification benefits from high resolution facial images of the targets.

The work in this thesis provides a facility to ease the interpretation of video data taken by a number of active cameras involved in the observation of the same scene. We explore methods to turn visual input into a strategy to follow given tasks in an optimal fashion.



Figure 1.2: Cameras relying on manual control might not be put to optimal use. Photo "homeland security" by Flickr user "bhikku", copyright Nigel Britton, made available under a Attribution-Noncommercial-Share Alike 2.0 license

The tasks and the relative importance of these tasks can be set by a controller, either human or artificial.

In this regard, this thesis proposes methods to control active cameras so that they best fulfil given, potentially conflicting objectives.

1.2 Approach

We address this problem in this thesis in an information-theoretic way, and build upon a decision-theoretic framework. We require that each of the objectives is expressed as an information gathering problem, i.e. its desired control should maximise the information resulting from the observation process. While this is intuitively appealing, it also has significant theoretical and practical benefits: the objectives become comparable. Each objective maximises information about one or another state, and this effectively puts the objectives on the same scale. This allows us to employ a fundamental result of utility theory, by which any mixture of the objectives can be addressed by their linear combination.

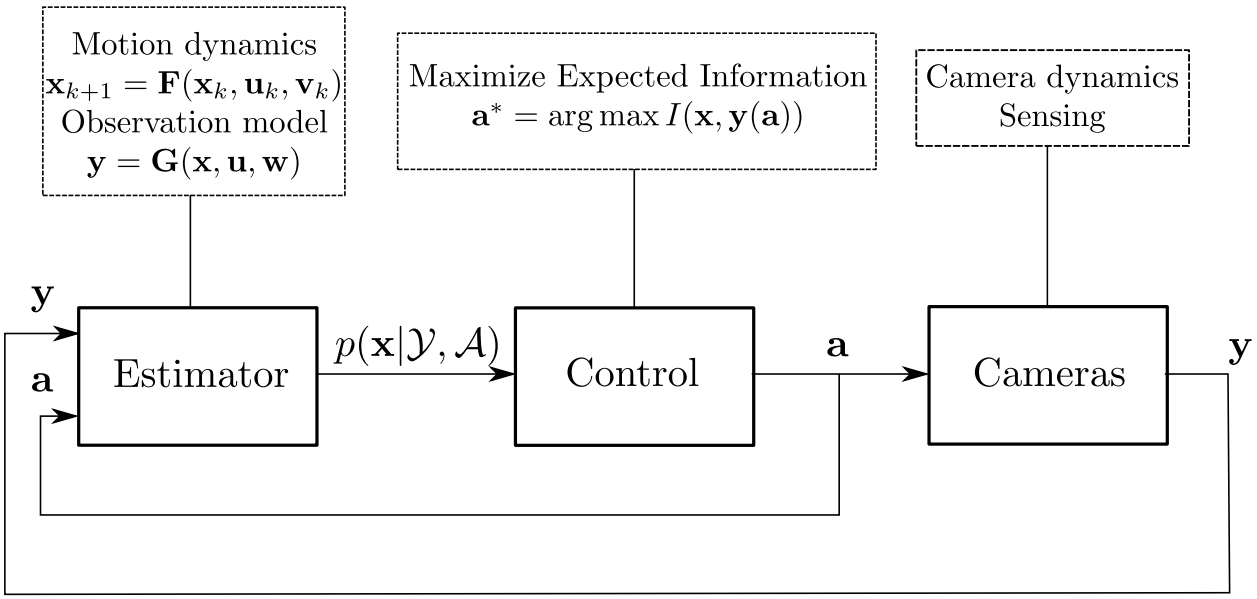


Figure 1.3: Schematics of the control approach, after Jacobs [79, p354]. The state vector **x** is estimated from the observations **y** by a set of cameras. The latter are controlled by input **a** which is chosen to provide maximum information about the state estimate. Unlike in linear, Gaussian models, the proposed control method makes use of the uncertainty in the estimate.

Control of any real-world quantity has to deal with the uncertainty of measurements, the state it tries to model, and its actions. Probabilistic and information-theoretic concepts lend themselves to model the uncertainty within these quantities, and provide a measure of the information they provide.

Information-theoretic metrics make use of the underlying probability of the state estimate and the measurement process. This is in stark contrast to standard control approaches, where the assumption of separability of estimation and control results in control that only takes into account the mean of the estimate. Figure 1.3 shows a block diagram of the proposed control approach. Contrary to control approaches that are separated, the full distribution of the estimate is fed into the controller.

Importantly, this thesis addresses the control of a real system; we need to embrace the concept of non-deterministic outcomes for given control inputs. This means the system

needs to react to observations directly after sensing time, and to take into account the inherent uncertainty in the sensing process.

In the robotics community, the problem is known as *information gathering problem*[151]. However, the focus there is on a static environment and mobile robots, where the objectives are exploration of an unknown, static environment, and localisation. These objectives are also conflicting: journeying far from known landmarks rapidly discovers new areas in the map, but risks getting lost as new landmarks are not reliable enough. The typical sensors used here are laser scanners, sonar and LIDAR (LIght Detection And Ranging).

In this thesis, we transfer this approach to the domain of surveillance and visual tracking. Not only do we focus on standard sensors and models in vision, such as cameras with limited field of view, person detectors and Kalman filters, we also have to take into account the dynamics of the state. Contrary to freely moving robots and a static environment, we have sensors that are fixed in their location, but a highly dynamic state space comprising the appearance and motion of targets under surveillance.

The relationship of this work to the area of computer vision is threefold. First and foremost, we focus on sensor data from cameras. Second, while the general framework for integration of objectives is flexible, we address the typical objectives of a surveillance system [85, 73, 106] of target tracking, detection and target identification.

Third, the control problem is addressed in the active vision sense, as set out by Bajcsy [12]. In her landmark paper she defined active vision as an intelligent data acquisition process. Control strategies thus should minimise a loss function while maximising information about the state. This comes with an understanding that the feedback into the control system is dependent on the context as well as the scene, and originates from sen-

sors that produce noisy or even wrong data. The imperative of this approach is that instead of processing and reshaping this imperfect data via hard constraints, its inherent uncertainty has to be embraced and fed back into the sensing strategy.

The traditional approaches to this problem come from stochastic control theory or phrase the problem as a partially observable, sequential decision making problem. The solution to this problem yields a policy, i.e. optimal control commands for every possible configuration that can occur. The size of the state space is the first issue. As we are interested in tracking multiple targets, the state space grows with the number of targets, which makes storage of optimal control commands difficult even in the very compact notation of a linearised position of targets.

Furthermore, the optimality of each of the commands depends on the stretch of time planned for. This means that not only the next best solution might be considered, but also the accumulation of information over a near future stretch of time, or planning horizon. Existing solution methods involve taking expectations over predictions of the state, which are hard to solve for any real world problems.

In the light of this, we take a common approach (e.g. [74]) and consider only *greedy* or *myopic* planning, i.e. we use new information immediately and plan only a single time step ahead at the time, which can be calculated on the fly for every configuration of targets presented to the system. The resulting policy will try to maximise the instantaneous information acquisition, a method also called *information surfing* [69].

Specific to this thesis, we focus on methods to control one or multiple active cameras. Within this work we assume that cameras have a known position and orientation, and observe a partially overlapping area, i.e. handover problems are limited due to known geometric constraints. Furthermore we assume that the network topology is known, and all

decisions are made centrally. This implies centralised sensor coordination and information fusion. Decentralised control is an extension to this problem that has been addressed in the robotics community, and results in "auctioning" and "bargaining" methods from game theory [98, 143]. The output of the methods proposed in this thesis can then be used as the underlying reward for every required subset of the cameras.

1.3 Methodology

Evaluating scheduling algorithms on live video data is difficult. For a fair comparison, each algorithm should run on the same input, which is hard to obtain with human actors, as human movement cannot be repeated exactly. Pre-recorded video can be used for evaluation if the resolution is high enough to support a "virtual zoom" approach, where the image is down-sampled and cropped to a desired field of view. A high resolution is required if object detectors or trackers are to be run on the down-sampled image.

In this thesis, we take a two-pronged approach. Quantitatively, we evaluate the proposed methods by simulation, based on ground truth data acquired from live surveillance systems. We evaluate the methods on basis of standard metrics in the surveillance community, which allows us to relate our results to previous work.

Qualitative experiments are run using an actual implementation of a surveillance system, in order to give a validation of the methods with real noise and time constraints.

1.4 Thesis Outline

Chapter 2 gives a short overview of related work from the system perspective and the planning approach. Literature relevant to specific topics is discussed in each of the subsequent chapters.

In chapter 3 we introduce the information theoretic metrics we build our control on, and give an informal reasoning behind the choice of this control approach. We furthermore address a practical problem when using the metrics just described and how to circumvent this. Lastly, we address alternative metrics used in the literature.

Performing some of the experiments in this thesis required a surveillance system to run and test the proposed control methods. This involved the installation of hardware, but also the development of a software architecture supporting real time control as well as subsequent inspection. We present the system that served as a testbed for our algorithms in section 3.9.

We first apply the information-theoretic metrics proposed to the control of a single, active camera for the purpose of target tracking. Chapter 4 provides a detailed account of the improvements we made to the method first presented by Denzler et al. [46], and how these improve the method's applicability and useful zoom range.

In chapter 5 we extend the previous method to the tracking of several targets, and introduce a term that encapsulates the uncertainty about a new, previously undetected target in the supervised area. This formulation uses only a single camera, and makes the implicit assumption that targets are detected once they are in the field of view. The approach in this chapter addresses three issues: firstly, how far to zoom onto the chosen target, minimising the risk of losing track; secondly, how to decide which of the detected actors to observe more closely and, lastly, how to explore the scene to search for new, yet undetected actors. We use an activity map to incorporate scene specific actor behaviour. This map keeps track of the likelihood that actors appeared in this area of the scene. The probability of making a new detection is obtained from the locations which are missed when a set of parameters is chosen. This acts as a counterbalance for the zoom onto the

actors. The best parameters are the ones which maximally reduce the uncertainty of all or a subset of actors and minimise the chance of an undetected appearance of a new actor.

Chapter 6 introduces the concept of the detector performance for a given target position. This not only gives a natural way to incorporate the usefulness of sensing a certain location into the objective function, but also supports the use of several cameras.

Surveillance systems often have the aim of identifying targets that move in the observed area. Chapter 7 presents an information-theoretic objective function that addresses exactly this problem; for this, we introduce the target's identity into the state and show that maximisation of mutual information with respect to this part of the state yields better performance than standard methods.

The thesis closes with a summary, the conclusions that can be made from the experiments and resulting control rules, and a number of future research directions that this work has triggered.

1.5 Contributions

The work presented in this thesis makes the following contributions.

- In the context of surveillance, we present the first information-theoretic approach to control of multiple cameras and show how this approach allows integration of arbitrary objectives (Chapter 3).

- We extend an existing method for information-theoretic target tracking and parameter selection to include the innovation covariance in the Kalman filter of the tracked object, effectively increasing available zoom range and robustness (Chapter 4, published in [135]).

- We present a novel extension of the term for tracking a single target to multiple targets (Chapter 5, published in [136, 137]).

- We propose a term that aims at detection of target. Whilst on its own, the resulting control is a "scanning" of the supervised environment. When putting this term in competition with terms from tracking, the camera behaviour alternates between tracking and exploration (Chapter 5, also published in [136, 137]).

- A novel objective function for tracking of targets using single and multiple cameras which incorporates observability constraints from scene knowledge (Chapter 6, published in [138]).

- We show how to turn detector output into a sensor model. This sensor model then yields the mutual information from observing a set of discrete locations (Chapter 6, published in [134]).

- We present a novel scheduling method that integrates the classification performance of a given identification system (Chapter 7, submitted to [139] and co-authored with Ben Benfold).

- We present a database based architecture for surveillance systems, and a client implementation that performs tracking as well as detection in parallel (Chapter 3.9), which serves as a testbed for live experiments. This has been co-authored with Bellotto et al. [17].

1.6 Related Publications

The publications created during this thesis were published at conferences highly relevant to either the fields of vision or robotics.

- Eric Sommerlade and Ian Reid, "Information theoretic Active Scene Exploration" Proc. IEEE Computer Vision and Pattern Recognition (CVPR), 2008 [136]

- —, "Information-theoretic Decision Making for Exploration of Dynamic Scenes", Proceedings of the 5th International Workshop on Attention in Cognitive Systems (WAPCV), 2008 [137]

- —, "Influence of Zoom Selection on a Kalman filter", IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2008 [135]

- —, "Cooperative Surveillance of Multiple Targets using Mutual Information", Proceedings of the ECCV Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2), 2008 [134]

- —, "Probabilistic Surveillance with Multiple Active Cameras", IEEE International Conference on Robotics and Automation (ICRA), 2010 [138]

- Eric Sommerlade et al., "Gaze Directed Camera Control for Face Image Acquisition", Proceedings of the International Conference on Robotics and Automation, 2011 [139]

- Nicola Bellotto et al., "A Distributed Camera System for Multi-Resolution Surveillance", Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), 2009 [17]

Furthermore, we started investigating the use of prior knowledge in the control problem. The resulting work has been covered in the co-authored publications with Patron and Reid[115], Baiget et al. [11], Breitenstein et al. [28], and Ellis and Reid [53].

Chapter 2

Related Work

2.1 Introduction

This chapter first gives a very short overview on related approaches to control in the context of surveillance systems, then addresses particular issues tackled in this thesis. Most of the more specific related work will be discussed in the chapter that addresses the corresponding problem.

2.2 Surveillance Systems

Ng and Ng [106] give an overview on sensor systems, and define the term sensor management, which encompasses the control problem of this thesis. They propose a classification of sensor management systems into three levels. Systems from the first level provide individual control for the sensors, such as direction and exposure control. Systems from the second level focus on the objectives, i.e. prioritisation of tasks such as target acquisition. The third level comprises management methods that provide dynamic sensor placement, and effective allocation of different sensors according to objectives. From this categorisation, the proposed method occupies levels one and two, whereas connections to the third level are made available. The architecture of the proposed approach can be classified after Ng and Ng as a centralised, hierarchical, multiple platform sensor system.

From an actual system implementation perspective, a more recent overview on visual sensor networks is given by Soro and Heinzelman [142], summarising properties such as resource requirements, performance, storage, accuracy of measurements and collaboration among single sensor nodes. They subdivide sensor management into the tasks of hardware architectures, networking, signal processing and collaboration. Only the last two points are addressed in this thesis.

Early surveillance systems with an active camera have been set up with a single, fixed supervisor camera and one or few pan, tilt and zoom cameras [15, 35, 45, 67, 72, 116]. This kind of master-slave-configuration is often used in the context of scheduling the supervision of targets, either as a dynamic discrete optimisation problem [10], or motivated by packet scheduling algorithms [35, 121]. Hampapur et al. [72] uses hand crafted rules to assign active cameras to actors, and chooses the zoom setting via geometric reasoning. The system uses multiple calibrated supervisor cameras for 3D tracking, and incorporates a head detector to focus the zoomed view onto the face of persons. Lastly, there are some systems with less common approaches, for example assigning 'virtual forces' to the camera movement and targets [1], or assignment of targets based on rules from cinematography [50]. Greiffenhagen et al. [67] use probabilistic reasoning in a system to control the pan and tilt and zoom parameters of an active camera such that the likelihood or a person's face in the camera's field of view is greater than a specified threshold.

A working example of collaboration between static and active cameras has been given recently by Krahnstoever et al. [85], where active cameras are steered according to the inputs from static ones. Soto and Song et al. [140, 143] separately address the tasks of tracking and classification with multiple cameras in a game-theoretic way, but each objective on its own. The collaboration among the cameras is obtained from a distributed method

that seeks to reach consensus among neighbouring nodes. Song and Roy-Chowdhury [141] address data association across multiple, static cameras. The underlying utility functions are not comparable (feature similarity across entry/exit points and path smoothness), and need to be combined in a multi-objective optimisation framework.

2.3 Tracking

Tracking single objects using active cameras has received plenty of attention in the vision community; earlier it focussed on single cameras or stereo heads [27, 40] and the processing constraints from hardware, now multiple cameras in much looser connection are controlled, addressing coordination and control delay in networks of cameras [49, 72, 143, 126].

Another approach to tracking problems has been taken by the visual servoing community. In general, image features or prior knowledge about the tracked object's structure is used to centre the camera's field of view as desired [31]. The error in the image features can then be related directly to the required control commands through the interaction matrix, which encapsulates the geometry of the controlled device and the camera. Here, the uncertainty in the tracking process is usually not included in this control approach.

Gans et al. [60] tracks multiple targets using a visual servoing approach. He addresses multiple objectives through a control law which requires weighting terms for mean, variance and visibility variation that are acceptable in the resulting control. The input feature to focus the camera is the mean position of the tracked targets, and the variance is used to obtain a zoom rule. The variance is estimated from the spread of the measured target positions.

Regarding the use of multiple zoom levels for classification, Smith et al. [133] uses several cameras at different, but constant focal lengths to infer about the activity of a person under surveillance. Locations of relevant objects are transferred into different views using epipolar geometry, and acquisition supported by the colour model obtained in the first view. The combined information from multiple zoom levels then facilitates activity recognition.

2.4 Detection and Classification

Planning of optimal positions of sensors has a long history; on one hand there is optimal placement for automatic manipulation [3], and next-best view planning for automatic acquisition of 3D-models [118]. The art-gallery problem [109] is one of the first methods specific to optimal coverage of locations: find the minimum number of guards and their locations in a polygonal environment, such that the environment's entire boundary (the art gallery's walls) is visible. González-Banos and Latombe [64] extend the standard algorithm from computational geometry to incorporate view-related constraints such as distance and the view's angle of incidence.

Instead of our own review, we refer to Abidi et al. [2], who give an exhaustive survey of methods specific to surveillance of extended areas (several kilometres) up to 2004; they address methods relating to coverage, sensor positioning (with varying mobility and elevation), zoom versus focus, and data fusion.

Placement of sensors using conditional entropy of the sensor locations has been proposed early on by Carla Currin et al.[37]. This criterion maximises the uncertainty among the participating sensors (e.g. by placing them as far apart as possible). As this approach does not consider the remaining uncertainty about the places without sensing, this place-

ment can be too wasteful – the sensors should stand as far apart as possible, but also maximise the proximity to all remaining locations.

Krause et al. [87] points out that placement of sensors using maximisation of mutual information between sensed and not sensed locations does exactly the latter, as it maximally reduces the uncertainty about the not sensed locations. Krause also shows that placement of sensors using mutual information is NP-complete (i.e. NP-hard, but a solution can be verified in polynomial time), and presents a greedy approximation.

Osborne[111] presents an approach to global optimisation that can be applied to the sensor placement problem. The example given in [111] outperforms Krause's method; instead of incrementally placing a growing set of sensors, the number of sensors is kept fixed but their location is optimised by exploiting the global optimisation framework. Krause's or Currin's methods could be used to bootstrap Osborne's approach. The number of sensors can be varied, if the underlying distance metric in their approach models cost of addition or removal of sensors.

The latter two approaches express the dependency of spatially neighbouring locations using Gaussian Processes. These models need to be learned or estimated from expert knowledge. While they allow inference about neighbouring areas from a limited set of locations, they make the measurements dependent given the actual state; they are jointly Gaussian. When more sensors are added, then less and less information is to be gained from the remaining locations. In fact, when all locations are being sensed, there is no information to be made. This means that the mutual information criterion used by Krause is not monotonic. However, in the case of independent locations and noisy measurements (as modelled in this work), more measurements will always result in more information.

2.5 Planning

The control methods in surveillance systems are often not very sophisticated, in that only the next best option is chosen, i.e. greedy planning is used. Also the uncertainty in the motion of tracked targets and the sensing process is often ignored. The simplest approaches use greedy planning over heuristics, such as the closest or oldest target [35, 67, 72, 121, 126].

Krahnstoever et al. [85] employ a probabilistic objective function for the identification of targets. They hypothesize plans of a maximum length for each camera and target, and search over the assignments of cameras to targets that yield the highest likelihood to capture a facial image of the target. The uncertainty in the target motion enters the planning method only at one location: For each hypothesized plan, target motions are addressed by predicting their position according to a constant velocity model by the time a camera needs to move this target into the centre of its field of view. The longer this time, the higher rises the uncertainty in the target's position, which provides an upper limit to the plan length, and longer plans are discarded.

Del Bimbo and Pernici [45] schedule target capture using the kinetic travelling salesperson problem, with a single, active camera and a supervisor camera. Akin to the original problem, the active camera is seen as an agent that needs to visit customers, with the additional challenge that the latter are now travelling as well, with a constant velocity. The approach does not make use of any uncertainties in the sensing or motion of the targets, and results cover different simulation settings.

There is very little other work on planning for control in visual surveillance, and even less that has probabilistic or information theoretic approaches. Soto et al. [143] proposes a distributed approach to camera control, but the underlying utility function addresses the

distance of the target to the camera only. Instead, the robotics community has addressed uncertainty in the sensing process.

Chhetri [32] addressed planning over several, discrete time steps and multiple, discrete sensor actions using exhaustive search methods, and branch and bound methods. The investigated problems are short timespan (less than 6 time steps) and very few sensor actions (less than 10), and for a single target only. This allows formulation of a graph based approach, i.e. the states can be easily enumerated. The graph is then searched using branch-and-bound methods. For larger graphs (as one would encounter with more sensor options, longer time spans, and more targets), this becomes infeasible as each node in the graph requires memory to store intermediate results such as covariance matrices or means.

Grocholsky [68] addresses motion planning of multiple mobile sensors using optimal control theory over information-theoretic objective functions. The focus of Grocholsky's work is decentralised decision making, i.e. communication of local decisions and finding agreements. He adds the uncertainty (in form of the Fisher information matrix) to the state space [68, p76], and phrases the control problem as an optimal control problem over a constant, finite time with an additional constraint on the objective function at the final time of the optimisation.

The assumption in Grocholsky's approach is that the state's evolution can be addressed by a ordinary stochastic differential equation. This is possible if the sensors used always yield a measurement. In this case, the covariance, or in Grocholsky's case the Fisher information matrix, always can be updated. In case of vision sensors, the measurement might fail if the target is not in the field of view, and the expected distribution will turn from unimodal into a mixture model.

Ryan and Hedrick [125], and Hoffmann and Tomlin [74] recently proposed methods that make a particle filter approximation to the state space. These particles are propagated through the dynamic system. Ryan approximates the information-theoretic measures (entropy and mutual information) by linear interpolation of the particle weights at particle locations in state space, and introduces extra particles to clamp the distribution to 0 in sparsely populated areas. Mutual information is then obtained from integration of this piecewise linear approximation. Hoffmann uses standard quadrature methods to evaluate the integral for mutual information.

As Ryan concentrates on detecting and tracking of a single target with a single unmanned aerial vehicle, she is able to address multiple look-ahead steps with a planning horizon of 7 time steps. Hoffmann makes a single step approximation, i.e. greedy search [74, p43]. Hoffmann concentrates on multi-sensor search of a single target in minimum time, this means the aim of the control problem is to minimise the expected number of future observations required to localise the target.

In the work by Singh et al. [131], a path planning method for multiple robots is proposed which maximises the information gain over all locations visited, while keeping the costs of visiting the places under a given budget. The space of possible locations is discrete and finite, and the information that can be gained at each location depends on the locations visited before only, i.e. the problem addressed has no temporal dependency. The planning is using a greedy heuristic over a set of the maximum number of nodes to be visited, and is solved by a branch-and-bound method. This off-line planning does not take into account newly collected data, and as such is non-adaptive towards changes in the environment that do not conform to the expectations. This is a feasible approach when dealing with

mostly static environments and known noise, but inappropriate given the spontaneous nature of human behaviour.

Spletzer and Taylor[145] present a method that can track several targets with several mobile robots on a ground plane. They propose to minimise the distance between a target and a robot, and take the expectation over the state and measurement space. This expected loss is minimised independently for robot and each control parameter. They approximate the expectation of this loss function using Monte Carlo quadrature. From the paper, it is not clear how the robots perform collaborative sensing.

Chapter 3

Background

This chapter gives a summary of the notation and mathematical identities used and referred to in this thesis, proceeds with a general definition of the control problem, and approaches to its solution are discussed. The information theoretic metric proposed is presented and evaluated on standard filters.

3.1 Introduction

Here we introduce the notation and prerequisites for the applications presented in the later chapters. The mathematical identities presented here are common in standard text-books, e.g. MacKay [94] and Bishop [20], and repeated here for convenience. We conclude the chapter with a argument for an information theoretic objective function, and present the metrics used throughout this thesis.

3.2 Notation

Throughout this thesis, we follow the notation given in table 3.1. In addition, we use superscripts for set and sequence indices if other indices are required, e.g. for time varying elements of a set: s_k^i is the i-th element in the sequence, at time step k.

For random variables, we follow the conventions of standard text books, e.g. Bishop [20]: For a set of symbols with N unique elements, $\mathcal{S} := s_1, ..., s_n$, a discrete random variable is a random instantiation from this set and is denoted simply as $s \in \mathcal{S}$. The probability

Table 3.1: Notation used throughout this thesis

Scalars are Latin or Greek lower-case characters	s, γ
Vectors are bold faced	$\mathbf{x}$, $\boldsymbol{\mu}$
Matrices are bold faced and upper-case	$\mathbf{P}$, $\boldsymbol{\Sigma}$
Identity matrix	$\mathbf{I}$
Sets and Sequences are calligraphic	$\mathcal{S} = \{s_1, \ldots, s_N\}$ $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
Indices:	
into matrices or vectors	$\mathbf{x}_i$, $\mathbf{P}_{ij}$
time varying entities usually use k	$\mathbf{x}_k$
shorthand for $\mathbf{H}(\mathbf{a})$	$\mathbf{H}_\mathbf{a}$

of encountering the symbol s_i is $p(s = s_i)$. For convenience, this is also written as $p(s_i)$ and $p(s)$, if no ambiguities arise. The probability of encountering any of the symbols in the set is always $\sum_i p(s_i) = 1$ and $p(s_i) \geq 0$. If there are dependencies between random variables (e.g. the chance of winning the lottery depends on choosing the right numbers, but can also depend on one's partner not forgetting to hand in the ticket), these are expressed by their joint probability $p(s = s_i, t = t_j) = p(s, t)$, and their conditional probability $p(s = s_i | t = t_j) = p(s|t)$, i.e. probability of s assuming value s_i, given that t is already known.

A continuous random variable is defined in a similar manner on a continuous domain, $\mathbf{x} \in \mathcal{R}^N$. Instead of a probability, a density function is defined with $\int_\mathbf{x} p(\mathbf{x})\, d\mathbf{x} = 1$, and $p(\mathbf{x}) \geq 0$.

3.3 Information theoretic Metrics

3.3.1 Entropy

Entropy was introduced by Claude Shannon [127] to measure the uncertainty about the outcome of a random process. His choice of the name comes from the similarity to expressions in statistical mechanics, in particular in Boltzmann's H-theorem and Gibbs'

entropy[127, p11]. Shannon derived his definition of entropy according to three requirements (continuity in the probability, increase of uncertainty with more events, and decomposability), which leads to the definition of the **information** of an outcome x as

$$h(x) = -\log p(x). \tag{3.1}$$

The entropy, or uncertainty about a random process, is the average information content

$$H(x) = -\mathrm{E}_x \{\log p(x)\} = -\sum_{\mathcal{X}} p(x) \log p(x), \tag{3.2}$$

i.e. the expected value of the information.

The entropy for discrete state spaces is non-negative, and reaches its maximum for a uniform distribution, i.e. where all symbols are equally likely to appear.

This discrete measure can be used in the context of classification. A classification method assigns probabilities that measure how likely a classified subject belongs to one of the known classes. The classification is very uncertain – and thus not useful – if the distribution of the returned values is close to uniform. If one of the likelihoods reaches 1, however, the system is certain and the result can be trusted. The corresponding distribution has a very low uncertainty, and is thus preferable to other, less informative observation settings.

3.3.2 Joint and Conditional Entropy

The joint likelihood of two concurring events x and y can be expressed as (using the product rule of probability theory) $p(x,y) = p(x)p(y|x)$. The joint entropy is

$$H(x,y) = \mathrm{E}\{-\log p(x,y)\} = \mathrm{E}\{-\log p(x) - \log p(y|x)\} = H(x) + H(y|x) \tag{3.3}$$

This is an example for Shannon's requirement for decomposability of the measure.

The term $H(y|x)$ is the conditional entropy, i.e. the entropy of a posterior distribution, when the state has been conditioned on a second event x. For a given, fixed event x_0, this entropy can be obtained directly as

$$H(y|x=x_0) = -\mathrm{E}_y\{\log p(y|x=x_0)\} \tag{3.4}$$

$$= -\sum_{\mathcal{Y}} p(y|x=x_0)\log(p(y|x=x_0)) \tag{3.5}$$

This entropy is a function of the event x_0. Conditional entropy yields an *a priori* measure for the uncertainty after making an observation, which describes the expected uncertainty for all possible observations:

$$H(y|x) = \mathrm{E}_x\{H(y|x)\} \tag{3.6}$$

From the decomposability of entropy one obtains the chain rule of entropy for N random variables $x_1 \ldots x_N$:

$$H(x_1,\ldots,x_N) = H(x_N|x_1,\ldots,x_{N-1}) + H(x_1,\ldots,x_{N-1}) \tag{3.7}$$

3.3.3 Entropy in Continuous State Spaces

The entropy $H(\mathbf{x})$ of a distribution models the informativeness of a probability distribution $p(\mathbf{x})$, which in turn describes our belief in the position of $\mathbf{x}$. For continuous random variables this is defined as [20]:

$$H(\mathbf{x}) = -\mathrm{E}\{\log p(\mathbf{x})\} = -\int_{-\infty}^{\infty} p(\mathbf{x})\log(p(\mathbf{x}))\,d\mathbf{x} \tag{3.8}$$

Note that this differs from the discrete case A.4. Unlike discrete entropy, differential entropy can be negative.

The requirements on entropy leave room for an arbitrary factor. Traditionally for discrete state spaces, this is the binary logarithm: the information is measured in bits. In

continuous state spaces, the chosen base is e and the information is given in "nats" [94, p265]. For optimisation, the factor can largely be ignored, however, care must be taken when comparing information from different state spaces. In this thesis, we always use the identifier log and choose the basis appropriately.

3.3.4 Entropy of Multivariate, Normal-distributed Random Variables

For a k-dimensional, normally distributed vector $\mathbf{x}$ with covariance $\mathbf{\Sigma} = \mathrm{E}\{\mathbf{x}^T\mathbf{x}\}$ and mean $\boldsymbol{\mu} = \mathrm{E}\{\mathbf{x}\}$, the probability density function is

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\mathbf{\Sigma})}} \exp(-\,{}^1\!/_2 (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})) \tag{3.9}$$

The entropy of this random variable is (the derivation can be found in appendix A.5)

$$\begin{aligned} H(\mathbf{x}) =& \mathrm{E}\{-\log p(\mathbf{x})\} && (3.10)\\ =& \,{}^1\!/_2 \log((2\pi e)^k \det(\mathbf{\Sigma})). && (3.11) \end{aligned}$$

The result shows that the entropy is influenced by the covariance of the state only.

3.3.5 Entropy in Mixed State Spaces

Some states can comprise a continuous part as well as a discrete part. The entropy of these states can be derived similarly as in the cases above, and the full details are presented by Nair et al. [104].

An example for mixed states is the blockage of one of two passages – it can be modelled as two queues that are selected by an independent random binary variable [104]. Another example is the identity and position of a subject, which are most often assumed as independent.

If we split the state $\mathbf{x}$ in equation 3.6 into a continuous $\mathbf{y}$ and discrete part $\mathbf{c}$, we obtain as mean conditional entropy by using $p(\mathbf{x}) = p(\mathbf{c},\mathbf{y}) = p(\mathbf{c}|\mathbf{y})p(\mathbf{y})$:

$$H_{\mathbf{a}}(\mathbf{x}|\mathbf{o}) = -\sum_{\mathbf{c}} \int p_{\mathbf{a}}(\mathbf{y}|\mathbf{o}) p_{\mathbf{a}}(\mathbf{c}|\mathbf{y},\mathbf{o}) \log(p_{\mathbf{a}}(\mathbf{y}|\mathbf{o}) p_{\mathbf{a}}(\mathbf{c}|\mathbf{y},\mathbf{o}))\, d\mathbf{y}\, d\mathbf{o} \tag{3.12}$$

$$= H_{\mathbf{a}}(\mathbf{y}|\mathbf{o}) + H_{\mathbf{a}}(\mathbf{c}|\mathbf{y},\mathbf{o}) \tag{3.13}$$

3.3.6 Mutual Information

For any two distributions $p(\mathbf{x})$ and $p(\mathbf{y})$, we can predict how much information about $\mathbf{x}$ an observation (or event) from $\mathbf{y}$ would generate on average. This *mutual information* can be interpreted as the difference of the expected uncertainty to the current uncertainty, and as a measure of the dependence between these random variables.

$$I(\mathbf{x},\mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}) = \mathrm{E}\left\{\log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right\} \tag{3.14}$$

This quantifies the amount of information we expect upon making a new observation. If the random variables $\mathbf{x}$ and $\mathbf{y}$ are independent, the information about $\mathbf{x}$ potentially gained from an observation is zero. Mutual information is symmetric to its argument and and non-negative, $I(\mathbf{x};\mathbf{y}) \geq 0$. The symmetry of mutual information yields the following identities

$$I(\mathbf{x};\mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x}) \tag{3.15}$$

$$I(\mathbf{x};\mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{y},\mathbf{x}) \tag{3.16}$$

$$I(\mathbf{x};\mathbf{x}) = H(\mathbf{x}) \tag{3.17}$$

Concavity of Mutual Information w.r.t. the number of arguments

Krause and Guestrin showed relevant properties of mutual information under common assumptions [86]. Assuming that observations $\mathbf{y}_i$ and $\mathbf{y}_j$ are independent given the state

variables $\mathbf{x}$ (i.e. $p(\mathbf{y}_1,\ldots,\mathbf{y}_N|\mathbf{x}) = \prod_{1..N} p(\mathbf{y}_i|\mathbf{x})$), it holds that on average, mutual information does not decrease if more evidence is added:

$$I(\mathbf{x};\mathbf{y}_i,\mathbf{y}_j) \geq I(\mathbf{x};\mathbf{y}_i) \tag{3.18}$$

Mutual information has the property of yielding *diminishing returns* for an increasing number of observations: whereas adding observations does not decrease the information, the increase from N to $N+1$ observations will be less than the increase from $N-1$ to N. This property can be shown by the concavity of $I(\mathbf{x};\mathbf{y}_1\ldots\mathbf{y}_N)$ w.r.t. the number N of $\mathbf{y}$:

$$I(\mathbf{x};\mathbf{y}_1\ldots\mathbf{y}_{N-1}) - I(\mathbf{x};\mathbf{y}_1\ldots\mathbf{y}_{N-2}) > I(\mathbf{x};\mathbf{y}_1\ldots\mathbf{y}_N) - I(\mathbf{x};\mathbf{y}_1\ldots\mathbf{y}_{N-1}) \tag{3.19}$$

Proof for this can be found in sections A.2 and A.3.

3.4 System Definition

We now give a more detailed description of the control problem we solve in this thesis, and argue for the use of mutual information as objective function. First, we abstract the uncertain nature of the sensing process and the environment as follows.

- We discretize time by sampling with the minimum frame period Δ of all the cameras, such that $t_k = k\Delta$. This allows us to use an index-only notation for all involved quantities.

- $\mathbf{x}_k := \mathcal{X}_k = \{\mathbf{x}_k^1,\ldots,\mathbf{x}_k^N\} \in \mathbb{X}$ is the state of the world at time step k, which is the number N of the targets, their identity, and position and velocity in the area under observation by at least one camera. If there is no ambiguity, then we often simply refer to the current state by $\mathbf{x}_k$, even though this state is a set. For all but chapter 7, the identity of the targets is ignored.



Figure 3.1: A supervised target reacting to the observation process by readying her camera to take a picture herself, and the result (Right picture by Lyndsey Pickup).

- The parameters of the observation process are the control inputs: $\mathbf{a} \in \mathcal{A}^C$ are the pan, tilt, and zoom values for all C cameras. Throughout this thesis, these parameters are also called "actions".

- The motion of the targets is described by the state dynamics, which is a model that takes the state of the world one time step into the future:

$$\mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{a}_k, \nu_k, k). \tag{3.20}$$

  This function takes as argument not only the current state and the observation parameters, but also the discrete-time process noise ν_k. We make the assumption that the noise is time-invariant during each time step, zero-mean, and has a covariance $\mathbf{Q}_k$. The identities of targets are not changing, and thus exempt from state dynamics. Furthermore, we assume that the observation is not influencing the behaviour of the observed targets. Although debatable (see figure 3.1, and Steve Mann's "Shooting Back"[1]), the interaction of the cameras with the subjects is not considered in this thesis.

[1] `http://wearcam.org/shootingback.html`, retrieved on 2011-01-09

- The observation process of all cameras is given by an observation function

$$\mathbf{o}_k = \mathbf{h}(\mathbf{x}_k, \mathbf{a}_k, \xi_k), \tag{3.21}$$

which maps the current state to all observations under the current observation parameters. These are the (noisy) positions of the targets in the camera views, as well as labels from an identification system. For the dynamic part of the state, the noise term ξ_k is modelled by an independent, identically distributed random variable with covariance $\mathbf{R}_k$. These observations can be actual positions of the observed target, or empty, if a camera was not able to successfully observe the target.

- A fundamental result of decision theory [57, 58, 107] is that a single, scalar function suffices to encapsulate the advantages of actions for a given objective. This means that there is no need to express rewards for successful actions in a different way than penalties for unsuccessful ones. We thus employ a single utility function that expresses a single objective of the observation process. This function maps the current action and current state to a reward $u : \mathbb{X} \times \mathbb{A} \rightarrow \mathbb{R}$:

$$u = u(\mathbf{x}_k, \mathbf{a}_k). \tag{3.22}$$

The sensing process as well as the world dynamics are inherently uncertain: The targets behave according to the model only up to a certain degree, and sensing actions might not be executed exactly due to mechanical constraints of the devices. The current target state is thus modelled probabilistically using a current distribution $p(\mathbf{x}_k|\{\mathbf{o}_1 \dots \mathbf{o}_k\})$, which relies on all previous measurements to yield an estimate. The measurements have limited accuracy, too; hence the state's distribution will only be constrained after incorporation of the latest observations.

This incorporation is the application of Bayes' rule for all observations made so far:

$$p(\mathbf{x}_k|\mathbf{o}_k) = \frac{p(\mathbf{x}_k|\{\mathbf{o}_1,\ldots,\mathbf{o}_{k-1}\})p(\mathbf{o}_k|\mathbf{x}_k)}{p(\mathbf{o}_k)} = \frac{p(\mathbf{x}_k^-)p(\mathbf{o}_k|\mathbf{x}_k)}{p(\mathbf{o}_k)}. \tag{3.23}$$

Here $p(\mathbf{x}_k^-)$ denotes the prediction of the state from the last time step $k-1$, by propagating the previous, fully updated state through the state dynamics.

The update of the state estimate by observations from various sources is called data fusion. In this thesis we address centralised data fusion, i.e. all data is processed by one unit, and the result is used to make a central decision. This is also called the "super Bayesian" approach [98, p37].

3.4.1 Expected Utility for Greedy Planning

At each time step, the system has to find an observation parameter for the next time step. The utility function in equation 3.22 implicitly defines the parameter that provides the most reward, and which parameter is thus the best for the current objective.

For planning approaches, however, the system's uncertainty in observation and motion has extensive consequences. Let us assume for this section that we are only interested in the immediate and next reward of an action, so called *greedy*, or *myopic* planning. As can be seen from equation 3.21, the choice of the next action, $\mathbf{a}_{k+1}$, only affects the next observation. This next observation, $\mathbf{o}_{k+1}$, is used to update the distribution over the next state $\mathbf{x}_{k+1}$, as in equation 3.23. The actual reward for the next action is then the expected utility over all possible states:

$$U^+(\mathbf{o}_{k+1}, \mathbf{a}_{k+1}) = \mathrm{E}_{\mathbf{x}}\{u(\mathbf{x}_{k+1}(\mathbf{o}_{k+1}), \mathbf{a}_{k+1})\}, \tag{3.24}$$

where we have made explicit the dependency of the future state's update on $\mathbf{o}_{k+1}$. This observation $\mathbf{o}_{k+1}$ is obviously also not known (as we are still at time step k), so all possible

outcomes of the next observation process need to be taken into account with their respective likelihood of appearance. This yields the final objective function for the current state and the next action:

$$U^*(\mathbf{x}_k, \mathbf{a}_{k+1}) = \mathrm{E}_{\mathbf{x}_{k+1}, \mathbf{o}_{k+1}}\{U(\mathbf{x}_{k+1}(\mathbf{o}_{k+1}), \mathbf{a}_{k+1})\} \tag{3.25}$$

$$= \mathrm{E}_{\mathbf{x}_{k+1}, \mathbf{x}_k}\{U(\mathbf{x}_{k+1}(h(\mathbf{x}^-_{k+1}, \mathbf{a}_k)), \mathbf{a}_{k+1})\}. \tag{3.26}$$

The best observation parameter is then given by maximisation:

$$\mathbf{a}^*_{k+1} = \arg\max_{\mathbf{a}} U^*(\mathbf{a}_{k+1}, \mathbf{x}_k) \tag{3.27}$$

3.4.2 Utility over Multiple Steps

In many planning problems, the utility of an action is not noticed in the next time step. This is the case where longer planning is involved, epitomised in the artificial intelligence community by Feldman and Sproull's "Hungry Monkey" problem [57]. Here the protagonist first has to place a box before being able to reach a banana, as any action before correct placement and final grasp yield negative utilities: the monkey gets hungrier.

Ideally, the utility function in equation 3.22 addresses these long term plans, and previous derivation can be extended to include multiple time steps into the future by a recursive definition, i.e. the utility function itself is the expected reward over the time step $k+2$. A full evaluation of such a utility is computationally expensive, and alternative solutions have been proposed.

The problem definition can be seen as an instance of a stochastic optimal control problem, where the aim is to obtain a control law for every point in the state space. If this gives the optimal control for a current point in state space, the principle of optimality [79] guarantees optimal behaviour for the next state.

When the problem can be phrased in differential equations without strong linearities, this control problem can be turned into a boundary value problem and solved numerically. For example Grocholsky [68] discusses the optimal control of an unmanned aerial vehicle (UAV) using information-theoretic objective functions. The UAV carries a bearing only sensor, which always yields a measurement. For sensors with a limited field of view, the objective function becomes highly non-linear, and is not linearisable any longer.

This non-linearity can be addressed using dynamic programming [119], e.g. the value iteration algorithm [151], but in the case of several targets, the curse of dimensionality strikes; the creation of look up tables is infeasible for larger state spaces. Reinforcement learning [149] tries to acquire the optimal control law by learning the transition function that takes the state from one step to the next given a certain action, and the corresponding reward. For this, the learning procedure has to explore the state space, which is equally hard. Not only because a similar look-up table is produced, but this table needs to be filled from sufficient amount of observations. This is difficult especially in the light of human activity, which is unlikely to repeat in the same manner exactly.

In machine learning, stochastic optimal control is often phrased as a sequential decision making problem, and as such it is an instance of a partially observable Markov decision process (POMDP). Early methods had to reduce problems drastically (e.g. Trémois and Le Cadre[155] propose target pursuit on a 5×5 grid for moving sensor and moving target with a deterministic observation function). Although nowadays there are approximate solvers available [88, 144, 151], large and continuous state spaces are still prohibitive (for example, the decentralised method in [144] reduces the environment to a graph of 21 nodes).

In light of this, we consider only *greedy* or *myopic* planning, i.e. we plan only a single time step ahead at the time. The resulting policy will try to maximise the instantaneous information acquisition, a method also called *information surfing* [69].

3.5 Utility of Information

Apart from the permission to employ a single, scalar utility function to express the utility of actions, another result from decision theory is that the maximisation of the expected utility yields a valid strategy, i.e. it achieves the aims expressed by the original utility function. Unfortunately, the design of such a utility function requires that a preference of one probability distribution over another can be made. For example, if a distribution p_1 is preferred to p_2, then the expected utility over p_2 must be less than the expected utility over p_1 for any combination of actions. This hands the burden of rationality back to the original decision maker, who has to design the utility function accordingly. Furthermore, the expected utility must correctly reflect the preference of a single outcome given a probability distribution, which requires that the preference ordering must meet a set of axioms.

Luckily, our aim is information maximisation and reduction of uncertainty. A rational decision is thus to prefer actions which result in a higher information. As shown by Manyika and Durrant-Whyte [98, p95], the axiomatic requirements on a utility function can be fulfilled by information theoretic metrics.

They propose the negative of Shannon's information content, conditioned on the outcome of the observation[98, ibid.],[68, p49]:

$$u(\mathbf{x}, \mathbf{a}) := \log p_{\mathbf{a}}(\mathbf{x}|\mathbf{o}). \tag{3.28}$$

The expected utility then reduces to the negative conditional entropy (see section 3.3.2)

$$U^H(\mathbf{x}, \mathbf{a}) := \mathrm{E}_{\mathbf{x},\mathbf{o}}\{\log p_{\mathbf{a}}(\mathbf{x}|\mathbf{o})\} = -H(\mathbf{x}|\mathbf{o}). \tag{3.29}$$

This can be interpreted as the maximisation of the negative uncertainty, or alternatively the minimisation of the remaining uncertainty in the state variable after observation.

3.5.1 Maximisation of Mutual Information

The objective function in equation 3.29 fills the criteria set out by Bajcsy [12], whereas control strategies should minimise a loss function and maximise the information gathered during the sensing process.

An alternative is to use mutual information as expected utility function,

$$U^I(\mathbf{x}, \mathbf{a}) = I_{\mathbf{a}}(\mathbf{x}; \mathbf{o}) = H(\mathbf{x}) - H_{\mathbf{a}}(\mathbf{x}|\mathbf{o}), \tag{3.30}$$

which has the quantity

$$u(\mathbf{x}, \mathbf{a}) := \log \frac{p_{\mathbf{a}}(\mathbf{x}|\mathbf{o})}{p(\mathbf{x})} \tag{3.31}$$

as underlying utility function. This is a more natural choice, as in contrast to negative conditional entropy, mutual information remains positive in continuous state spaces, and is invariant to a change in variables.

The optimal choice of the observation parameter $\mathbf{a}$ is the one that maximises the mutual information

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} U^I(\mathbf{x}, \mathbf{a}) = \arg\max_{\mathbf{a}} I_{\mathbf{a}}(\mathbf{x}, \mathbf{o}) \tag{3.32}$$

This can be interpreted as choosing the action which makes the distribution $p(\mathbf{x}|\mathbf{o})$ most peaked relative to $p(\mathbf{x})$. Since the a priori probability of the state does not depend on the chosen parameter ($p(\mathbf{x}|\mathbf{a}) = p(\mathbf{x})$), this is equivalent to a maximisation of the negative conditional entropy

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} -H_{\mathbf{a}}(\mathbf{x}|\mathbf{o}) = \arg\max_{\mathbf{a}} U^H(\mathbf{x}, \mathbf{a}) \tag{3.33}$$

3.6 Combination of objectives

This section describes how to combine conflicting objectives in a surveillance system, for example tracking and acquiring new targets.

Early detection of an object is an important aspect of a surveillance system, as well as obtaining higher resolution imagery of the targets. These two objectives are mutually exclusive, and the importance can vary. For example, it might be of utmost importance to register all targets entering the scene as soon as possible. Further objectives are conceivable, e.g. identification of targets, their motion pattern, their meeting points.

As proposed by Manyika[98, p129], any multi-objective decision problem can be expressed as a simple linear combination of the individual utilities involved, if each utility or value assigned to these objectives is based on information theoretic measures, and if the utility for each objective comes from the same underlying state and is independent from the state underlying the other utilities (again Manyika, [98, p108]).

This makes the assumption that the utilities assigned to the pursuit of each objective are comparable, i.e. have some meaning to the decision maker that makes a given amount of one utility better than another. In Keeney and Raiffa's words [82, p68], utilities are not compared if they are in totally different units. The standard approach to decision problems with multiple objectives, as proposed by these authors, is the maximisation of the product of the single utilities. This has been used by e.g. Cook et al. [34] to decide about the actions of a group of vehicles, where the utilities under consideration are a preference over stealth of operation (or not being detected), and a preference over exploration of the environment.

Instead, when the utility is obtained from the involved probability distributions in an information-theoretic manner, the resulting values are placed on the same (relative) scale

and can be compared. The information returned for each of the objectives passes the test proposed by Keeney and Raiffa (in the form of [29, p117]) for appropriateness of a linear combination of objectives:

Take two single utilities A and B, with best and worst outcomes a^*, a_* and b^*, b_*, when keeping all other involved single utilities constant. Now consider the combined events $W := [a^*, b^*]$, $F := [a_*, b_*]$, $M_1 := [a^*, b_*]$, and $M_2 := [a_*, b^*]$. If there is no preference for either a 50% chance of W or F, or for a 50% chance for M_1 or M_2, then the utilities can be combined additively.

Along the line of Manyika's research, Bourgault et al.[26] use a linear combination of utilities to arbitrate accuracy of robot localisation with the accuracy of the map created by the same robot. This violates the conditions set out by Manyika, as the information obtained about the map depends on the accuracy of the positional estimate of the robot. Is it still conceivable, though, that one could be equally satisfied by robot operation if the outcome is – with equal chances – a total success or total failure, or a perfect map and poor robot position and conversely.

3.7 Measuring information

The information-theoretic measures – entropy and mutual information – depend on the distribution of the underlying state and observations. These can be modelled using sufficient statistics, such as normal distributions. If these statistics are accurate and there exists a closed form solution for the entropy, then the resulting metric will be accurate as well.

For example, if the transition and observation functions, as well as the noise terms can be modelled as Gaussian random variables, the entropy and mutual information can be

obtained from the explicit form of the entropy in equation 3.11. This is the case for linear systems and the standard, linear Kalman filter [13].

In the case of approximations, the resulting entropy can deviate from the actual information. For example in the extended Kalman filter (EKF), the transition and observation functions are linearised by a first order Taylor approximation [13, 157]. This effectively approximates the involved probability density functions by Gaussian random variables and transforms them by the linearised system equations. These approximations can introduce not only a large bias, i.e. deviation of the expected from the actual mean, but also a large difference to the actual covariance of the estimated state. Difficulties with EKFs in exploration using bearing-only SLAM have been reported in the literature [130], in the case of translational freedom, however, the tendency to steer towards the landmarks seems to outweigh the effects of turning sideways to the target.

As an example, we show how the mutual information from the observation of a single target is influenced by the observation model, and the parametrisation of the state and observation space. We model a target at constant position with a Gaussian random variable, and observe the target with either a pin-hole or a spherical camera model. The target is then moved away from the optical axis towards the right, while maintaining the same distance. This has the same effect as rotating the camera to the left. The mutual information is obtained from the state entropy $H(\mathbf{x})$ and conditional entropy $H(\mathbf{x}|\mathbf{y})$, both in Gaussian form.

For the latter, we linearise the involved projection, and assume an infinite field of view, i.e. the observation is certain to be made. We obtain the resulting information gain through the following means:

- straightforward propagation of the input covariance through the linearised system equation, as in the extended Kalman filter case, and analytical evaluation of the resulting mutual information from the resulting covariance,

- statistical approximation from a set of sample points in the form of the unscented Kalman filter (UKF), see van der Merwe and Wan [157], and van der Merwe[156] for details. We also compared the Iterated Sigma-Point Kalman filter proposed by Sibley et al.[129], but the results were indistinguishable from the UKF case. Again, the resulting covariance directly yields mutual information.

- direct numerical approximation of the mutual information by Monte Carlo simulation. Samples are generated from the initial distribution and propagated through the non-linear system. The resulting observations then contribute to the covariance according to the original sample weights. The last method, proposed by Boers et al.[23], estimates the mutual information directly from the observations, without an intermediate step through a covariance matrix. In both approximations we use 1000 samples in the particle filter.

The coordinate systems we analyse here are a planar, pin-hole projection model with observations in planar or spherical coordinates, as well as a spherical projection model with observations in spherical coordinates. In the case of observations in the latter representation, we naturally discard the depth information, as a single camera is a bearing-only sensor (i.e. no depth measure can be obtained from a single point). The choice of these projection functions and coordinate systems has effects on the underlying Jacobians that are needed for the update of the covariance matrix. These Jacobians are given in detail in appendix C.2.

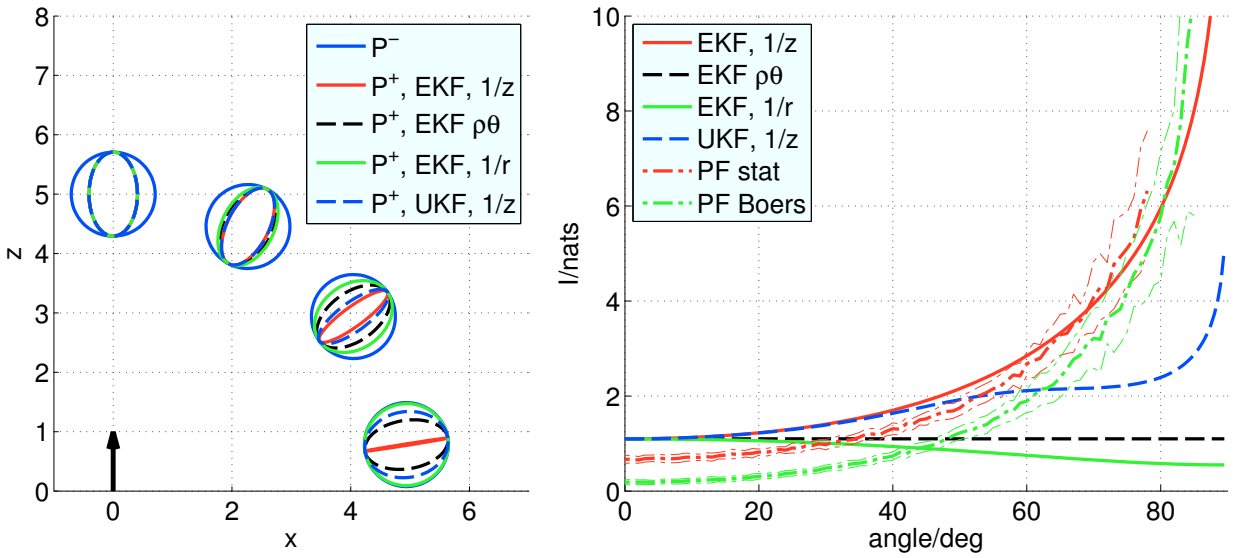


Figure 3.2: Left: camera at origin, and original and updated covariance ellipses for varying angles. Right: mutual information for different parametrisations and observation functions. The thinner lines around the particle filter curves (PF stat and PF Boers) indicate a single standard deviation.

The state space of the target is thus $[x\ y\ z]^T \in \mathbb{R}^3$, and the observations are $[\hat{x}\ \hat{y}]^T \in \mathbb{R}^2$ for planar, pin-hole coordinates, and $[\rho\ \theta]^T \in \mathbb{R}^2$ for spherical coordinates.

The initial covariance of the target is $\mathbf{P}^- = {}^1/_2\,\mathbf{I}$, and the observation noise is $\mathbf{R} = 10^{-2}\mathbf{I}$. Figure 3.2 shows a comparison of the updated covariances and the resulting mutual information for varying angles of rotation. As can be seen, the resulting covariances are similar for small deviation from the optical axis, but deviate quickly. For the pin-hole camera model, the EKF approximation (EKF 1/z) actually attains an infinite mutual information at 90°, as entries of the Jacobian of the observation function become infinite.

The mutual information from the UKF (UKF 1/z) does not diverge as quickly as the EKF, but also increases with greater angle. This is confirmed by the particle filter based methods (PF stat and PF Boers). Note that these yield non-monotone results due to the random choice of particles. The spherical projection model (EKF $1/r$) instead yields less information with increasing angle.

The only correct information – in the sense that information should remain constant for any view angle – is returned when using the parametrisations of spherical coordinates (EKF $\rho\theta$).

The important conclusion of this section is that control of cameras using pin-hole observation models and the wrong parametrisations results in cameras that look away from the target, i.e. keeping them as close to the boundary of their field of view as possible.

3.8 Other Performance Metrics

The increased information about a state should reflect itself in some properties of the resulting observations of the target. Unfortunately, there is little consensus about the right metrics for surveillance systems [45]. For tracking alone, there exists a plethora of metrics that address the quality of the trajectories obtained, see e.g. Kao et al.[81] and references. Not only are these hard to directly maximise; they usually do not incorporate the uncertainty in the tracking process, i.e. an accurate tracker will yield the same results as a fairly uncertain one with the same mean. Furthermore, there are no common metrics regarding the quality of observation due to increase in zoom. Recently, Kao et al. [81] presented an information-theoretic metric to judge the output of a tracking system, which also partially addresses the probabilistic output. This method yields better values for targets that are tracked at higher accuracy in the data association. If the increased resolution in the tracking process yields a better state estimate, then this method can be used to judge the performance of the system under scrutiny.

However, the metric does not rise if the increased resolution addresses a part of the target state that provides no improvement in the accuracy of the target's trajectory. This is the case for identification, where the actual performance of the system is better measured

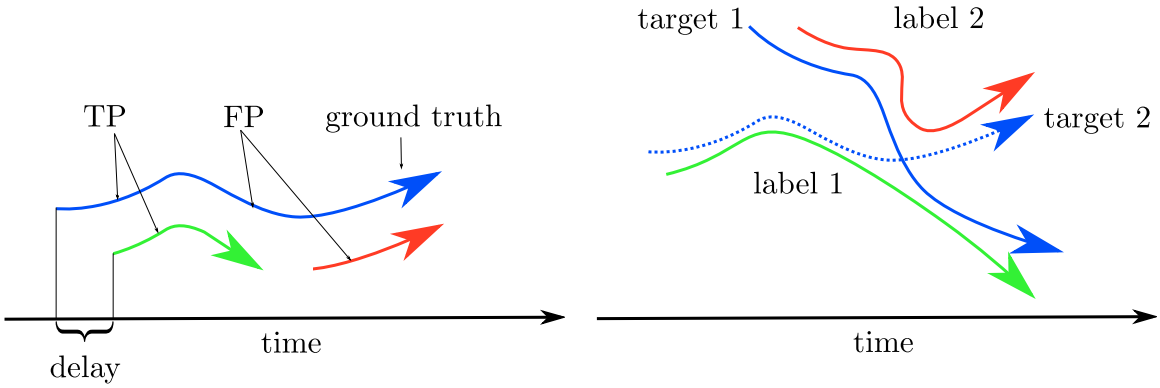


Figure 3.3: Left: An original trajectory results in two measured trajectories. The first one has a delay in acquisition, the second one is not associated with the original, and counted as a wrong trajectory, i.e. false positive. Right: an example of an identity switch. The original trajectories cross and are wrongly tracked and the labels reported by the system confused.

by the percentage of correctly and incorrectly identified subjects. We are interested in such a metric because higher resolution benefits identification and classification tasks, as supported by recent studies[117].

To address this, we go back to the metrics in the form presented by Yin and Makris [162]. Because they explicitly incorporate the output of tracking processes on image data, they can provide a metric for an increase in resolution. Figure 3.3 shows the concepts of track association, delay (latency) in target acquisition, and identity switches due to data association failure. The latency is a measure for the delay of the detection of a target in the scene, e.g. when the camera is currently zoomed onto another target.

The idea is to compare the original data with the results from the tracking process. The original data, or ground truth, is either obtained from annotation by humans or trivially obtained if the system is evaluated by simulation. Among the output from visual surveillance systems are trajectories, i.e. a list of estimates that belong to the same target, as well as a list of the area the target occupies in each frame of a video sequence, and all output belonging to a single target can be called “track”, or “system track”, whereas the ground

truth takes the name "truth track". A system track has a start time k_s^s, and end time k_e^s, correspondingly for a truth track: k_s^g, k_e^g.

The resulting tracks are compared to the ground truth tracks by association: if a system track is close enough to a truth track under a given metric, they are associated. If several system tracks associate to a single truth track, all but one of them will be false positives (FP). This can happen if a system track gets fragmented or split, e.g. due to tracking loss. Similarly, a missed detection of a target results in a false negative (FN).

We use the concept of spatial and temporal overlap as metrics for track association. Temporal overlap O_t is the fraction of time steps both the ground truth track and system output exist, e.g. if there is a delay in acquisition of a target, the system output will start this amount of time later than the ground truth track. If the absolute time both tracks coexist is

$$T^{\cap} = \max(0, \min(k_e^g, k_e^s) - \max(k_s^g, k_s^s)), \tag{3.34}$$

and the absolute time any of both tracks exists is

$$T^{\cup} = \max(0, \max(k_e^g, k_e^s) - \min(k_s^g, k_s^s)), \tag{3.35}$$

the resulting temporal overlap is

$$O_t = \frac{T^{\cap}}{T^{\cup}} \tag{3.36}$$

Spatial overlap is the relative area the observations share with the ground truth. Per time step k, it is defined of the fraction of the intersection to union of the observation areas:

$$o_k = \frac{\|A_k^s \cap A_k^g\|}{\|A_k^s \cup A_k^g\|}, \tag{3.37}$$

and the average spatial overlap is

$$O_s = \frac{1}{k_e^g - k_s^g} \sum_{k=1..T} o_k. \tag{3.38}$$

A track is considered a false positive if the average spatial coverage is above a threshold, but the temporal overlap is too small. A track is considered a false negative if it is overlapping either spatially or temporally below given thresholds, and a true positive if it fulfils both criteria.

The average spatial overlap is less than one if the camera is not constantly observing the target, or only seeing a part it. For example, if a target is tracked at half the size during half of the time it resides in the scene, its average spatial coverage would be 0.25.

The overall coverage is the sum of all average spatial overlaps for all ground truth tracks, and is the relative increase of object area due to zooming. This metric measures the average observed area, relative to the ground truth value. Successfully observing the whole scene, i.e. all contained targets, with a zoom setting of 2 would result in an overall coverage of 2.

An example is shown in figure 3.4, where the areas of a single target in the ground truth data (left) are compared side by side (or time step by time step) with the areas of the same target while zooming. This is shown for a longer sequence and two targets in a top-down view in the right part of figure 3.4.

3.9 System Architecture

This section gives a short overview on the actual implementation of the live system. Details of the system have been published in collaboration with Bellotto et al. [17].

The underlying criterion for the system's architecture was the requirement for a facility that stores and provides access to recorded data, and that simplifies the communication between distributed clients. Furthermore, the architecture should support encapsulation

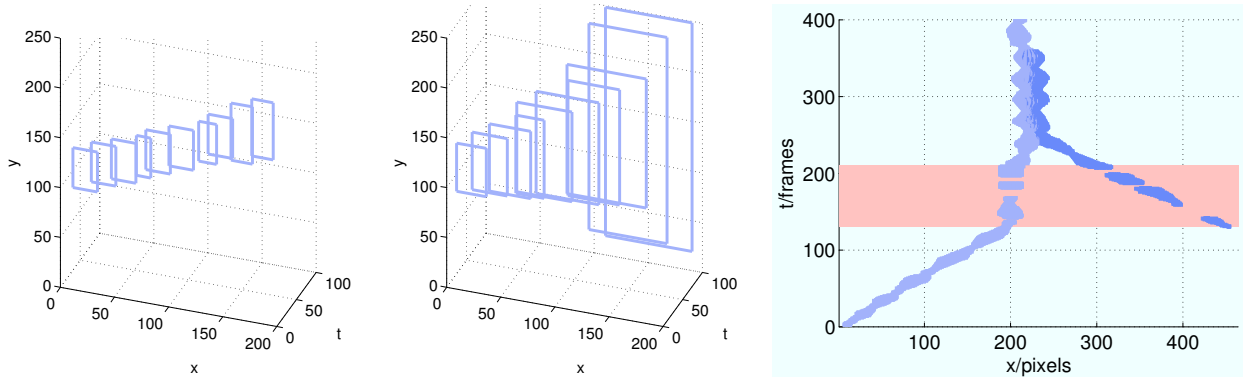


Figure 3.4: Left: a set of detections in image space at different time steps. Without zoom, only the change in size of the target is observed. Centre: When zooming, the effective resolution of the target is increased. Right: top-down view of detections for a sequence with continuously changing zoom and pan of the camera. Between frames 130 and 200 (shaded red), the camera zooms out and acquires the second target (dark blue), then periodically moves between the targets until they meet.

of each client's functionality: if a single client stops working, the rest should be able to keep running.

To address these requirements, we selected an architecture which is based on a central supervising unit which communicates with its clients through the backbone of an SQL-database. The clients are semi-autonomous tracking processes, but also a data fusion process, as well as a control process. The database maintains tables for observations from clients, results from the data fusion process, image data, control commands, as well as scene knowledge and target information. All data is stored in a global coordinate system shared by all clients of the system. A schematic of the system is shown in figure 3.5.

The active clients obtain system information, such as calibration information and control commands by polling the database. Once the clients receive the appropriate command, they start tracking or acquisition of targets with the parameter settings that have been sent.

After detection or tracking, the clients send observations back into the database, after transforming these into the world coordinate system. A data fusion process polls these

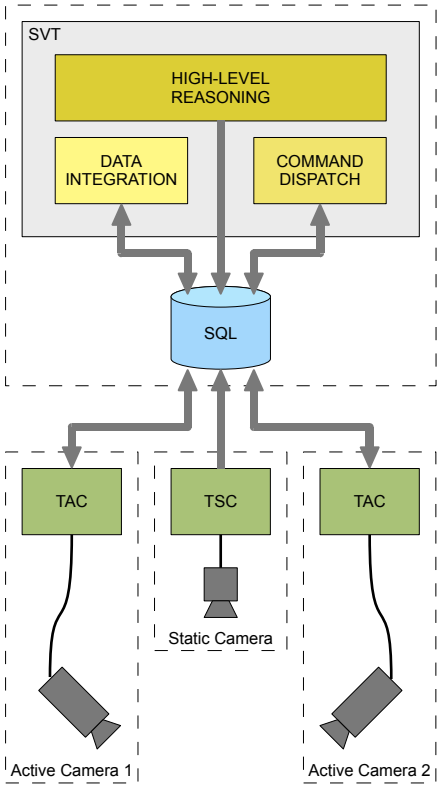


Figure 3.5: System architecture. High level clients such as the data association and control processes are summarised by the supervisor tracker ("SVT").

observations from the appropriate table and populates another with the results. The latter are then used by the control process to derive control commands, which are sent back to the active clients. This sounds like a long round trip time, but unless many clients send image data into the database at the same time, the times are in the order of a few milliseconds, with spurious longer delays [17].

In summary, the key components of our architecture are the following ones:

- a tracking client with a static camera (TSC), provided with a wide-angle lens;
- two tracking clients with actives camera (TAC), comprising a FireWire camera with integrated zoom lens (2× Sony DFW-VL500 and Imagingsource DFK 21BF04-Z2) and a Directed Perception DTU-D46 pan-tilt unit;
- a data fusion process,

- a control process,
- SQL database shared by the above four components; the present implementation comprises 4 tables: one for observations/estimations generated by the trackers and the data fusion process, and another one for stabilized images produced by the TAC, and finally one to store calibration information for all the trackers.

By virtue of the SQL database, all clients can communicate with simple SQL query statements that are sent over TCP/IP.

3.9.1 Passive Visual Sensing

The static camera of our system (TSC) is used for real-time human detection on wide-angle image sequences. This data is used for two purposes: Generation of "Track and Detect" commands, and ground truth creation during if the active clients run in "Autonomous Track and Detect".

To detect targets with the supervisor camera, we use an implementation of the Lehigh Omnidirectional Tracking System (LOTS) algorithm [71]. This is based on a background subtraction technique that uses two grey-scale images of the background and two per-pixel thresholds. The background subtraction well suits our current installation of the static camera because the targets, being observed from the top, do not overlap in this case (see figure 3.6).

3.9.2 Active Visual Tracking

The active clients periodically poll the database to determine if there is a new control command from the supervisor.

There are three different control types implemented:

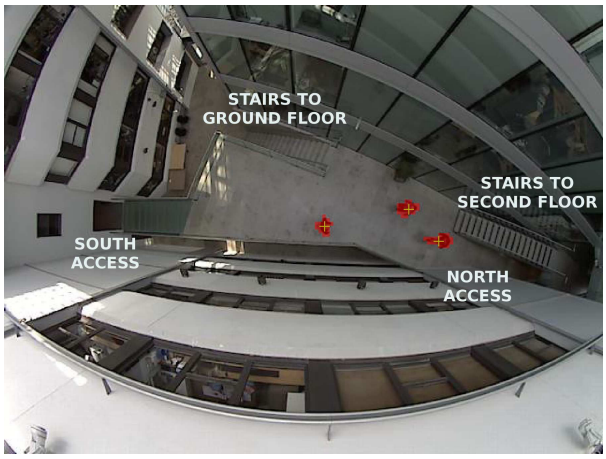
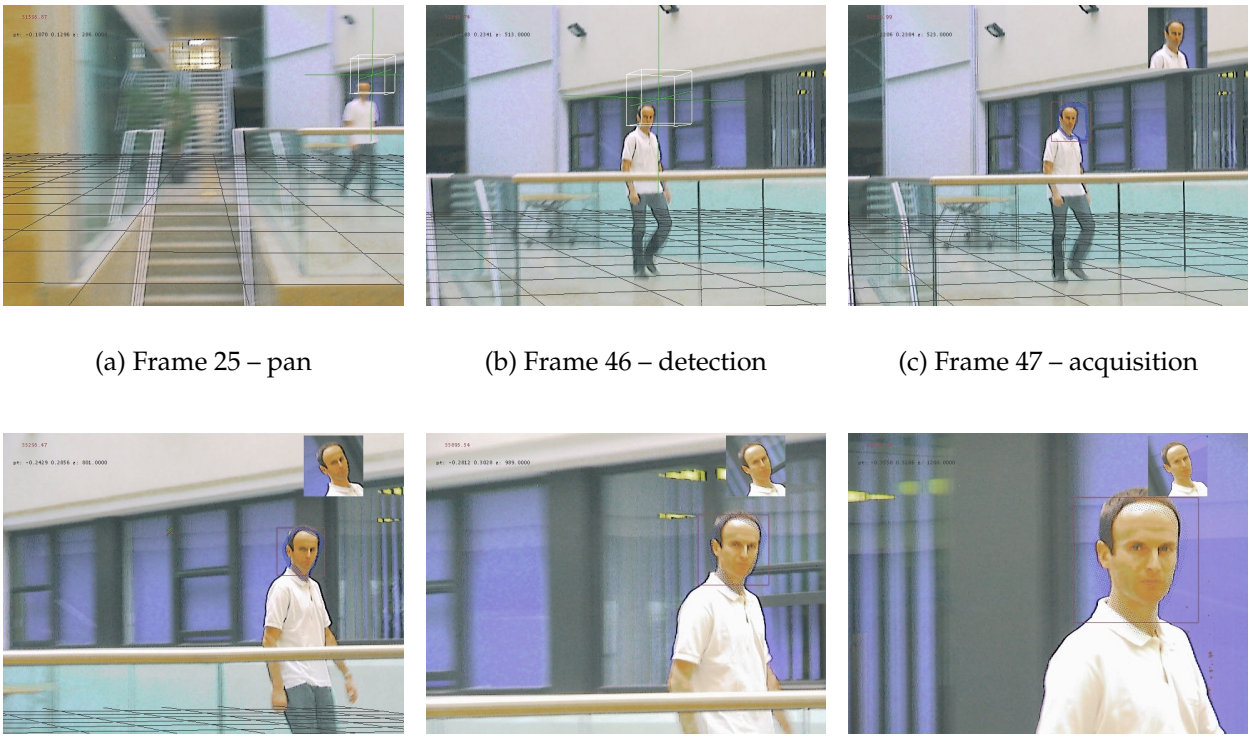


Figure 3.6: Target detection using LOTS background subtraction. The centroid of each red cluster, corresponding to a pedestrian, is marked with a cross.

- Direct pan, tilt, zoom demand. The client adjusts the parameter accordingly. This mode serves for slaving of the cameras only, no data is sent back into the database.

- "Track and detect" a target. The client receives a bounding box in world coordinates and velocity information from the database. This describes the expected region in space where the target was detected when the data was written into the database. The client adjusts the parameters to maximise the projected area of the bounding box, and runs a target detection method (typically the OpenCV implementation of the Viola-Jones boosted face detection algorithm [158]), and initialises a level set tracker on an eventually found target [19]. The client then follows the target and sends the target coordinates into the data base until a new command is received, or the target is lost. While tracking, the zoom setting is adjusted to provide high resolution face data.

  An example of target acquisition and tracking at long distances is shown in figure 3.7. In the first two frames, the white box outlines the latest target position as received from the database. Note the motion blur in the first frame 3.7(a), induced by higher velocity demands for initial centring of the target.



(a) Frame 25 – pan (b) Frame 46 – detection (c) Frame 47 – acquisition

(d) Frame 67 – zoom in (e) Frame 83 – zoom in (f) Frame 115 – max zoom

Figure 3.7: Multi-resolution tracking over long distances. Zoom factors are changed from 1 to 2.5 in frame 25, 4.5 in frames 46 and 47, up to 10 in frame 115. Note the stabilised close-up image in the upper right corner once the target is acquired, and which is stored in the database. To speed up initial acquisition of the target, the active tracker is not always run until full convergence (see face boundary in 47). This can result in subsequent rotation of the close-up image.

- "Autonomous Track and Detect": The client receives an observation region, and a list of targets and target indices which are present in the scene. The client gets to know about where to search for a target, and is the only source for new detections. The detection and tracking needs to be considerably more robust without guidance from the supervisor camera. We use a multithreaded approach to address slower detection and online level set tracking. For this, we employ a background thread which runs a fast implementation of the Histogram-of-Oriented Gradients detector [120], and a foreground thread, which runs the level set tracker, but adds a strong prior.

Table 3.2: Overview of vision processes used in the system

passive tracking	LOTS background separation [24]
data association	Kalman filter tracking and Nearest Neighbour data association
active target acquisition	detection using HOG[120] and Haar features[158]
active tracking	PWP tracking [19]
low level camera control	PID control over velocity estimates
high level camera control	maximisation of mutual information, scheduling strategies

The data fusion process merges the observations using a bank of Kalman filters [16], such that the control process can calculate optimal parameter settings. The control process also provides several methods to select targets. Either using the first come, first serve approach, or using the mutual information based metrics we are going to discuss in the subsequent chapters of this thesis.

Table 3.2 gives an overview of the processes and methods employed by the system.

3.10 Conclusion

In this chapter we gave an overview of the involved objective functions for the control of cameras, their properties and care that needs to be taken when optimising them. We furthermore presented the performance metrics used in other publications which facilitate comparison. Lastly, we gave an overview on the elements of the setup for experiments in this thesis. Whereas this thesis focuses on the objective functions, a considerable amount of work went into the creation of the system. The system is covered in more detail in the publications [17, 166].

Chapter 4

Tracking

In this chapter we address methods for adjusting zoom on a camera while tracking a single target. We focus on related works which take into account uncertainty – one by adjusting the zoom based on the deviation from the expected motion, and the other one by choosing the focal length which best reduces the expected uncertainty.

We point out the weaknesses of each of these methods, and present a new method that is a combination of both, which builds upon the first information-theoretic measure and is made more stable by using a estimate of the innovation. We show in detail how this increases the usable zoom range while still maintaining track, and how the underlying parts and the previous approaches limit the use of potentially available zoom range. Parts of this chapter were published in [135].

4.1 Introduction

In many application areas – such as sport events, surveillance, and patient monitoring – zoom control can be seen as a simple example for arbitration of different interests. One interest is to obtain the maximum resolution of a target to facilitate classification. Examples are identification of people, close-ups to disambiguate specific gestures, or properties such as view direction. The second interest is to minimise the risk of losing a target once it has been detected. Here zoom is an important factor. When a target remains static, the zoom can be safely increased. Once a target starts moving, small mistakes in following the object can result in a loss of sight. For example, following an object with a fixed zoom

telescope is extremely hard once this object begins to move. Intuitively, the optimal setting for the zoom at any instant is the one that best compromises between minimising the chance of losing the target, and maximising the resolution of the target.

4.2 Related Work

There is little work on zoom control using a single camera that addresses the uncertainty in the sensing process. Recently, Al Haj et al. [7] addressed zoom control by employing a standard PD-controller[1], but addresses the uncertainty in the state of the camera parameters by estimating them in the same way as the target's position through an extended Kalman filter. However, uncertainty in the parameter estimate is not used to change the parameters in the control.

The *focal length dilemma* is a term coined by Denzler et al. [46]. It addresses the specific issue of balancing the preference for a greater zoom (or longer focal length) with the risk of losing track of the target. This trade-off between resolution and tracking error has also been addressed by Tordoff and Murray in two separate works. In [154] they considered kinematic uncertainty, while in [153] they considered size preservation. In Fayman et al. [55], zoom selection is driven by keeping the ratio of focal length and distance of the target constant. The control input is either the autofocus sensor of a camera, or the optical flow of the target while the zoom setting is kept constant.

The important difference is that Tordoff and Murray, and Fayman et al. use the constraint of keeping the object within certain limits to the image boundary and size preservation, whereas Denzler et al. make use of the predicted behaviour and expected visibility of the target. The zoom control is different, since an object next to the image border that

[1] A Proportional-Derivative-controller reacts proportionally to the error and the derivative of the error in the desired values

is moving closer to the centre of the field of view is unlikely to get lost, and hence does not require zooming out.

4.3 Robust Entropy-based zoom selection

In the next two sections, we recapitulate the approaches of Denzler et al.[46] and Tordoff and Murray [154]. An analysis in section 4.4 on a simple model points out their strengths and weaknesses, we then combine the former into our model described in section 4.4.1. The chapter finishes with an in-depth discussion of the zoom limit imposed by the noise characteristics.

4.3.1 Entropy-based zoom selection

The zoom selection method by Denzler et al. for tracking a single target takes an information-theoretic approach, phrased as a decision problem. The decision to make is the choice of an action, which in this case are the relevant observation parameters of the system, e.g. focal length and view direction.

Figure 4.1 shows a sketch of this approach. The action (i.e. observation parameters) for the next time step is chosen as follows: before making an observation $\mathbf{o}_k$ at time k, we select the best parameters (or action) $\mathbf{a}_k$ for this observation. This choice is based on the latest estimate $\mathbf{x}_k^-$ of the state of the target. This action $\mathbf{a}_k$ summarises all different parameters; in the case of controlling a single camera these are pan, tilt, and zoom. After an observation has been made, the predicted state $\mathbf{x}_k^-$ is updated using Bayes' rule. Assuming the system has a Markovian property, the resulting posterior is used as current estimate.

We want to choose the parameter that yields the most informative observation to the state, i.e. the observation that maximally reduces uncertainty about the state. The least

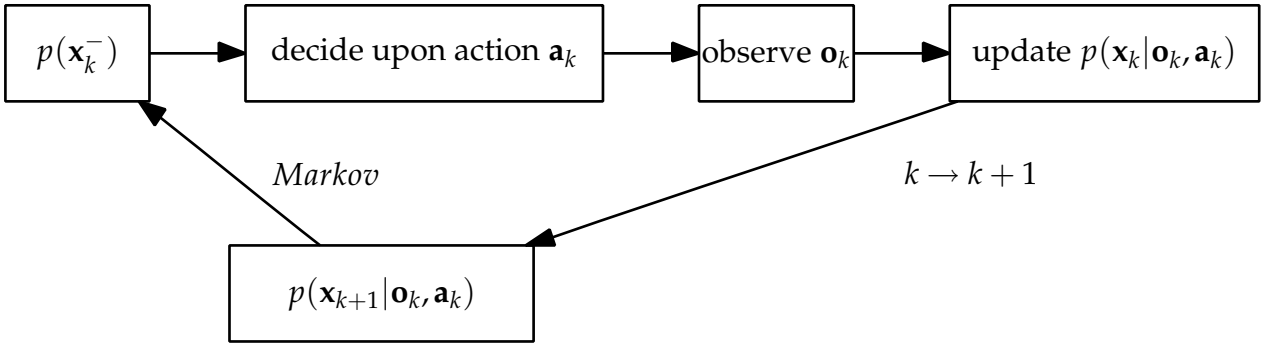


Figure 4.1: Overview of the sequential decision process. Starting with a current estimate of the state of the target (top left), the optimal observation parameters are chosen. After the observation (e.g. target position or loss of track), the state is updated.

useful is the parameter which gives an observation that is independent of the state, e.g. when it does not influence the certainty of the position of the target. Instead, a suitable objective function should reward an observation that is as discriminative for the estimation of the state as possible. For example, observing a person frontally usually yields more information about the person's identity than from behind. If we were to find a person's identity, a sensible choice of observation angle would be a frontal one. Finally, as the target behaviour and measurement process are both subject to randomness, we need to take the expectation over all possible observations and all possible states.

From information gain to conditional entropy

These considerations naturally lead to the mutual information between the next observation and the state in the formulation from section 3.3.6. Here the random variables in question are the predicted state $\mathbf{x}_k^-$ and corresponding predicted observation $\mathbf{o}_k^-$ at time k:

$$I_{\mathbf{a}_k}(\mathbf{x}_k^-;\mathbf{o}_k^-) = H_{\mathbf{a}_k}(\mathbf{x}_k^-) - H_{\mathbf{a}_k}(\mathbf{x}_k^-|\mathbf{o}_k^-) \tag{4.1}$$

The optimal parameter is then

$$\mathbf{a}_k^* = \arg\max_{\mathbf{a}_k} I_{\mathbf{a}_k}(\mathbf{x}_k^-;\mathbf{o}_k^-) \tag{4.2}$$

As shown in section 3.5.1, if the parameter $\mathbf{a}$ does not influence the uncertainty in the current state estimate, the maximisation in equation 4.2 is equivalent to a minimisation of the conditional entropy:

$$\mathbf{a}_k^* = \arg\min_{\mathbf{a}} H_{\mathbf{a}}(\mathbf{x}_k^- | \mathbf{o}_k^-) \tag{4.3}$$

The result can be interpreted as the action (or set of parameters) which results in the least remaining uncertainty about the state. This minimisation is the formulation proposed by Denzler et al. in [46].

Tracking using Kalman Filter

This zoom selection process is now put into a Kalman filter context (see appendix C).

As a short summary, we denote $\mathbf{x}_k^+$ for a state which has been updated with the latest observation $\mathbf{o}_k^-$, and $\mathbf{x}_k^-$ the state which has been predicted by the Kalman filter, but not updated because no observation has yet been made. The analogous notation is used for the covariance matrices, $\mathbf{P}_k^+$ and $\mathbf{P}_k^-$, respectively. The innovation $\boldsymbol{\nu}_k = \mathbf{o}_k - \mathbf{o}_k^-$ updates the prediction according to the Kalman gain $\mathbf{K} = \mathbf{P}_k^- \mathbf{H}(\mathbf{a})^T (\mathbf{R}_k + \mathbf{H}(\mathbf{a}) \mathbf{P}_k^- \mathbf{H}(\mathbf{a})^T)^{-1}$:

$$\mathbf{x}_k^+ = \mathbf{x}_k^- + \mathbf{K}_k \boldsymbol{\nu}_k \tag{4.4}$$

and the covariance according to:

$$\mathbf{P}_k^+ = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k^- \tag{4.5}$$

The observation matrix $\mathbf{H}(\mathbf{a})$ reflects the dependency on the current zoom or other parameter $\mathbf{a}$, the covariance matrices $\mathbf{R}_k$ and $\mathbf{Q}_k$ describe the Gaussian noise of the measurement and the process, respectively.

The results are easily augmented to apply to non-linear extended Kalman filters, but for the sake of clarity we use the notation of the linear version.

The conditional entropy in equation 4.3 can be written as

$$H_{\mathbf{a}}(\mathbf{x}_k^-|\mathbf{o}_k^-) = -\iint_{-\infty}^{\infty} p_{\mathbf{a}}(\mathbf{x}_k^-, \mathbf{o}_k^-) \log(p_{\mathbf{a}}(\mathbf{x}_k^-|\mathbf{o}_k^-))\, d\mathbf{x}_k^-\, d\mathbf{o}_k^- \tag{4.6}$$

in the continuous form.

Since in a Kalman filter context all random variables are assumed to be Gaussian distributed, the entropy of such a variable $\mathbf{x} \in \mathbb{R}^n$ with $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{P})$ reduces to (as shown in section 3.3.4)

$$H(\mathbf{x}) = \frac{n}{2} + \frac{1}{2}\log((2\pi)^n|\mathbf{P}|), \tag{4.7}$$

showing that the uncertainty of $\mathbf{x}$ depends only on its covariance matrix, not on the actual updated mean.

The conditional entropy in equation 4.6 is obtained by averaging over the domain of all observations. This domain can be split into the area inside (v) and outside ($\neg v$) the image. When the target is inside of the image, an observation is made and the state can be updated to $\mathbf{x}_k^+$, resulting in the entropy $H(\mathbf{x}_k^+)$. If the target is outside of the image, the entropy $H(\mathbf{x}_k^-)$ is obtained from the predicted state only. Both these entropies are independent of the actual observation (see equation 4.7), only on the assumption whether the target will be observed or not. When rewriting the integral in 4.6 as

$$\begin{aligned} H_{\mathbf{a}}(\mathbf{x}_k|\mathbf{o}_k) = -\int_{-\infty}^{\infty} & p_{\mathbf{a}}(\mathbf{o}_k) \\ & \int_{-\infty}^{\infty} p_{\mathbf{a}}(\mathbf{x}_k|\mathbf{o}_k) \log(p_{\mathbf{a}}(\mathbf{x}_k|\mathbf{o}_k))\, d\mathbf{x}_k\, d\mathbf{o}_k \end{aligned} \tag{4.8}$$

and applying previous observations, this simplifies to

$$\begin{aligned} H_{\mathbf{a}}(\mathbf{x}_k|\mathbf{o}_k) &= w(\mathbf{a})H(\mathbf{x}_k^+) + (1 - w(\mathbf{a}))H(\mathbf{x}_k^-) && (4.9)\\ &= w(\mathbf{a})(H(\mathbf{x}_k^+) - H(\mathbf{x}_k^-)) + H(\mathbf{x}_k^-) && (4.10)\\ &\propto w(\mathbf{a})(\log|\mathbf{P}_k^+| - \log|\mathbf{P}_k^-|) + C && (4.11)\\ &\propto w(\mathbf{a})(\log|\mathbf{I} - \mathbf{K}\mathbf{H}_{\mathbf{a}}|) + C, && (4.12) \end{aligned}$$

where the entropies are averaged by the likelihood of making an observation $w(\mathbf{a})$, i.e. the chance of the target being within the observation region. It is worthwhile to emphasise that none of the parts of the criterion in equation 4.3 depends on future observations, since the term

$$w(\mathbf{a}) = \int_v p_{\mathbf{a}}(\mathbf{o}_k)\, d\mathbf{o}_k = \int_v \int_{\mathbf{x}_k} p_{\mathbf{a}}(\mathbf{o}_k|\mathbf{x}_k) p(\mathbf{x}_k)\, d\mathbf{x}_k\, d\mathbf{o}_k \tag{4.13}$$

is influenced only by the current likelihood of the observation, which is a Gaussian distribution

$$p_{\mathbf{a}}(\mathbf{o}_k^-) \sim \mathcal{N}(\mathbf{H}_{\mathbf{a}}\mathbf{x}_k^-, \mathbf{R} + \mathbf{H}_{\mathbf{a}}\mathbf{P}_k^-\mathbf{H}_{\mathbf{a}}^T) \tag{4.14}$$

about the projected mean of the state prediction.

This formulation of visibility is inspired by [46], and further details can be found there. For axis aligned Gaussian distributions $p(\mathbf{o})$, the integral in equation 4.13 has a simple solution in the form of the error function. In the case of multiple dimensions (e.g. actual image coordinates), a product of error functions is obtained, where each error function is evaluated at the axis limits of the image. The assumption of axis alignment of the distribution of the observation is usually given in a surveillance scenario (i.e. pedestrians are observed mostly upright).

Example

We give an example for a simple, contrived system, which nonetheless already exhibits the relevant features we are going to discuss. Imagine a target in front of a camera. Both the target as well as the measurements shift in a random, erratic way. The measurement is magnified according to the focal length f chosen, but we assume that the measurement error remains the same. This is a valid assumption if the measurement is extracted by some vision process (e.g. corner detection).

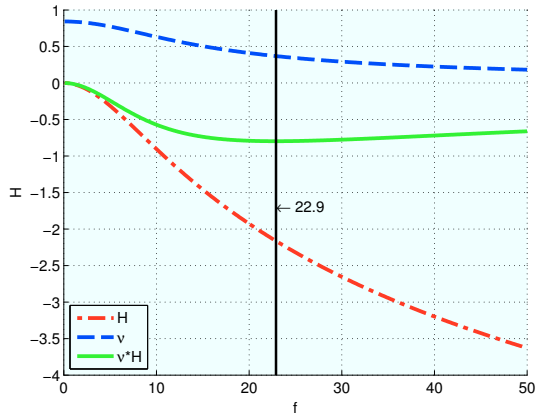


Figure 4.2: Exemplary comparison of uncertainty reduction over varying focal length. The positional uncertainty keeps decreasing (red/dash-dotted), but the chance of making a successful observation goes down (blue/dashed), yielding a minimum. This is the optimal zoom setting for the next observation.

Figure 4.2 shows the influence of the two factors in equation 4.12 on the overall, expected entropy. The expected visibility of the point approaches 0 for rising focal lengths, as the uncertainty in the position of the target is amplified by the observation parameters in equation 4.14.

The uncertainty in the position drops with higher magnification, as a successful observation at this level serves better to locate the target. However, the likelihood of successfully making an observation nears zero with rising focal length, bounding the expected entropy at a defined minimum. The more unlikely making a successful observation becomes, the more the expected entropy approaches the predicted state's entropy.

Another example for the behaviour of this entropy based zoom control is shown in figure 4.3, which shows the development of the conditional entropy H and the probability of making an observation, w, as well as the images that result from the choice of the zoom parameter.

The first frame (a) shows the 1σ - covariance ellipse of the location right after initialisation on a newly detected target. Due to the high initial uncertainty of the location, the

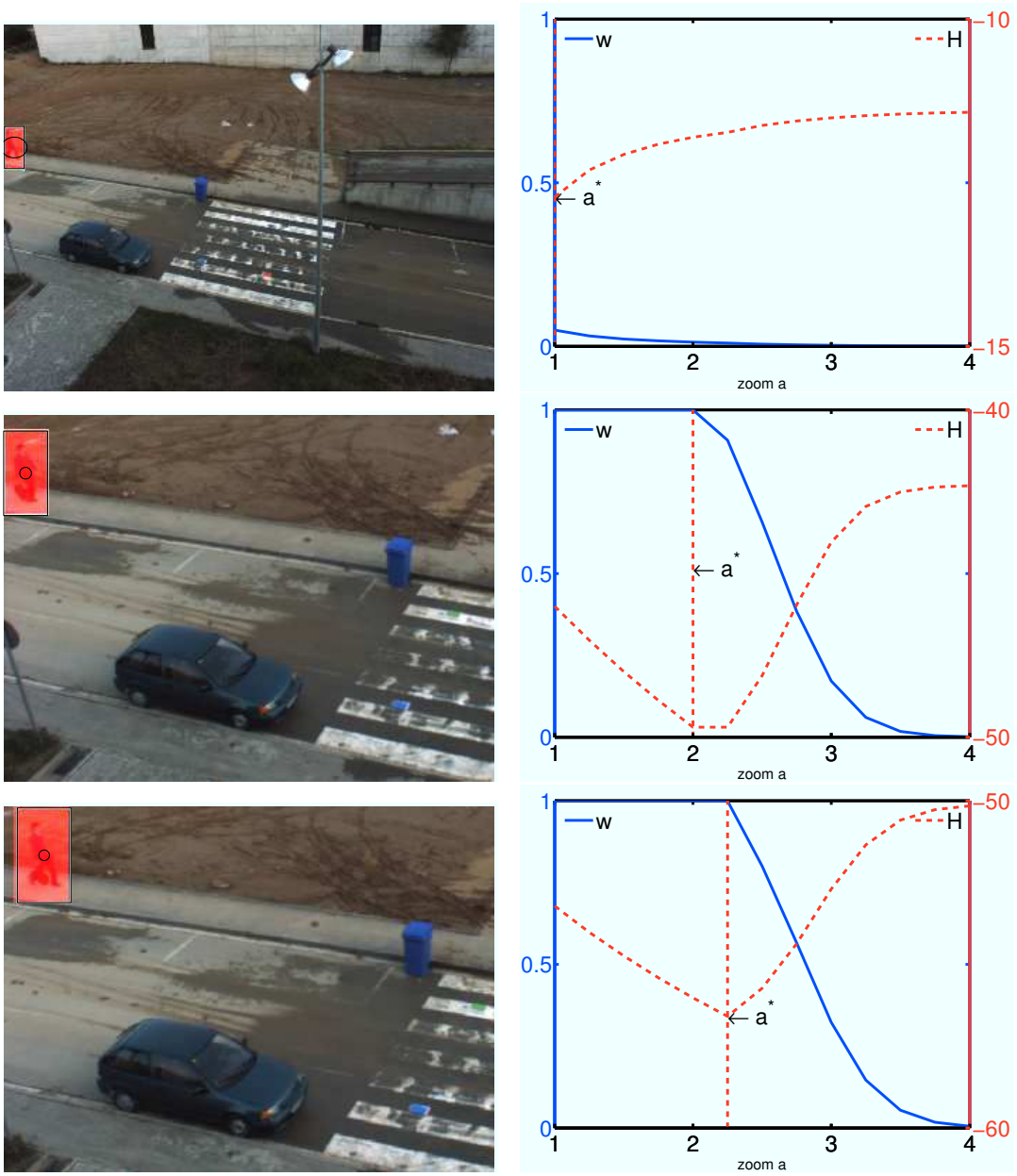


Figure 4.3: Visibility term w and entropy H for given levels of zoom for frames 442, 447 and 458 of the HERMES Outdoor sequence, camera 1. (top) after the initialisation of a Kalman filter on a new object. (centre) The covariance gets smaller, and the confidence in the visibility rises. The camera zooms in. (bottom) The camera pans to follow the object.

probability of making an observation is highest when not zooming in. The fifth frame (b) shows the decreased covariance ellipse, and that the confidence in the making an observation in the next frame rises. The camera zooms in, but is limited by the visibility of the bounding box of the target. If the camera zoomed in too far, the bounding box would be cropped. In the 16th frame (c), the camera zooms in further and starts panning to follow the object. Once the target moves towards the centre of the frame, the maximum zoom will be selected.

The straightforward intuition behind this type of control for setting the zoom of a camera is that the parameter chosen arbitrates the risk of losing the target from a zoom that is too high, with the improved accuracy of tracking at higher zoom levels.

4.3.2 Innovation based zoom selection

In the approach of Tordoff and Murray [154] – who also use a Kalman filter to track an object – the covariance of the innovation is used to specify a confidence interval ζ on the fixation error ν, such that the fixation error is required to remain below the observation boundaries ψ:

$$p(|\boldsymbol{\nu}_k| < \psi) \geq \zeta. \tag{4.15}$$

For a confidence of $\zeta = 1 - 10^{-6}$, this results in the zoom rule

$$f_{k+1}^2 \approx \frac{\psi^2}{24||\operatorname{covar}[\boldsymbol{\nu}]_k||_2}. \tag{4.16}$$

Note that the matrix $2-$norm yields the largest uncertainty in any direction. The covariance of the innovation is estimated by keeping a running average:

$$\operatorname{covar}[\overline{\boldsymbol{\nu}}]_k = \gamma \boldsymbol{\nu}_k \boldsymbol{\nu}_k^T + (1-\gamma)\operatorname{covar}[\overline{\boldsymbol{\nu}}]_{k-1}. \tag{4.17}$$

We would like to point out that this is the sample innovation covariance, not the one produced by the Kalman filter.

Tordoff also pointed out that a change in zoom affects the fixation error and the Kalman filter dynamics, if the measurement error is dominated by zoom-independent noise, e.g. when the only source of positional error is from a pixel based localisation method, ignoring errors in positioning the camera and setting the focal length.

Tordoff investigated the influence of zoom in a tuned filter, i.e. which performs optimally at a given zoom setting. To keep balance between measurement and process noise in such a tuned filter, the change in the observation model has to be taken into account explicitly, and requires that the innovation, state covariance and process noise be scaled inversely to the zoom:

$$\mathbf{P}_k^- = (f_k^2 / f_{k+1}^2)\mathbf{P}_k^- \quad \mathbf{Q}_{k+1} = (f_k^2 / f_{k+1}^2)\mathbf{Q}_k \tag{4.18}$$

$$\boldsymbol{\nu}'_k = \boldsymbol{\nu}_k / f_k \tag{4.19}$$

This external change to the covariance is not used in the previous approach. Instead, the filter is tuned in such a way that it keeps track at the highest possible zoom level, thus overestimating the measurement noise when zoomed out, resulting in oversmoothing of the estimate.

The adjustment of the process noise according to the focal length as proposed by Tordoff must take place after the zoom selection process with the entropy based approach, as the aim of the former method is to scale the process noise such that it remains constant despite changes in the focal length. An example is given in figure 4.4, where the entropy of the updated covariance is increasing with zoom, yielding a minimum at the lowest zoom level. Furthermore, measurement noise that increases with focal length effectively decreases the optimal zoom range. This means that if the influence of the variable measurement noise is too great, the reduction in entropy due to improved localisation disappears.

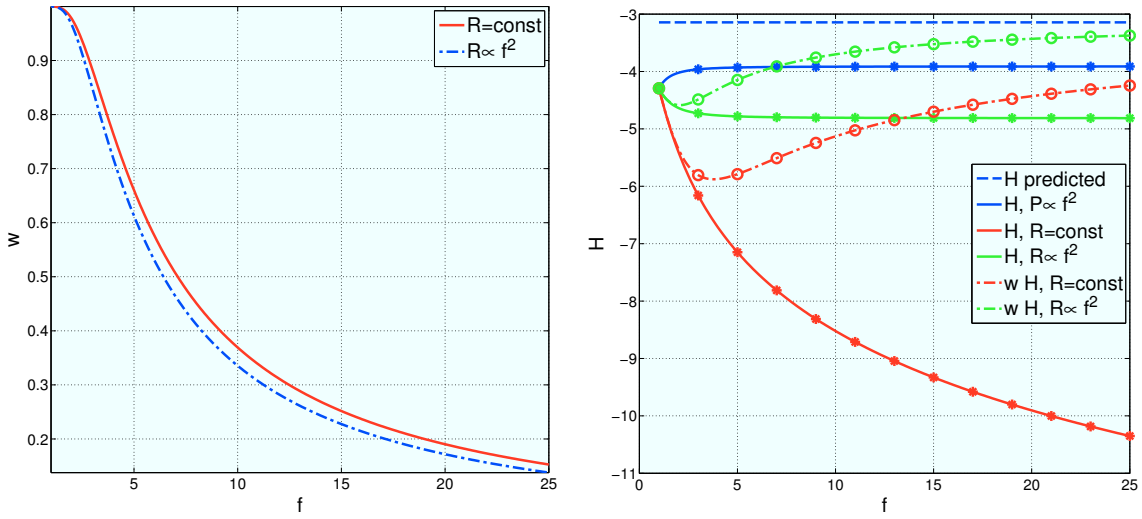


Figure 4.4: Influence of focal length dependent observation noise on visibility term w and entropy H. The updated (solid lines, starred) and expected entropy (dash-dotted, circled) are shown. The variable observation noise is modelled as R and $0.5(R + f^2R)$ and results in a lower optimal focal length. The solid blue line shows the influence of a process noise adjustment on the objective function. This is the proposal by Tordoff to keep the filter tuned. If this is done *before* the zoom selection, this will result in the minimum focal length, as the chance of observation is highest, and the covariance smallest. The constant predicted entropy is shown as a dashed blue line, and is the limit of the expected entropies for no visibility.

4.4 Analysis

A vital factor to the success of Denzler's entropy-based zoom control approach described in section 4.3.1 is that the Kalman filter correctly models the movement of the object and the noise characteristics of the motion model. While it is comparatively simple to model the behaviour of inanimate objects, this is much harder or can be impossible in the case of living beings, or objects operated by humans. It certainly is possible to find some upper bounds on the maximum velocity of pedestrians, but this might be limiting performance if the object under scrutiny is a speeding car. We therefore should expect a failure of the model, and investigate what happens if the model is wrong.

To compare the performance of zoom control by the two approaches, we run several variants of the filter on synthetic data. For clarity and comparability we restrict ourselves

to a simple model which has been introduced by Tordoff and Murray [154]: A line-camera is tracking an object at a constant distance, with a supposedly constant angular velocity. The camera is rotatable and the predicted position is used to keep the tracked object in the centre of the image.

The state of the object is described as $\mathbf{x} = \begin{pmatrix} \phi & \dot{\phi} \end{pmatrix}^T$ with a linear motion model $\mathbf{x}_{k+1} = \mathbf{F}\mathbf{x}_k + \mathbf{u}_k + \mathbf{q}_k$. The plant model for a given discrete time step Δt is [13]

$$\mathbf{F} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \tag{4.20}$$

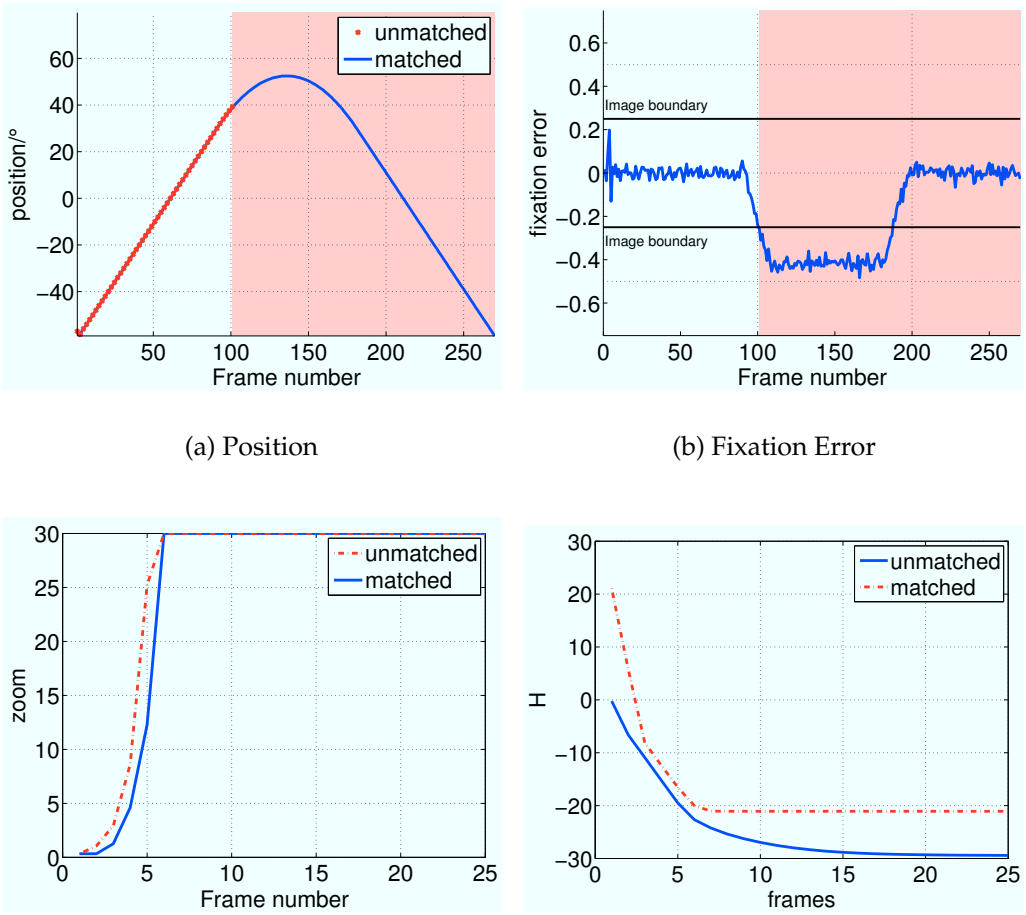
The known input $\mathbf{u}_k$ is the direction the camera is looking: $\mathbf{u}_k = \begin{pmatrix} -\theta & 0 \end{pmatrix}^T$, and the camera direction θ is set to the predicted position ϕ of the object. The process noise $\mathbf{q}_k$ is a zero mean Gaussian noise sequence with covariance $E[\mathbf{q}_k \mathbf{q}_k^T] = \mathbf{Q}_k$, which approximates the size of the unmodelled acceleration $\ddot{\phi}$. The observation model assumes a small angular error, hence a linear model

$$\mathbf{o}_k = \mathbf{h}^T \mathbf{x}_k + \mathbf{r}, \tag{4.21}$$

with $\mathbf{h} = \begin{pmatrix} f & 0 \end{pmatrix}^T$ suffices. In this observation model the value f is the zoom value, or action $\mathbf{a}$, as described in the section on entropy based control. A zero mean Gaussian noise sequence $\mathbf{r}$ models the observation noise with a covariance of $E[\mathbf{r}\,\mathbf{r}^T] = \mathbf{R}$.

The object motion starts at an angle of $-60°$ with constant velocity of $30°/s$. Once the target reaches $60°$, the object accelerates with $-20°/s^2$ until it attains its final velocity of $-30°/s$. It basically moves from left into, partly across, and back out of the field of view. A plot of this trajectory is shown in figure 4.5(a). The initial velocity for the state estimate is set to zero, whereas the position is initialised to the actual ground truth value. The image border is arbitrarily set to $-0.25 \ldots 0.25$. In all experiments we assume an observation noise of four percent of the image width ($\mathbf{R} = \sigma_r^2 = 0.02^2$).

To demonstrate the behaviour of the zoom control algorithms, we let the filter run once with a process noise with standard deviation of $20°/s^2$, and a second time with a hundredth of this, making it a biased estimator, or unmatched filter.



(a) Position

(b) Fixation Error

(c) Zoom levels

(d) Entropy

Figure 4.5: Tracking behaviour of Kalman filters with entropy based zoom control over time. The biased Kalman filter (red, stippled) should follow the ground truth position of the target in (a) as the matched, unbiased filter does, but results in fixation error shown in (b) and loses track (shaded area). (c) shows that the biased filter is overconfident and zooms in on the target earlier than in the matched case. (d) shows the remaining uncertainty in the state estimate per time step.

The results of the entropy based zoom control are shown in figure 4.5. The first plot, (a), details that matched filters, as well as unmatched, have approximately the same performance; i.e. the position follows the ground truth, but the unmatched filter loses track once the target passes the image boundary (shown in (b)), which happens a few frames

after the velocity changes. Figure 4.5(c) and (d) show how the entropy is minimised in each frame in both versions, and the zoom is increased up to the maximum value due to the decreased uncertainty. The entropy of the unmatched filter sinks faster than for the matched one, due to the increased trust in the motion model. This leads to a faster increase of focal length, as well.

The loss of track can be explained by investigating the mean conditional entropy term in equation 4.12. Since $\mathbf{P}_k^-$ stays constant during the minimisation, the relevant part can be rewritten as

$$H_{\mathbf{a}}(\mathbf{x}_k|\mathbf{o}_k) = c_2 + w(\mathbf{a}) \log|\mathbf{I} - \mathbf{K}\mathbf{H}_{\mathbf{a}}| \tag{4.22}$$

$$= c_2 + w(\mathbf{a}) \log\left|\mathbf{I} - \mathbf{P}_k^- \mathbf{H}_{\mathbf{a}}^T (\mathbf{R} + \mathbf{H}_{\mathbf{a}} \mathbf{P}_k^- \mathbf{H}_{\mathbf{a}}^T)^{-1} \mathbf{H}_{\mathbf{a}}\right| \tag{4.23}$$

c_1, c_2 are constants irrelevant to the minimisation, and $\mathbf{K}$ in equation 4.22 is the Kalman filter gain, which expands to equation 4.23. Our model has a one dimensional observation space; i.e. $\mathbf{R}$ and the observations are scalar. This simplifies the equation even further:

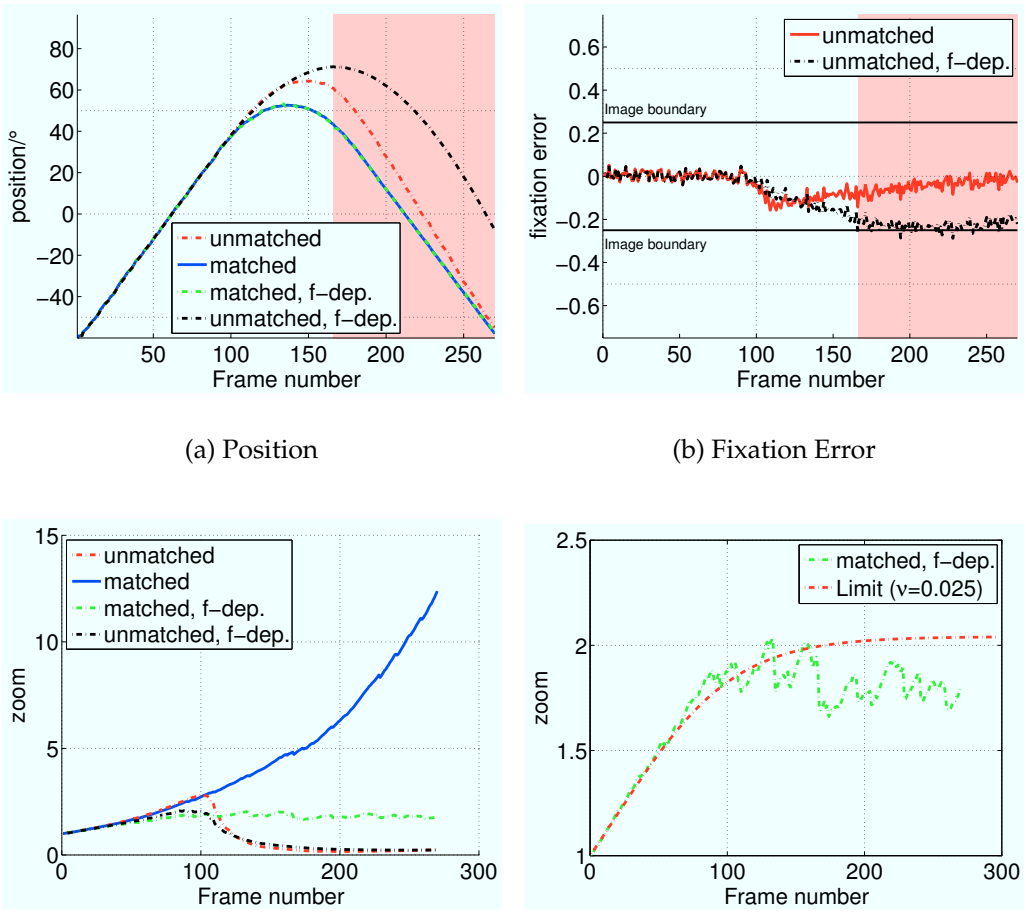
$$H_{\mathbf{a}}(\mathbf{x}_k|\mathbf{o}_k) = w(\mathbf{a}) \log\left|\frac{1}{1 + \mathbf{h}\mathbf{P}_k^- \mathbf{h}^T/\sigma_r^2}\right| + c \tag{4.24}$$

Since in this model the predicted observation is always 0, the visibility factor $w(\mathbf{a})$ reduces to

$$w(\mathbf{a}) = \mathrm{erf}\left(\psi \sqrt{2(\sigma_r^2 + \mathbf{h}\mathbf{P}_k^- \mathbf{h}^T)}^{-1}\right). \tag{4.25}$$

Both factors of the minimisation criterion depend on *a priori* values only, hence the inability to change the behaviour of the zoom control if the target is about to leave the observable region. This points out that entropy-based control requires a fitting process model and process noise characteristics.

Figure 4.6 shows the behaviour of the covariance based zoom control, both with and without the zoom adaption of the filter dynamics as stated in equations 4.18-4.19. The



(a) Position

(b) Fixation Error

(c) Zoom levels

(d) Bounded zoom

Figure 4.6: Tracking behaviour of Kalman filters with covariance based zoom control over time. The resulting estimated positions are shown in (a), where the unmatched/biased and zoom-dependent ('unmatched, f-dep.') filter loses track (shaded area). The biased Kalman filter results in fixation error shown in (b), and keeps track if adapted to zoom change. (c) shows the change in zoom levels for all variants, and that the zoom range is bounded. This is shown by comparison with the ideal case in (d).

fixation error in figure 4.6(b) shows the loss of track of the unmatched, non-adapted filter. The matched versions of the filter keep track, as well as the adapted version of the unmatched filter. The zoom rises steadily, yet slowly in the adapted and matched filter case. This behaviour is emphasised in the zoom-dependent and matched filter case. Whereas the filter keeps track, the covariance criterion is too restrictive. As can be seen in figure 4.6(c), the zoom rises much slower than in the case of entropy based control, and never reaches the maximum. The speed and maximum of the zoom are actually bounded

by the running average imposed on the innovation covariance: When assuming a stationary white noise sequence for $\boldsymbol{\nu}$, i.e. covar$[\boldsymbol{\nu}]_k = \boldsymbol{\nu}_0^2$, the difference equations 4.16 and 4.17 result in

$$f_{k+1}^2 = \frac{A}{(A - \boldsymbol{\nu}_0^2)(1-\gamma)^k + \boldsymbol{\nu}_0^2} \quad A \approx \frac{\psi^2}{24}, \tag{4.26}$$

which is limited by $f_\infty^2 \approx \frac{\psi^2}{24\boldsymbol{\nu}_0^2}$. Figure 4.6(d) details this behaviour for a non-adapted, matching filter. The covariance is kept constant and set to the average value of the innovation covariance. This problem is less apparent in the matched and adapted case, because here the innovation covariance is inversely scaled with the current focal length before updating the last estimate. Still, the increase of zoom depends on the damping γ.

The second restriction of this approach is due to the use of observations for control. The observations do not reflect the expected dynamics of the object. Since the zoom rule specifies a confidence interval on the innovation, the zoom is simply reduced to keep the object in the centre of the observation region. This is not necessarily the best thing to do. Consider the setting given by Tordoff and Murray, where a cameraman observes a gnu on a veld. If the cameraman knows with confidence that the gnu will move to the right, because the gnu is running already, she will certainly not zoom far out if the gnu is at the left image boundary.

4.4.1 Combination of Previous Methods

The problems of the two approaches are loss of track in the entropy-based control, and both a slow increase and limit of zoom in the innovation covariance based method. A naïve solution would be to choose the minimum of either approaches, but since the entropy based approach is not bound by actual fixation errors, the zoom setting would simply be imposed by the innovation covariance approach. Also, the control rule addresses one observation parameter – zoom – only, and restricts the selection to a radial

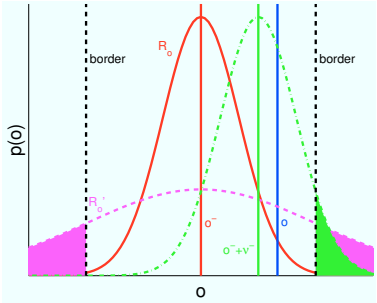


Figure 4.7: Inclusion of the innovation term in entropy based control. Shifting the mean (green, stippled) or increasing the covariance (cyan, dashed) decreases the visibility part of the objective function, demanding to zoom out

observation area. The entropy based control rule instead can address multiple observation parameters and non-uniform observation domains. We therefore discuss two improvements to the entropy based approach which make it more robust to wrong filter dynamics.

Basically, there are two ways to influence the zoom selection process of the entropy method. One is to rectify the false estimation of the mean value, and the other one is to increase the uncertainty according to the actual innovation sequence. Both these approaches are sketched for a constant observation parameter in figure 4.7. In the case of a matching filter, the predicted observation $\mathbf{o}_k^- = \mathbf{H}_a \mathbf{x}_k^-$ coincides with the actual process. Now assume an unmatched filter, which gets measurements around $\mathbf{o}_k$. The first method we propose follows the innovation sequence with an innovation estimate $\boldsymbol{\nu}_k^-$. The second approach adjusts the covariance of the observation, $\mathbf{R}_k$, according to the measured innovation covariance. Both approaches reduce the likelihood of making an observation, which is given by the total area of the distribution function within the observation region. To increase this likelihood (and to decrease the area outside of the image borders) a smaller zoom value must be chosen.

4.4.2 Estimation of the innovation sequence

In the first approach, we incorporate the fixation error directly into the visibility term $w(\mathbf{a})$ in equation 4.13 by adding an innovation term to the observation likelihood:

$$p_{\mathbf{a}}(\mathbf{o}_k) \sim \mathcal{N}(\mathbf{H_a}\mathbf{x}_k^- + \boldsymbol{\nu}, \mathbf{R} + \mathbf{H_a}\mathbf{P}_k^-\mathbf{H}_\mathbf{a}^T) \tag{4.27}$$

When there are fixation errors, the chance of making an observation will decrease, and the zoom is decreased. This leaves the zoom control intact as long as the filter is matched; i.e. results in a faster increase of zoom demand. For this approach, the innovation of the next time step $\boldsymbol{\nu}_{k+1} = \mathbf{o}_{k+1} - \mathbf{H}\mathbf{x}_k^-$ is added to the predicted observation. This innovation is not yet available and needs to be from the same parameter setting as the one being currently evaluated. By making use of the pseudo-inverse $\mathbf{H}_k^+$ at time k we approximate:

$$\mathbf{o}_{k+1} \approx \mathbf{H}_{k+1}\mathbf{H}_k^+\mathbf{o}_k \tag{4.28}$$

In the case of the observation model used in the experiments above, this gives $\mathbf{o}_{k+1} = \frac{f_{k+1}}{f_k}\mathbf{o}_k$, which is the same result as obtained for the single parameter case of zooming.

Note that the introduction of the innovation in the visibility term makes the entropy highly dependent on the last observation. Similar to the covariance based approach, we use a running average. Contrary to the limit of zoom speed imposed by the running average, this only addresses the observability term $w(\mathbf{a})$, which is close to 1 if the filter is matched. Since the observability term is influenced, small changes have a huge influence and the control reacts with zooming out once the target is nearing the border of the observation region.

When a target is observed for the first time, the zoom is bounded by the initial value of the state covariance matrix. Additionally, the predicted observation value can be initialised with a maximum of the observation domain, i.e. with the left border if the target

enters the scene from the left. This initialisation is of course more difficult in observation spaces of higher dimensions.

Figure 4.8(a) shows the performance of the filter modified accordingly, once in a focal length dependent (red), and independent version (green). For both versions of the filter, the zoom level rises as long as the filter is matched, and reaches the maximum, as seen in inset 4.8(c). This behaviour can be explained by the fourth plot (d) of that same figure. The entropy continually sinks, and the visibility term rises. The visibility attains its maximum as soon as the initialisation phase of the running average has passed. Once the motion pattern of the target changes, the visibility drops, and the zoom is set to minimum.

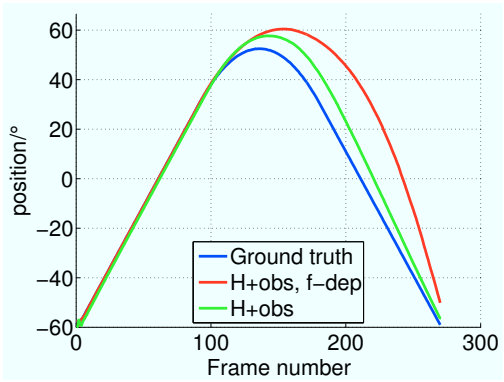
The zoom change affects the state estimate shown in figure 4.8(a), but with a considerable delay in the case of the focal length dependent version (red). This is also portrayed by slow increase of the entropy in (d). This behaviour stems from the loss of importance of the measurement error when compared to the state covariance matrix. The latter is amplified by the observation model, and data observed at a lower zoom has a higher influence on the state estimate. When f is rising, the constant measurement noise loses importance. When f is decreasing, the gained trust in the process model is slowly decreased.

This change in filter dynamics results in a state estimate (in figure 4.8(a)) with a smaller error than in the original version. Once the target accelerates, the entropy based zoom criterion yields a smaller setting, keeping the target within the image boundary. As soon as the filter has recovered, the zoom is increased again.

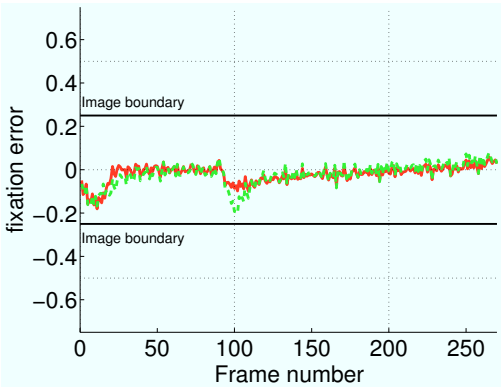
Integration of innovation covariance

Unfortunately, the previous method is only usable under the condition that $\mathbf{H}_{k+1}\mathbf{H}_k^+$ has full rank[2] and suffers from the need of appropriate initialisation. The second method we

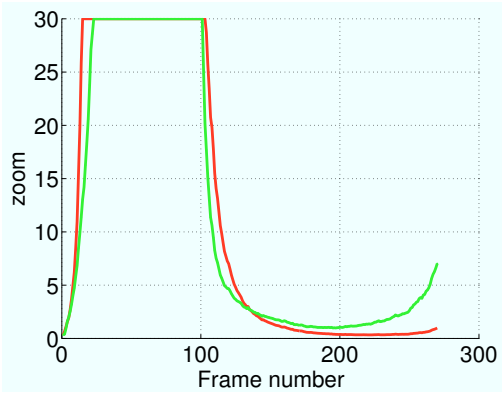
[2] For the case of a rotating camera in the example of this chapter, the observation matrices $\mathbf{h}_1 = [0\,1]^T$ and $\mathbf{h}_2 = [1\,0]^T$ can be contrived. Any observation $\mathbf{o}_1$ would get mapped to 0.



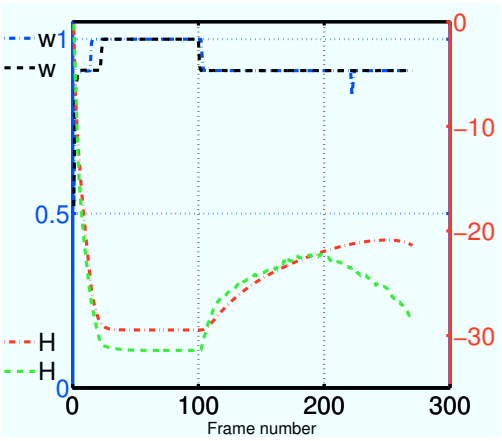
(a) Position



(b) Fixation error



(c) Zoom levels



(d) Entropy H and visibility term w (blue: f-dep.)

Figure 4.8: Position, fixation error, zoom levels and entropy of entropy-based zoom control, incorporating the predicted observation only (red), and combined with the focal length adaption (green)

propose avoids these pitfalls and incorporates the covariance of the fixation error into the entropy term in equation 4.12, arguing similarly to Mehra [101], that the innovation sequence contains the missed information useful for innovation adaptive estimation.

We keep track of the innovation covariance independent of the varying observation parameters, which thus needs to be normalised by the pseudo-inverse of the observation model. The resulting matrix is finally used to update a running average

$$\mathbf{C}_k = \gamma \mathbf{H}_k^{+} \mathbf{o}_k \mathbf{o}_k^T \mathbf{H}_k^{+T} + (1-\gamma)\mathbf{C}_{k-1}, \tag{4.29}$$

similar to equation 4.17.

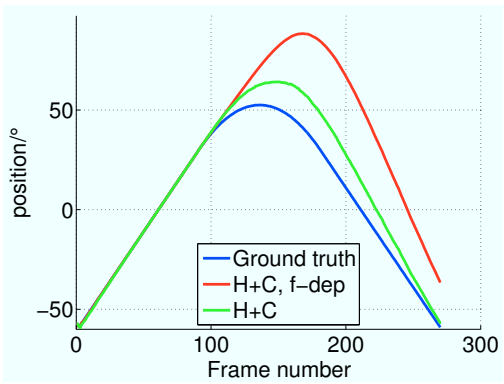
Instead of directly working with this matrix for zoom selection, we instead replace $\mathbf{R}$ in the entropy calculation with

$$\mathbf{R}' = 0.5\mathbf{R} + \alpha 0.5\mathbf{H}_{k+1}\mathbf{C}_k\mathbf{H}_{k+1}^T. \tag{4.30}$$

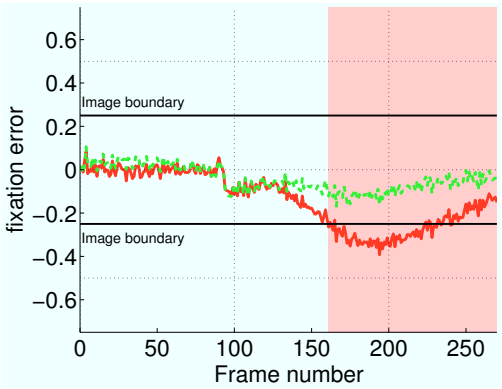
A rise of $\mathbf{C}$ effectively penalises an increasing zoom for a non-matching filter by back-projecting the running average of the covariance with the hypothetical observation parameters $\mathbf{H}_{k+1}$. This penalty is controlled by the factor α, $\alpha \geq 1$.

Note that the term $\mathbf{R}$ is not only changed in the visibility term, but also in the calculation of the entropy of the Kalman filter. This is necessary since an increase in uncertainty flattens the Gaussian in equation 4.13, reducing the impact of the visibility term in the overall conditional entropy calculation.

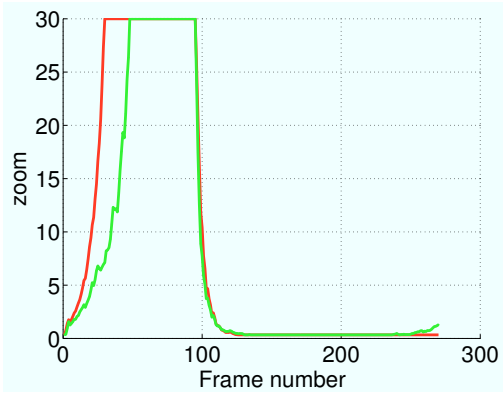
The behaviour of this modification is shown in figure 4.9, again with and without Kalman filter adaption to focal length change. Apparent is the loss of track in the unadapted case. Even though the zoom is set to the minimum, the fixed dynamics of the filter are too slow and the visibility term sinks. The factor α is chosen in such a way that at the smallest zoom level the modified observation covariance $\mathbf{R}'$ attains the original value $\mathbf{R}$, but can also be used as a safeguard value. In figure 4.10 the influence of this value on observation error and zoom selection is shown for varying α. In this setting, the smallest zoom level is $1/3$, i.e. the appropriate α is 9. Apparent is the influence on the maximum zoom level, which is not reached on average for higher values of α before the motion of the target changes, but keeps track in all cases. Figure 4.11 shows a comparison of the approaches presented in this chapter. Most notable is the effect of the combined approach that we proposed in section 4.4.2. This has a far higher influence on the visibility than the approach using innovation covariance, which is only addressing the spread of the Gaus-



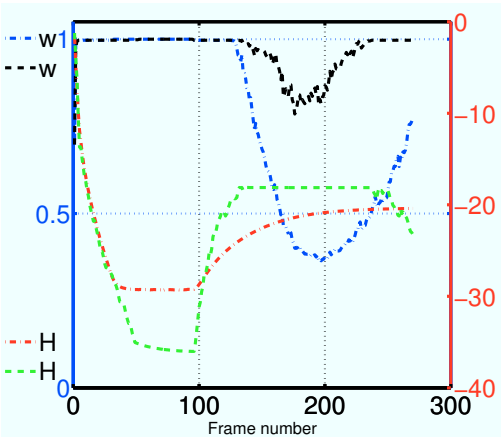
(a) Position



(b) Fixation error



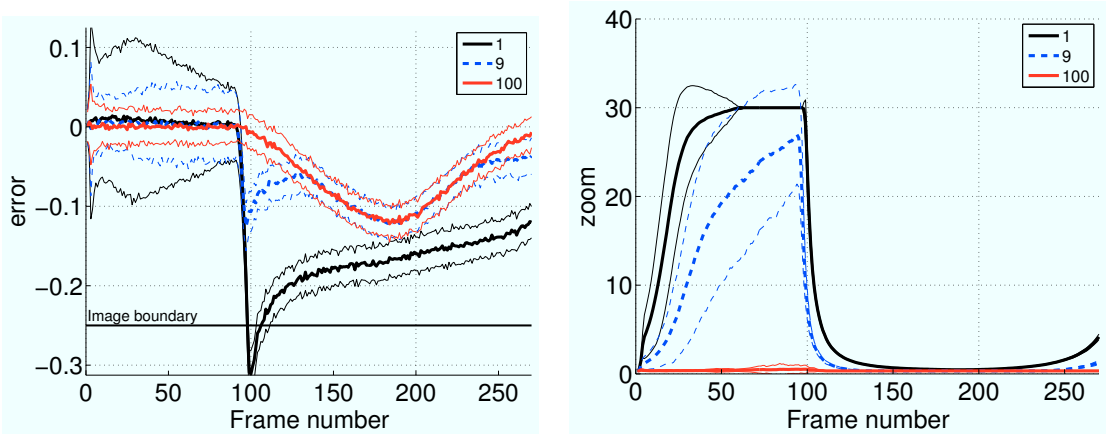
(c) Zoom levels



(d) Entropy H and visibility term w (blue: f-dep.)

Figure 4.9: Position, fixation error, zoom levels and entropy of entropy-based zoom control, incorporating the innovation covariance only (red), and combined with the focal length adaption (green), for $\alpha = 9$

sian, but not the mean value. Since the application of the former method is limited, we restrict ourselves to the use of the approach incorporating the innovation covariance. It tracks as well as the method proposed by Tordoff, but uses a wider range of zoom levels.



(a) Errors

(b) Zoom levels

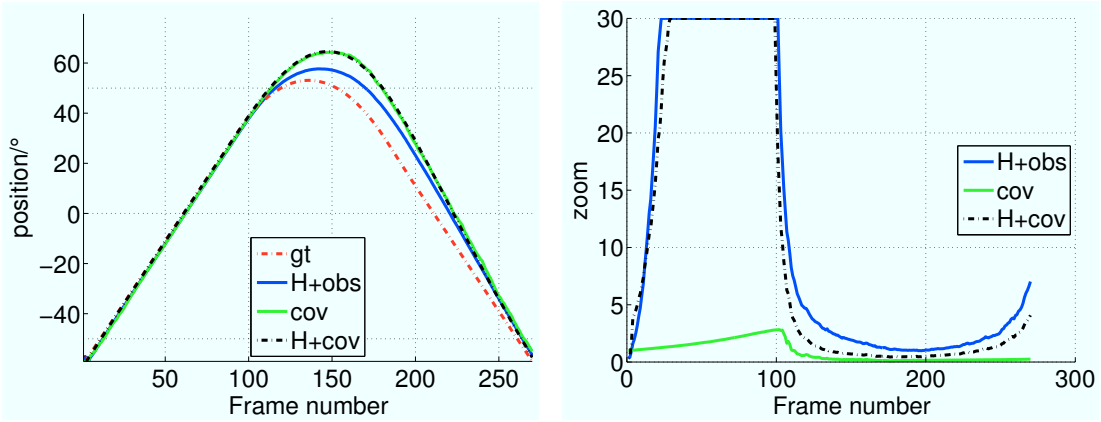
Figure 4.10: Mean performance (and standard deviation) of the filter modification based on the innovation covariance for varying $\alpha = \{1, 9, 100\}$, for 100 runs of the filter.

4.5 Discussion

Since all of the filters performed well with a higher process noise, one might ask why a small value is beneficial. Figure 4.12 shows the resulting zoom settings of the entropy based zoom control with over-estimated process noise $\mathbf{Q}$. All other settings of the model are kept the same, also the maximum zoom setting of 30.

The further the process noise is increased, the more limited is the zoom control. Figure 4.13 shows the minima of equation 4.24 for varying ratios of process to measurement noise covariances. They define the maxima of the zoom obtainable by the entropy based zoom control approach. For example, in order to be a useful zoom criterion for a maximum zoom of $40 \approx 10^{1.6}$, the ratio $\frac{P}{R}$ must not be greater than 10^{-2}, otherwise the minimum would limit the zoom range.

As an example, figure 4.14 shows the behaviour of the filters presented in the previous sections. We used the measurement noise as given in section 4.4, and the steady-state



(a) Position (b) Zoom levels

Figure 4.11: Comparison of the zoom control methods based on entropy with the improved observation ('H+obs'), innovation covariance ('H+cov'), the original covariance method ('cov'). 'gt' refers to ground truth.

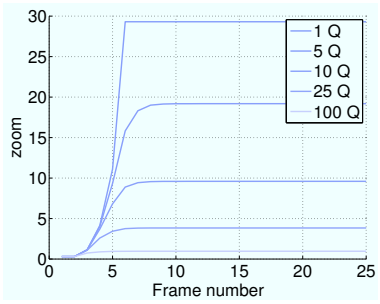


Figure 4.12: Zoom settings resulting from entropy based control with varying process noise covariance. The less trust exists in the process model, the smaller the maximum zoom setting.

solution of the Riccati equation for the state's covariance matrix ([79], p306f). This is finally approximated by $\sigma^2 = \sigma_r \sqrt{(||\mathbf{Q}||_2)}$.

This shows how the zoom effectively is limited by the conditional entropy – even with perfect visibility the zoom would never be set beyond the respective minimum. The interpretation of these findings are explained quite easily — if the process noise is too high, trust in the movement of the target is not high enough to risk zooming in. In the case of the cameraman observing a speeding object, zooming makes sense as long as she is sure the object will continue to move in a direction that can be guessed reliably.

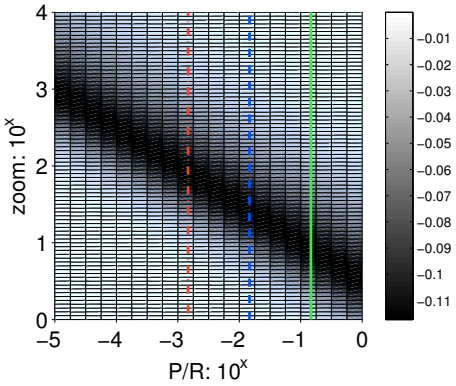


Figure 4.13: Minima of mean conditional entropy for varying ratio of measurement noise to state covariance, over various levels of zoom. Entropies for unmatched (red, stippled), matched (blue, dashed) and a filter with $Q^* = 100Q$ (green) are emphasised. Note the double logarithmic axes.

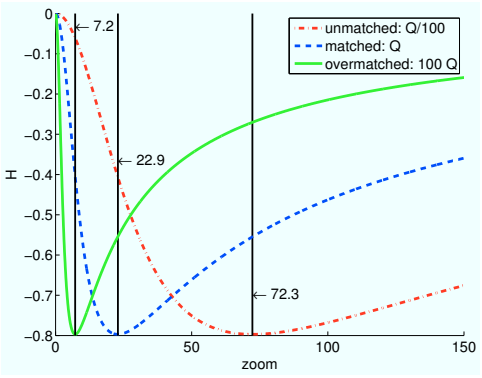


Figure 4.14: Mean conditional entropy vs. zoom level for unbiased and biased filter with $P = Q$. The biased filter would have a zoom limit of 18.75.

4.5.1 Live experiments

We implemented the zoom control method presented in section 4.4.2, which uses the innovation covariance. We qualitatively evaluated the zoom behaviour for a target without motion model, similar to the example given in section 4.3.1.

To this end, we used a calibrated Imaging Source FireWire camera with 16 different zoom settings. The camera was calibrated at every zoom setting using the homography-based calibration by Zhang[108, 165], yielding focal lengths from 1000 to 15000 pixels at a resolution of 1024×768 pixel. A target is manually selected and tracked with an implementation of a mean-shift based approach [33, 77]. The target has a diameter of about $5cm$ and is only half visible at maximum zoom. The results of this experiment can be seen in figure 4.15, where screenshots of live views are placed next to plots of the determinant of the innovation covariance, the absolute distance to the image centre, and the zoom value chosen.

We run the experiment with two different process noises, while the standard deviation of the observation noise was kept constant at 5 pixel. The smaller process noise is used twice, once without, and once with change to the innovation covariance as in our proposed method.

The process noise's standard deviation is initially specified at 0.5cm in (at a distance of 100cm from the camera), and the uncompensated focal length selection method by Denzler [46] was used. The tracking performance resulted in moderate zoom behaviour, i.e. the maximal zoom value attained was 4. The method reacts to a motion of the target towards the boundary of the field of view by zooming out. A screenshot at the maximum zoom level can be seen in the top row of figure 4.15. The red shaded regions denote zooming due to proximity of the image border. Note that the camera does not zoom in onto the target even if there is little motion in the scene.

To obtain a higher maximal zoom, we intentionally underestimate the standard deviation of the process noise by $1/800 = 0.00125$cm, which results in an immediate zoom onto the target (see figure 4.15, middle row). The confidence in the target's position results in the highest zoom level. This does not change even when approaching the image border, and results in target loss.

Lastly, we maintain the process noise setting and choose to adapt the innovation covariance as described in section 4.4.2. The value α is set to 1, and the temporal adaption γ is set to 0.5 for increasing determinants of the innovation covariance, and $0.5/8$ for the decreasing case. This results in slower zoom in, and faster zoom out behaviour for abrupt changes. As can be seen in the last row of figure 4.15, the maximum zoom is also attained if the target remains static for some time, but quickly decreases upon nearing the image border or when unexplained motion is observed (green shaded regions).

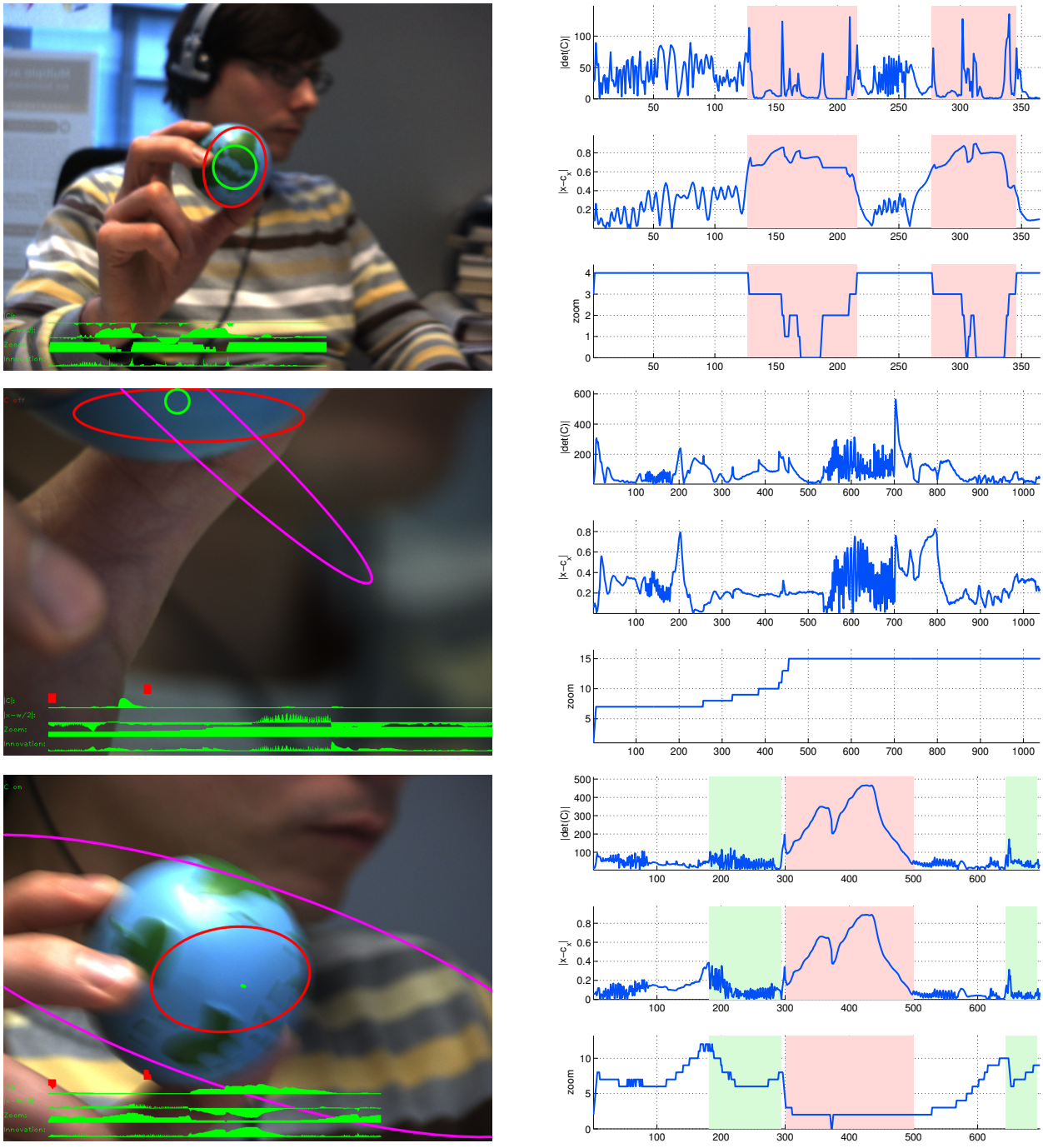


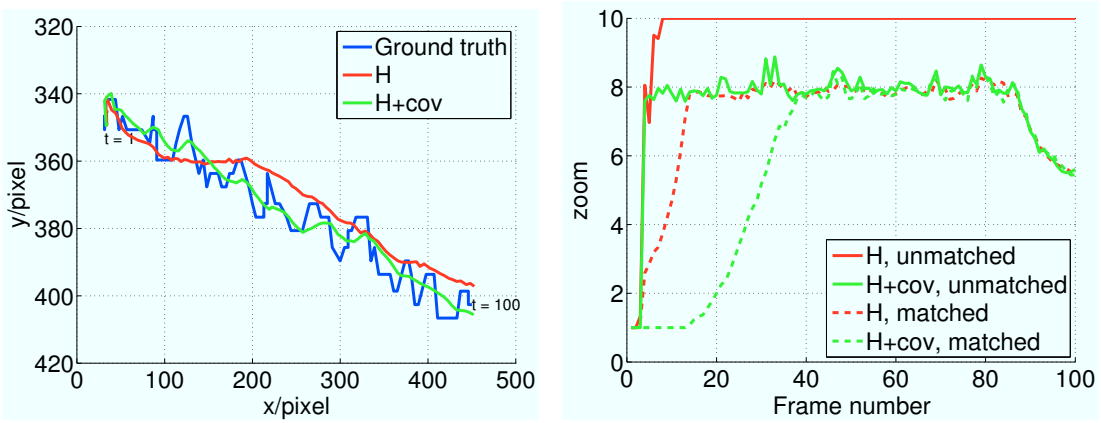
Figure 4.15: Left, top to bottom: Live views (including annotations) with over- and underestimated process noise (top and centre), and underestimated but adapted filter (bottom). Right: Corresponding normalised innovation covariance, distance from image centre, and zoom reaction for each filter variant. Top: Note the response to deviation from the centre, and limited zoom. Centre: The zoom plot has an increased maximum zoom, but no reaction to unexpected motion. Bottom: the same maximum zoom level as before is attained, but shows responsiveness to unexplained motion (green) and proximity to image border (red).

4.5.2 Application in Visual Surveillance

We applied the presented zoom control to tracking of humans in video sequences, as shown previously in figure 4.3.

For the purpose of comparison of zoom control methods, we use the same observations for all methods, which are extracted beforehand from a high-definition video sequence by a method based on background subtraction [124]. Every frame is then a subsampled part of this video input, imitating a PTZ camera. We assume a constant velocity motion model, and the observation is the bounding box of the detected object.

In this setting and with maximum zoom level of 10, we run an experiment with a matched and an unmatched filter, where the process noise is underestimated. Here, we compare the entropy based zoom control in the original form with the innovation covariance based modification presented in section 4.4.2. Figure 4.16 shows a comparison of the zoom level chosen by both methods for the matched and unmatched case, as well as the resulting trajectories of the centre of the target obtained by the biased filters. Both control rules for matching filters zoom to the maximum zoom level which allows observation of the whole target, and gradually zoom out once the size of the target increases. Again, the entropy based control rule applied to the unmatched filter results in a fast zoom onto the target up to the upper limit. Since the control rule assumes a matching filter, the zoom level remains set to this setting. The control rule which includes the innovation covariance, however, incorporates the error in the observation process and adjusts zoom accordingly. The trajectories also show that the improved control method results in a better estimate. This comes at the price of a slightly delayed zoom.



(a) Position

(b) Zoom levels

Figure 4.16: Comparison of entropy based zoom control without ('H') and with innovation covariance ('H+cov') both for a matched and unmatched filter on a surveillance sequence.

4.6 Conclusion

We presented novel zoom control methods which use an information–theoretic criterion, but also integrate an estimate of the fixation error to make robust predictions of the expected observability. A thorough analysis of previous work has shown that the ordinary entropy based control rule needs a matching filter, and its resulting zoom rule is bound by an overestimated process noise. This demands a filter which is not necessarily tuned to minuscule accuracy, but can react to fixation errors instead. To make the filter robust with regard to an incorrect motion model, we introduced the predicted fixation error into the visibility term, and inserted the innovation covariance into the mean conditional entropy. Lastly, we have discussed the Kalman filter's sensitivity to zoom changes if it is not matching the noise characteristics.

Chapter 5

Multiple Targets and Scene Exploration

In this chapter we address the exploration of the environment under surveillance by a single, active camera, which is an objective competing with the goal to track targets optimally. To this end, we extend the work in previous chapter to multiple targets, and propose a term that measures the knowledge about the environment, and compare the proposed method with other approaches.

For this, we address the information gain from several targets, and introduce an extra objective to explore the supervised area, which means letting go of tracked targets if the expected mutual information for observation of other areas is higher. We show that the resulting objective function results in a better tracking and detection performance than standard methods.

Parts of this chapter have been published in [136, 137].

5.1 Introduction

Object detection and tracking of objects in video data are crucial elements for further reasoning in modern vision-based systems. In that regard, surveillance systems need to address the task of tracking multiple targets. Furthermore, the acquisition of high resolution imagery of these targets is a prerequisite or benefit for tasks such as identification or classification [117]. If a supervisor camera provides sufficiently accurate positional measurements of the positions, then further scrutiny of the desired target is straightforward by directing long, fixed focal length cameras at it. This setup however constrains the

region of activity – where active cameras can observe and track targets after such instruction – to the region covered by the supervisor camera. To this end, multiple supervisor cameras can be installed [85], and several active cameras instructed. Furthermore, a decision has to be made which target is the most useful to observe if there are more targets than cameras.

This chapter proposes a different approach that harnesses the flexibility of a single, active zoom camera which can serve both objectives: by adjusting the zoom parameter it can provide overview images, as well as close-ups of targets. Instead of using a supervisor camera, we add to the camera control objective function a term that demands exploration of the supervised area for targets that might have appeared in regions that have not been observed for some time. When this term for exploration outweighs the tracking term, the camera control stops following targets and selects promising regions for detection of new targets.

Unfortunately, the exploration of the scene conflicts with a close-up inspection of objects of interest: the longer the camera is focussed on other areas of the supervised area, the more rises the uncertainty about the position of previously tracked targets, until the objective of tracking targets outweighs the exploratory one. Similarly, zooming into a part of the scene decreases the field of view of the camera, and areas with possibly interesting behaviour are not covered any longer. Furthermore, and along the lines of the previous chapter, active zoom control needs to strike a balance between the maximum attainable zoom onto an object and the risk of losing lock. The resulting behaviour is an alternation of wide angle observations and close-ups of the targets identified. We address this problem with the information theoretic toolkit developed in chapter 3 and propose an objective function that balances both goals of target tracking and detection.

The next section addresses work related to the exploration of environments and the tracking of multiple targets. In the subsequent sections we develop the term that addresses the exploration and give an quantitative analysis of the resulting surveillance process.

5.2 Related Work

When there are more targets to be observed than sensors available, a decision has to be made which target to observe with which sensor. This camera assignment problem is phrased as a dynamic optimisation problem by Bagdanov et al. [10]; specifically Isler *et al.*[78] address the computational issue of assignment of a single target to a single camera. Zhang and Qi [164] use mutual information from a Dynamic Bayesian Network to decide upon use of different sensors for a specific target type. Takemura and Miura [150] focus on camera assignment to targets and planning over multiple time steps, but do not address the uncertainty of the sensing process inherent in vision systems.

Del Bimbo et al.[44] use visual odometry and particle filters to steer a single, active camera, but the actual control command comes from a supervisor camera. In other work on multi camera control by the vision community [10, 72, 121], all use at least one specific supervisor camera and specific, hand-crafted rules to control the individual sensors, for example choosing the zoom setting via geometric reasoning. Hampapur et al. [72] also incorporates a head detector to focus the zoomed view onto the face of persons.

Active control using reinforcement learning has been addressed by Naish et al. [105] for tracking of targets; Bagdanov et al. [9, 45] apply this method to the acquisition of close up imagery of a single target; Derichs [47] in particular addresses long term planning of camera motions for classification. The underlying problem in reinforcement learning

is to learn a mapping $Q(\mathbf{s}, \mathbf{a})$ that specifies the expected reward for any element $\mathbf{s}$ from the state space and $\mathbf{a}$ from the action space. This mapping has to be learned. When little structure is known about the problem, then Monte Carlo Learning is one approach, which basically samples the action space. Occurrences of different samples $\mathbf{s}$ from the state space are assumed sufficiently dense and repeating. The resulting mapping is then the expectation over all different returns for all state and action pairs. In the case of multiple target tracking, even the use of a single camera already poses a very high-dimensional state space, which cannot be explored by current reinforcement learning methods.

Davis et al. [41] use detected motion from randomly chosen pan/tilt settings to learn a map which is then used to select future camera parameters. The authors propose several methods to navigate through the learned map, but all goal locations are chosen randomly by assuming the map entries stem from an unnormalised probability distribution. Another kind of activity map can be found in Gould et al. [65]. Here, a sophisticated perceptual model is learned and used to drive the focus of attention. Objects are classified in a close-up view which is selected from a wide angle view having a high chance of containing classifiable objects. The actual distribution is given by a previously trained Bayes network.

Sequential camera scheduling in a decision theoretic framework has also been used to generate a series of fixation points. Gu et al. [70] obtain these with a set of thresholds to accommodate for the heuristic modelling of the costs and rewards obtained in from directing the eye fixation point. Similar to our work, Paletta and Fritz [112] model these gains with entropy as a measure of local information content and use reinforcement learning to obtain a general policy which needs shorter sequences of fixations to reduce uncertainty than selecting these points in a random fashion.

Sensor Placement

A related problem is the choice of sensor placement, which is the one-off choice of where to install sensors for a permanent installation. This problem has similarity to ours, although we have to solve the problem at every time instant, but with significant restrictions (i.e. direction only) on the placement.

This problem is addressed by a body of literature; for cameras with multiple overlapping views, Park et al. [114] propose to use only the best camera for a tracking task, where best is defined by distance of the tracked target to the view frustrum. The encompassing approach is to minimise the installation cost with respect to a maximisation of the quality of the recorded data [54, 160, 161].

Given a floor plan, Yao et al. in [160] use a binary occupancy grid with N locations to label positions that can be seen by at least one camera. This grid is then turned into a binary vector $\mathbf{b} \in 0,1^N$. For each possible camera position and one parameter setting x_j, a row vector $\mathbf{a}^T$ is set up with a_i corresponding to the evaluation of an objective function, e.g. if this setting can see the corresponding location b_i. This yields a binary vector $\mathbf{x} \in \{0,1\}^J$ for all J different camera positions and parameter settings. Given a cost w_j for each parameter setting, the overall cost of the installation is then

$$c = \mathbf{w}^T \mathbf{x}. \tag{5.1}$$

The approach taken is then either the maximisation of the visible area given a maximum cost $C_{\max}$:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \sum_i A\mathbf{x} \text{ subject to } c \leq C_{\max} \tag{5.2}$$

or, similarly, the minimisation of the installation cost with respect to a minimum coverage overall $c_{\min}$ or for a minimum required set of locations $\mathbf{b}_c$.

The costs are then derived from observation likelihoods, occlusion probability, camera handoff of targets, and camera overload. In the latter case the target appearances are modelled using a Poisson process [161]. Even though the method addresses placement of pan tilt zoom cameras, it does not address the active handling of targets, as the camera assignment and placement is fixed.

This approach neither takes into account target tracking accuracy nor detection likelihoods, as the objective function only addresses single cameras for a single location. Most importantly, it also ignores the diminishing returns property of prolonged observations, i.e. temporal dependencies.

Similarly, Krause et al. [87] have argued for the maximisation of mutual information as the means to solve sensor placement tasks. Nevertheless much of the sensor placement work finds an optimum for a static environment with temporal average target behaviour, and does not address active parameter changes to measure short term behaviour. An exception is Bodor et al. [22], who use their method to place a robot for optimal surveillance; however, this is only single agent system where no zoom parameters are adjusted.

The discretisation of the ground plane into occupancy grids to drive exploration was introduced to the robotics community by Moravec and Elfes [51, 102], who also used information-theoretic criteria to explore a static map in [52]. Bourgault [26] combines this with the information acquired from the extended Kalman filter in a SLAM system, using greedy planning.

Target Search

The aim of single target detection can be seen as an instance of a "search and track" problem. In the works by Bourgault et al. [25] and Furukawa et al. [59], the continuous target space is discretised either in a regular grid or a set of elements defined by shape

functions. Recently, Hollinger et al. [75] proposed a graph-based approximation to a set of connected rooms, and applies Krause's [87] sensor placement algorithm. Tisdale and Ryan [125, 152] address the search problem using a particle filter approach over the search space. All methods assume a largely homogenous detector performance in the supervised region, and assume a static environment apart from a single, dynamic target.

To the best of our knowledge, there is no prior work that addresses the two objectives of detecting targets as well as following them with a single sensor. The long term planning methods obviously do not address the control required to actively track targets and proscribe only a single sensor setting. The target search methods address only the single objective of finding a target as quickly as possible, and make the assumption of a limited number of moving targets, but an otherwise static environment. Apart from this, the approach from Ryan and Hedrick [125] is phrased in an information theoretic sense and as such could be included in our framework.

This chapter introduces a discretisation of the search area where targets can be detected, and affixes each location with a likelihood that a target appears. In the next section, we extend the objective function for tracking a single target from previous chapter to several targets, and show that it results in a round-robin schedule for target observation. In subsequent sections we introduce the term that drives the exploration, and present experiments which take into account scene activity.

5.3 Multiple Targets

We now extend the target tracking problem of chapter 4 to multiple targets. Assume a single camera and several targets. For several targets, we have a set of state vectors $\mathcal{X} = \{\mathbf{x}_1 \dots \mathbf{x}_N\}$, and a set of observations $\mathcal{O} = \{\mathbf{o}_1 \dots \mathbf{o}_M\}$.

We take the same approach of choosing the optimal observation parameter by maximising the information gain (see equation 4.2). The information gain can then be expressed as follows:

$$\mathbf{a}_k^* = \arg\max_{\mathbf{a}_k} I_{\mathbf{a}_k}(\mathcal{X}; \mathcal{O}) \tag{5.3}$$

In general, this maximisation could include the data association problem, which relates any of the M expected observations to the target positions. As we are interested in the expected observations, these come from the predicted target positions, observed by the parameter $\mathbf{a}$ under evaluation. These expectations can also result in several assignments of observations to targets, and can include the failure rate of the underlying sensor model (e.g. the actual vision based detector that is employed), as well as false alarms. Depending on the data association method used, the influence on the state estimate will vary. For example, in the formulation of the Joint Probabilistic Data Association Filter [13], this influences the covariances according to the proximity of the observations.

For the remainder of this work, however, we take the data association for given, i.e. we assume independence of the targets and thus ignore information gain from data association. Each target thus generates a single observation and this observation is with certainty associated with its source.

The equation 5.3 thus reduces to a sum:

$$\mathbf{a}_k^* = \arg\max_{\mathbf{a}_k} \sum_{n=1..N} I_{\mathbf{a}_k}(\mathbf{x}_k^n; \mathbf{o}_k^n) \tag{5.4}$$

Note that this is not equal to the sum over entropy of a single target, as when tracking several targets $\mathbf{x}_1 ... \mathbf{x}_N$, the uncertainty in each state, $H(\mathbf{x}_i)$, influences the importance of a target in the sensing process. For example, if the state is known well enough and the increase in uncertainty over the next time steps small, the camera can focus onto a

different target, and be fairly confident that the original target can be picked up again later.

5.3.1 Prioritisation of Targets

In this section we show that the use of mutual information as an objective function prioritises the target that has not been observed for the longest time, which results in a round robin scheduling of the targets.

To demonstrate this, we create a simple scenario, comprising a set of N targets and a single camera. Assume that for a camera setting a_0, all targets are visible, and have been observed long enough such that the covariance of the targets' positions have converged to the steady state solution. The state of each target i at time step k is denoted as $\{\mathbf{x}_{i,k}, \mathbf{P}_k\}$.

We further simplify this scenario by allowing only $N+1$ observation settings[1]: The camera can either focus on all targets at the same time (parameter a_0), or zoom onto each of the targets exclusively with setting a_i. This will then exclude all other targets from observations. The covariances are thus predicted for all unobserved targets, and updated for the observed ones.

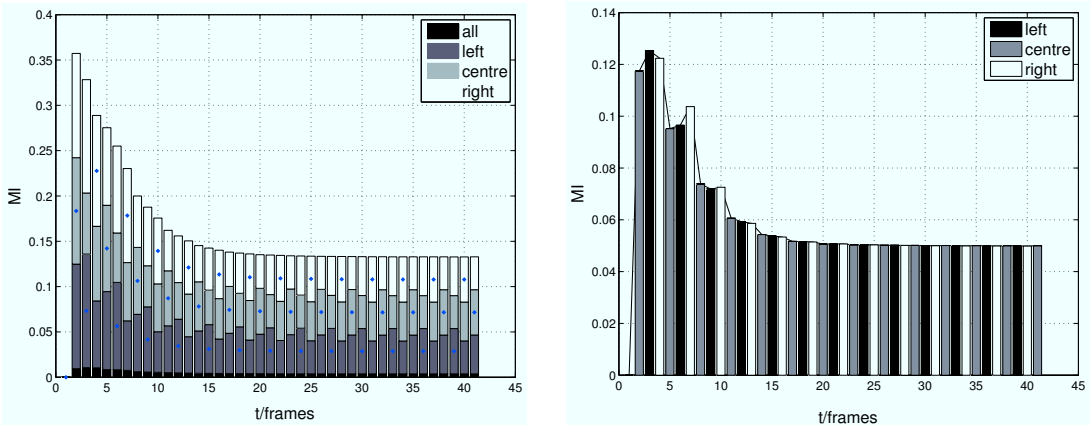
The information gain when keeping all targets in view (a_0) is 0 because the Kalman filters are in the steady state. Upon zooming onto a single target, the information gain from any of the unobserved targets is

$$I = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{o}) = \log(|\mathbf{P}^-|/|\mathbf{P}^-|) = 0 \tag{5.5}$$

The information gain from the single, observed target is

$$I = \log(|\mathbf{P}^-|/|\mathbf{P}^+|) = -\log(|1 - KH|) \tag{5.6}$$

[1] this is of course relaxed in later implementation



(a) MI for all views

(b) MI per chosen view

Figure 5.1: Mutual information of three targets as a function of time. Figure 5.1(a) shows the objective function for a zoom onto each of the targets, and the reward when staying zoomed out (in the first frame). The optimal choice is denoted by a blue dot in the figure, shown in figure 5.1(b). Note the round-robin cycle of the chosen target selection. The mutual information is smallest for the target that has just been observed at highest resolution, while it rises for the other targets.

The parameter setting with maximum information gain is thus a zoom onto a single target. Initially, all targets provide the same amount of information and the first one can be chosen. In the next step, the information gain from keeping this single target under closer scrutiny is smaller than observing any of the ones which have not been observed, and the camera will observe another target. The focus starts anew onto a different target, because the information gain from the target observed last is the smallest. Hence the targets are observed in a round-robin fashion. Figure 5.1 illustrates this for three artificial objects which are tracked in turn. The figure shows plots of the effective resolution of each target as a function of time, clearly showing that we obtain an automatic and natural scheduling of attention between the targets.

5.4 Appearance/Disappearance of Targets

In the objective function in equation 5.4, no term is included that requires the system to look for new targets. Instead, no matter how long a single target has been tracked, if it remains in the supervisable domain of a camera, this camera will include the target in the observation. The following sections present an approach to introduce a sense of anticipation into the objective function. That is, the longer the camera has been focussed on a smaller region of the area, the higher the uncertainty about the number of targets in its environment.

For an analogy, imagine a street scene where pedestrians show up regularly, but unpredictably. The street has a lot of houses and doors - nearly everywhere a person can appear. The absolute times of two appearances are independent random variables - the number of appearances before an occurrence is independent of the number of the following ones. Whereas information can be gathered by tracking a single target, the uncertainty about the other, currently not observed areas is increasing. With the term *exploration* we describe the objective to observe these areas and the deliberate stop of tracking a currently known target. The incentive to do so is introduced in the objective function by a dedicated term.

In the following sections, we present two methods to model this term for exploration. We then combine both objectives, the tracking and the exploration term, into one objective function. With this we address the goals of tracking targets as well as exploration of a supervised area for new targets without the need of a supervisor camera. The next sections present a simplification for the case of a single, active camera, whereas the approach in chapter 6 extends the mutual information from tracking and exploration to several cameras, taking into account the uncertainty in the detection process.

5.4.1 Scene Exploration with a Single Camera

The approach we take in this section is to model the likelihood of the appearance of a single target in the environment that is currently not under supervision. This likelihood influences the information to be gained from initialising a new target, I_0.

The resulting objective function is thus

$$\mathbf{a}_k^* = \arg\max_{\mathbf{a}_k} \left(I_0(\mathbf{a}_k) + \sum_{n=1..N} I_{\mathbf{a}_k}(\mathbf{x}_k^n; \mathbf{o}_k^n) \right) \tag{5.7}$$

The first term in this sum provides incentive to zoom out and to search the scene for new arrivals if the likelihood of a new appearance is high enough.

We now derive the information $I_0(\mathbf{a}_k)$ to be gained from a detection in the area covered by the observation parameter $\mathbf{a}_k$. Furthermore, we show how to model the appearance rates for this exploration term, and lastly, we compare the proposed method with other, standard methods on publicly available data sets.

5.4.2 Poisson Process

To model the likelihood of the appearance of a new target, we discretise the supervised area into a set of disjoint locations. We assume that targets appear randomly, and independently, and thus model these appearances with a temporally homogeneous Poisson process (i.e. the arrival rate does not vary over time). The waiting time T until the next appearance of an object at location $\mathbf{y}$ thus has an exponential distribution with the appearance rate $\lambda(\mathbf{y})$, i.e. we assume different appearance rates at every location, as this allows us to model entrances or sources for targets in the surveillance area. Applying the Poisson distribution, the probability of no appearance at $\mathbf{y}$ after having waited for time t is

$$p(T > t, \mathbf{y}) = \frac{(\lambda t)^0 e^{-\lambda(\mathbf{y})t}}{0!} = e^{-\lambda(\mathbf{y})t} \tag{5.8}$$

The chance of an activity (one or more appearances) h_k at location $\mathbf{y}$ since the last observation $t_0(\mathbf{y})$ is

$$p(T < (t - t_0(\mathbf{y})), \mathbf{y}) = p(h_k, \mathbf{y}) = 1 - e^{-\lambda(\mathbf{y})(t - t_0(\mathbf{y}))} \tag{5.9}$$

We denote the current field of view as $\mathcal{F}_k = \{\mathbf{y}_{k1}, \ldots, \mathbf{y}_{kN}\}$, that is all N different locations that are visible at time step k for the current observation parameter. Assuming that all probabilities of appearance at these locations are independent, the probability of making at least one observation is

$$\begin{aligned} p(h_k|\mathcal{F}_k) &= 1 - \prod_{\mathbf{y} \in \mathcal{F}_t} 1 - p(h_k, \mathbf{y}) && (5.10) \\ &= 1 - \prod_{\mathbf{y} \in \mathcal{F}_t} e^{-\lambda(\mathbf{y})(t - t_0(\mathbf{y}))} && (5.11) \\ &= 1 - \exp(-\sum_{\mathbf{y} \in \mathcal{F}_t} \lambda(\mathbf{y})(t - t_0(\mathbf{y}))). && (5.12) \end{aligned}$$

This assumes that if an object has appeared, then it will be detected by the system if it is in the field of view $\mathcal{F}$ of the camera. We assume that once an object is detected, the object is tracked by a Kalman filter.

We now derive the information I_0, which is part of the first term in equation 5.7. It is the information to be gained from a new target track upon detection.

We make the simplifying assumption that at any time step, only a single target can appear. This underestimates the expected information gain from acquiring new targets if several appear. [2]

[2] The exact term does not have a closed solution due to the inhomogeneity and different observation times of the Poisson processes involved. For a single location, however, the actual expectation is

$$\begin{aligned} I^* &= \sum_{k=1..\infty} p(N(\delta T) = k)\, k\, I_0 & &= \sum_{k=1..\infty} e^{-\lambda \delta T} \frac{(\lambda \delta T)^k}{k!}\, k\, I_0 && (5.13) \\ &= \sum_{k=1..\infty} e^{-\lambda \delta T} \frac{(\lambda \delta T)^{k-1} (\lambda \delta T)}{(k-1)!} I_0 & &= \sum_{k=0..\infty} e^{-\lambda \delta T} \frac{(\lambda \delta T)^k}{(k)!} (\lambda \delta T) I_0 && (5.14) \\ &= \lambda \delta T I_0 && && (5.15) \end{aligned}$$

However, this constraint on the locations under surveillance proved too limiting to the problem investigated in this chapter.

The value I_0 is the mutual information between the location of the target and the event of appearance within the field of view. It can be expressed as the difference of the state's entropy H_u and the entropy of the state conditioned on the appearance:

$$I_0 = H_u - H(\mathbf{x}^{n+1}|h_k). \tag{5.16}$$

The entropy of the state before appearance H_u is a constant equal to the uncertainty in the state of the undetected object. It can be interpreted as the entropy of a uniform distribution of a target that is in the currently unsupervised areas, or has not appeared yet, and is the best estimate as we currently are not tracking this target. In practice, we set it to the logarithm of ten times the covariance of a newly initialised tracker.

Lastly, the second part of the information is the conditional entropy $H(\mathbf{x}^{n+1}|h_k)$. For this we take the expectation over appearance and non-appearance:

$$\begin{aligned} H(\mathbf{x}_k^{n+1}|h_k) &= p(h_k|\mathcal{F})H_{tracked}(\mathbf{x}_k^{n+1}) \\ &\quad +(1-p(h_k|\mathcal{F}))H_u. \end{aligned} \tag{5.17}$$

This takes into account the chance that a target has appeared in the field of view ($p(h_k|\mathcal{F})$), weighting the entropy $H_{tracked}(\mathbf{x}_k^{k+1})$ of the target right after instantiation of a new tracker. To this adds the initial uncertainty weighted by the probability that the target has not appeared ($1 - p(h_k|\mathcal{F})$). We now have identified the elements of $I_0(\mathbf{a})$ and are ready to apply equation 5.7.

The behaviour is predictable: if all known targets are sufficiently well tracked, a view $\mathcal{F}_k$ is chosen which reduces the entropy by tracking a new object, weighted by the chance of its appearance. This view can exclude other, already tracked objects - the uncertainty in their position rises, increasing their entropy $H(\mathbf{x}_{k+1}^n)$.

An example of this behaviour is shown in figure 5.2.



Figure 5.2: Example for zooming in onto a target, and zooming out to explore the environment in a downsampled image sequence. Only the area lined in red is visible to the camera at any instance; the darker the area, the longer ago is its latest observation.

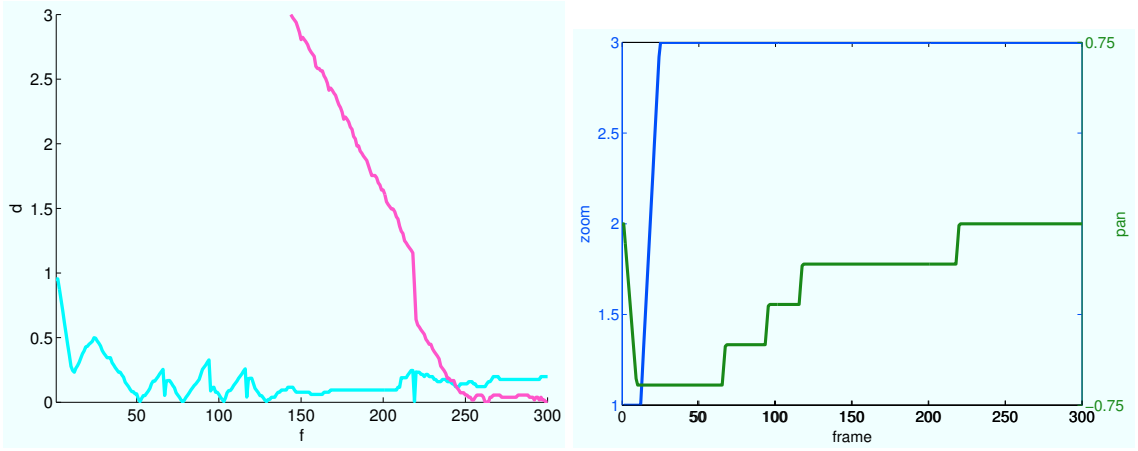
Figure 5.3 shows plots of the current zoom setting per frame and the horizontal distance of actors to the camera's focus normalised to the current width of the visible region. Any distance below 1 indicates a visible object. The figure shows the effect of the Poisson process on the detection of the second (right) actor in the sequence shown in figure 5.2. If appearance is not modelled (see figure 5.3(a)), the camera simply follows the object by adjusting the pan setting accordingly. In the case of introducing the scene uncertainty (see figure 5.3(b)), the camera scans the area of the currently tracked object while there is sufficient certainty about its next position. This local scanning behaviour results in an earlier detection of the second object.

5.4.3 Learning scene activity

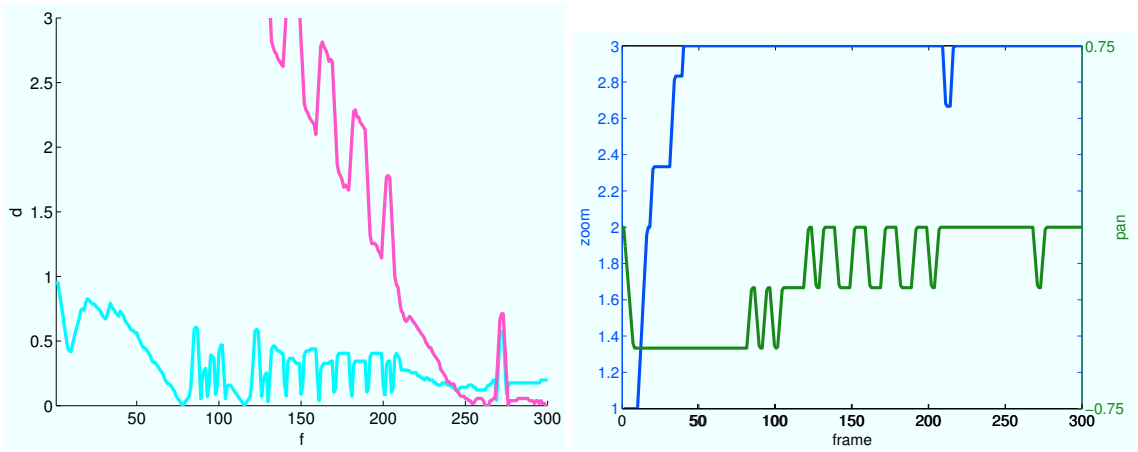
In most scenes there are areas where fewer appearances will occur, e.g. the appearances and disappearances of pedestrians are limited by walls, or parts of a camera's view can be blocked. In this section, we propose a simple method to learn these locations and appearance rates.

Target appearances happen near entrances to the scene. Such entrances (as well as exits) can be learned from long term observations by fitting a Gaussian mixture model to end points of trajectories [95, 147], or spatially extended regions [100]. The activity connecting these regions is then learned using neural networks [80], Gaussian processes [53] or Hidden Markov Models[96], which model transitions between learnt trajectories. These methods all rely on a working tracker, i.e. they are susceptible to occlusion of targets as well as tracker failures. A recent approach uses spatial histogramming of KLT features for scene activity classification [43].

Here we take a simpler approach. As we modelled the likelihood of appearance on the basis of independence between the locations, we exploit this by trivially learning the



(a) Appearance not modelled



(b) One appearance per minute, 25fps

Figure 5.3: (a), Left: Horizontal distance of actors to the view centre, right: zoom and pan setting per frame. (b) shows the earlier detection of an actor due to the Poisson process.

number of appearances per location, that is the number of time steps this location has been occupied by an appearing target. For the evaluation, we learn this on the pixel level of the ground truth data, yielding an appearance map. An example of such a map is shown in figure 5.4 for both viewpoints in all sequences of the EC-funded CAVIAR "shopping mall" data set [3].

[3] EC Funded CAVIAR project/IST 2001 37540, found at URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/



Figure 5.4: Appearances in corridor and frontal view of the CAVIAR "shopping mall" data set.

5.4.4 Evaluation

We evaluate the methods proposed in this chapter by simulation based on ground-truth data. We use the annotations supplied with the CAVIAR test case scenarios and add Gaussian noise of 1 pixel to the labelled bounding boxes. This also removes other sources of error in the evaluation, e.g. from detection, tracking and data association. Each detection is assigned a Kalman filter, which is used to obtain the uncertainty of the tracking as described in section 4.3.1. The Kalman filter uses an observation noise of 1 pixel, and process noise of 0.05 units. A track is lost if for more than 10 frames no observation has been made, the target leaves the maximum field of view, or the expected measurement does not overlap with the actual measurement.

5.4.5 Metrics

The metrics we use are the latency, the fragmentation of a track, and the overall coverage of all tracks compared to the ground truth, as detailed in section 3.8.

We use the threshold of 0.25 for the average spatial overlap, and the tracks must have a temporal overlap of at least 1/6. This means that to assign a ground truth track to an

output track, they must both coexist for at least a sixth of their common time, a half of the ground truth observations must be made with at least half of their area.

5.4.6 Experiments

To evaluate the influence of the exploration term on the tracking of targets, we performed two experiments. The first experiment addresses the effect of the appearance rate on the observation of single targets. The second experiment is concerned with the early acquisition of targets, if the camera cannot observe the whole scene at the same time. Here, the appearance rate has been determined a priori from the data set (see figure 5.4).

In the experiments, we approximated the information gain in equation 5.4 using the entropy difference, i.e. reduction in entropy in one time step:

$$\mathbf{a}_k^* = \arg\max_{\mathbf{a}_t} \sum_{n=1..N} \Delta H^n \tag{5.18}$$

$$= H_{k-1}(\mathbf{x}_{k-1}^n | \mathbf{o}_{k-1}^n, \mathbf{a}_{k-1}) - H_k(\mathbf{x}_k^n | \mathbf{o}_k^n, \mathbf{a}_k). \tag{5.19}$$

This is not exactly the mutual information, as the first term is not using the predicted state estimate. This entropy difference thus underestimates the information gain, as due to process noise, the uncertainty of the predicted state is always greater than the updated one.

Furthermore, the choice of the objects to be tracked is considered. Instead of focussing onto all targets, we investigate which subset Ω of the currently tracked targets yields the highest decrease in uncertainty

$$\mathbf{a}_t^* = \arg\max_{\mathbf{a}_t, \Omega} \sum_{\Omega} \Delta H \tag{5.20}$$

The experiments evaluated the performance of the scheduling using maximisation of mutual information for single, all and a subset of targets with a maximum number of 3 targets. In the latter case, if a further target was within the bounding box spanned by the

bounding number of targets in the subset, it is added to the evaluation. We also evaluated a straight minimization of the sum of all targets' conditional entropy, ignoring the amount of uncertainty of each target. Lastly, we evaluated standard rule-based scheduling methods, i.e. random selection of targets and the first come, first serve rule (FCFS) as used by [121].

For the first experiment we assume a minimum zoom setting which allows to observe the whole scene. Such a setup simulates a "virtual zoom" camera, which simply downsamples or resizes image regions from a sensor with a higher resolution. Another example for such an input is high definition video, which can be processed much faster by restricting analysis to the relevant parts of the image[14]. In this experiment, we set the maximum zoom (i.e. magnification factor) to 3, which corresponds to a tripling of the focal length. The evaluation compares different levels of modelled scene activity by assuming a spatially homogenous appearance rate, and varying its value λ. The average activity of the scene locations was chosen as one appearance every minute, every 12s and full second, or 5 appearances per second, respectively.

The number of false positives, shown in table 5.2(a), shows that with entropy based scheduling methods more tracks are not assigned to any of the ground truth tracks, i.e. yield a higher fragmentation of tracks. This is effectively the number of targets which needed to be initialised again due to a longer focus onto other targets. The false negatives in table 5.2(b) shows how the number of completely missed target goes down when the appearance rate is high enough. For both of these figures, the absolute number of targets in all sequences is 324. While the methods barely differ in the latency (table 5.3(a), left), the advantage of the methods presented here can be seen in table 5.3(b), where the entropy based scheduling methods result in a better coverage of the targets.

Table 5.1: Data association performance of scheduling methods with respect to appearance rate. The method selects either a *subset* of targets, *all* targets, or only a single one (1). False positives (FP): falsely reacquired targets and False Negatives (FN): missed targets, both out of a total of 324. First come, first serve ("FCFS") and random selection are independent of appearance rates.

$\lambda\ (1/s)$	Subset	All	1
1/60	14	14	12
5/60	14	15	13
1	13	19	12
5	14	16	9

(a) False positive tracks. FCFS: 11, Random: 15

$\lambda\ (1/s)$	Subset	All	1
1/60	1	6	5
5/60	1	6	5
1	0	0	6
5	0	0	5

(b) False negative tracks. FCFS: 4, random: 3

Table 5.2: Latency and resolution performance of scheduling methods with respect to appearance rate. The method selects either a *subset* of targets, *all* targets, or only a single one (1). Latency in frames (left) and Observation area relative to ground truth (right) for constant appearance rates. First come, first serve ("FCFS") and random selection are independent of appearance rates.

$\lambda\ (1/s)$	Subset	All	1
1/60	4.94	4.78	4.83
5/60	5.02	4.88	4.84
1	5.11	4.93	4.87
5	5.03	4.96	4.80

(a) FCFS: 4.90, Random: 5.11

$\lambda\ (1/s)$	Subset	All	1
1/60	1.69	2.18	1.65
5/60	1.62	2.11	1.61
1	1.59	1.65	1.58
5	1.58	1.53	1.56

(b) FCFS: 1.13, Random: 1.38

The second experiment compared the algorithms with a minimum zoom value of 2 and a maximum of 4, which requires an exploration, or scanning, of the scene. The camera settings continuously have to balance the reduction in uncertainty for a few targets with the risk of missing a target in the area currently not observed. The result is shown in figure 5.5. In addition to the standard methods of scanning ('scan'), random target selection ('rnd') and first-come-first-serve ('FCFS'), we added a background-only policy ('bg'), which results from searching for a new target only, without taking any tracking based utility into account. This last method, as well as the methods described in section 5.3 – all

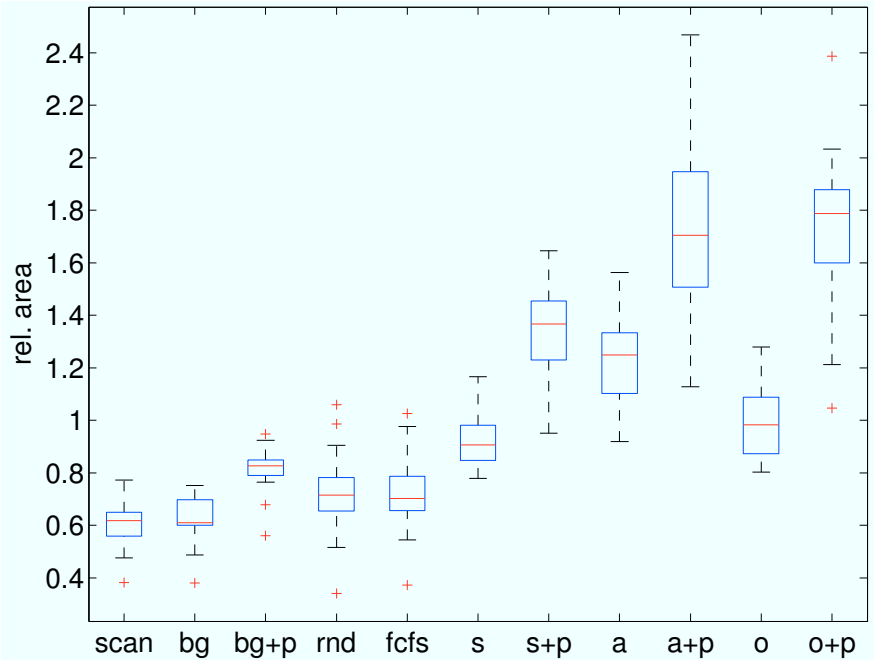


Figure 5.5: Second experiment. Box plot of area covered by different scheduling methods on CAVIAR data set. Measured area is relative to unzoomed ground truth. See text for explanation of labels.

('a'), single('s'), subset ('o') – have been evaluated with and without local scene activity (label augmented with '+p' in the latter case).

The poor performance of scanning, background-only, random and FCFS is easily explained. The first two methods simply scan the parameters in a more or less sensible fashion. They do not react to detected targets at all. FCFS profits from the tie-breaking rule of observing the oldest target next, but both FCFS and the random rule fixate onto an object only for a fixed time, not considering the state of the objects already visited or the duration the rest of the scene has been without observation. Apparent is the increase of the overall coverage when using the local appearance rates. The points of high activity are more often visited than the less active areas of the scene.

5.5 Conclusion and Future Work

This chapter presented a method of scene exploration combined with zoom control. We extended the information theoretic framework pioneered by Denzler et al. [46], in which the choice of zoom, pan and tilt settings is driven by the maximal expected decrease in uncertainty augmented by the likelihood of making an observation. To control the exploration of the scene, we added the uncertainty of a potential, yet unobserved target to this criterion. The chance of an appearance of a target is modelled by local Poisson processes; the chance of making an observation thus rises with the time passed since the last observation of this location. This acts as a counterbalance to the zoom-in behaviour, and yields behaviour in which a target is tracked while its surrounding area is maximally covered by observations.

The zoom onto a current target is discouraged once the expected decrease in uncertainty is higher for a new, potential target which has not yet been detected. We extended this reasoning to multiple targets. Here, the potential acquisition of a new target must provide more information than a subset of targets which can be observed simultaneously. We evaluated the performance of this scheduling policy with respect to existing and new metrics. These were in particular the analysis of latency of the target detection, the increase of observed area, and the number of missed targets.

The proposed method yields a higher detection rate and captures targets at a higher average resolution than other, standard methods used to control active cameras.

5.5.1 Future Work

The approach of modelling an undetected target seems extensible to multiple cameras. One drawback, however, is that the current formulation does not take into account multi-

ple observations of the same location. The formulation in equation 5.12 has a dependency on the field of view and the latest observation time only. This favours overlapping views of several cameras, as the information to be gained from each observation is assumed the same. This is obviously not a good assumption – if a single observation provides enough information to make a decision, then a second observation is not needed.

Furthermore, the actual likelihood of making a detection depends on the resolution of the analysed region of interest. In a visual surveillance context, it is unlikely that observation of a location at a resolution of a single pixel can yield meaningful results.

These issues are addressed in the next chapter, where we will introduce the concept of detector performance with respect to the parameters of the camera.

Chapter 6

Cooperative Scene Exploration

In this chapter we address surveillance with the objectives to detect and track targets with multiple, active cameras. We show that using information theoretic objective functions, these devices can be coordinated to fulfil both tasks in a cooperative fashion. A single camera can address a supervisor functionality if other cameras are better suited to pursue the tracking objective, and turn back to tracking if this behaviour is more germane to the situation it is presented with.

To this end, we build on top of the previous chapter, and extend both terms in the objective function, i.e. the target tracking and target detection parts, to multiple cameras. We show using simulated and real data that this joint information space is sufficient to trigger the desired tracking behaviour, and that a preference for a given behaviour can be chosen by a single, scalar parameter.

Parts of this chapter have been published in [134, 138].

6.1 Introduction

The benefits of multiple cameras in visual surveillance are manifold. For example, several cameras allow a better localisation of targets: as a single camera is a bearing-only sensor, the highest uncertainty in the sensing process is along the line of sight between the camera and the target, and observations from different angles allow us to constrain this uncertainty. Furthermore, the cameras can cover a greater area, and are a prerequisite if the supervised region has a non-trivial geometry, e.g. corners and walls.

As pointed out in the previous chapter, typical surveillance tasks such as identification or classification benefit from a higher resolution of the captured targets, which remote cameras with a wide field of view cannot provide. A solution is the installation of several active cameras with a longer focal length and pan/tilt/zoom (PTZ) functionality.

Often, the resulting systems still employ supervisor cameras to maintain an overview of the scene. In this chapter we propose to exploit the versatility of the active cameras, and explore the viability of surveillance without dedicated supervisor cameras. The active cameras thus need to take over the functionality of the supervisor cameras, but also have to react to the targets being tracked, and are thus in a similar situation as discussed in the previous chapter. This should happen in a collaborative manner; i.e. if one camera tracks a target, another one can safely search the environment for new appearances.

In this chapter, we extend the objective function presented in chapter 5 to take into account the information gained by other cameras. The exploratory term is rephrased in two ways: firstly, it incorporates the actual likelihood of making a detection for a given camera parameter; secondly, the underlying model for the appearance of a target at a given location is changed to become continuously more uncertain without observation. These two modifications allow us to address the case of multiple observations of the same location.

The chapter closes with an evaluation on pre-recorded data, as well as a reference implementation of a live system.

6.2 Tracking with multiple cameras

In this section we extend the mutual information between the state and a single observation to several observations from multiple cameras. In order to compare the information

gain of the same observation region from different cameras, they need a common reference, or coordinate system, in which data from all cameras is integrated. In this regard, we represent the state of a target in the scene in ground plane coordinates, common to all cameras. For this the position of the cameras needs to be acquired. Furthermore, we assume that these cameras are intrinsically calibrated and that the controls have a negligible positional error. Whereas this is a strong requirement – the camera calibration needs to be obtained for every parameter setting of the camera, or the current position estimated using visual odometry – the benefit is the ease of data fusion in a common groundplane.

The mutual information per target is obtained using the approach of Deutsch et al. [48], where the conditional entropy is estimated by a sequential approximation of running a single Kalman filter with variable number of potential observations.

6.2.1 Merging of Kalman Filters

We maintain the assumption of target independence, thus the mutual information is a sum over the information contributed by each of the targets. The mutual information of the state and several observations is decomposed similar to equation 4.1, but now for all C observations from all C cameras:

$$I(\mathbf{x};\{\mathbf{o}_1 \dots \mathbf{o}_C\}) = H(\mathbf{x}) - H(\mathbf{x}|\{\mathbf{o}_1 \dots \mathbf{o}_C\}) \tag{6.1}$$

For the conditional entropy, all possible combinations of observations need to be taken into account. When using Kalman filters, this results in 2^C different possible covariance matrices, as well as 2^C different weighted terms in the expectation, one for each different covariance matrix.

For example, the conditional entropy for 2 cameras is

$$H(\mathbf{x}|\mathbf{o}_1, \mathbf{o}_2) = w_{11}H(\mathbf{x}_{11}) + w_{10}H(\mathbf{x}_{10}) + w_{01}H(\mathbf{x}_{01}) + w_{00}H(\mathbf{x}_{00}) \tag{6.2}$$

The values w_{00}, w_{01}, w_{10} and w_{11} correspond to the likelihood of making no observation (w_{00}), a single observation in one of the cameras (w_{01},w_{10}), and observing the target in both cameras (w_{11}). Each of these terms is expressed in the same way as the visibility term in equation 4.13.

When keeping in mind that this sum needs to be evaluated for all possible parameter combinations (for active cameras with 3 parameters, pan, tilt and zoom, this results in a parameter space of 3^C dimensions), it is clear that this becomes unwieldy very quickly for several cameras.

To simplify this calculation, we use the approach by Deutsch [48], where instead of having a Kalman filter with variable number of observations, each camera's observation is treated independently, and the updated state fed into the next filter.

This "sequential" Kalman filter makes a single prediction step, taking target state $\mathbf{x}_{k-1}^{+}$ and covariance matrix $\mathbf{P}_{k-1}^{+}$ to a predicted position $\mathbf{x}_{k}^{-}$ and $\mathbf{P}_{k}^{-}$, taking into account the uncertainty of the motion model. This prediction is updated once for every observation made by each camera, which is valid if the measurement noise of the cameras is uncorrelated. This is a common assumption [13, 48]. However, for each update, the observation matrix $\mathbf{H}$ is linearised anew at the estimate produced by the incorporation of the previous observation, and the covariance matrix from the last update is used.

Unfortunately, no future observations are available at the point where a decision has to be made about the control for a camera, and thus no updated covariance. As discussed in section 3.4.1, the best we can do is to use the expected covariance. The expectation comprises two elements: one is the updated covariance, which is obtained with the likelihood of making an observation, given by the visibility term $w(\mathbf{a})$ in equation 4.13; the second is the previous covariance, with the likelihood of not making an observation. This

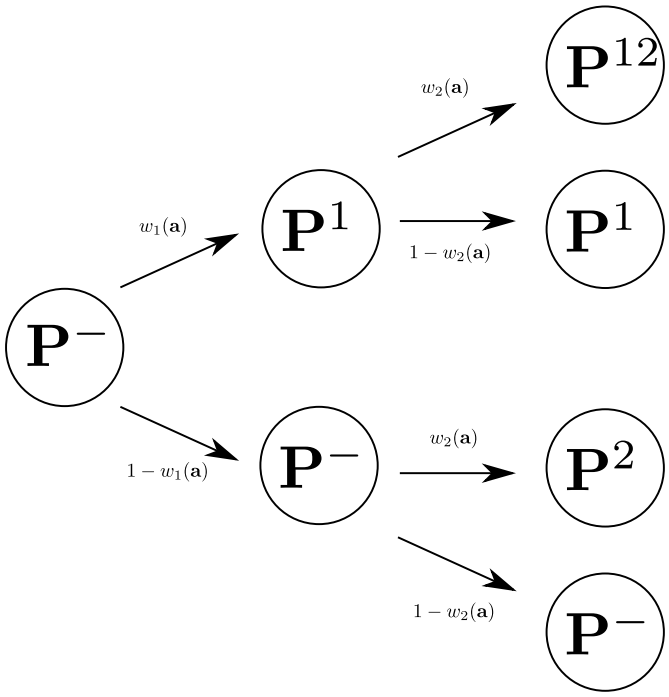


Figure 6.1: Multiple updates to the covariance results in a mixture model of 2^C elements. The first update gives two results, the updated covariance corresponding to a successful observation by the first camera ($\mathbf{P}^1$), and the original covariance ($\mathbf{P}^-$) if the observation is not successful. The next update results in four possible outcomes, e.g. $\mathbf{P}^{12}$ if observations were made by both cameras.

yields a mixture of Gaussians with two elements, located at the same predicted state $\mathbf{x}_k^-$. Propagating this through the set of filters results in a Gaussian mixture of 2^C elements. This is sketched in figure 6.1.

Thus the second simplification is to approximate this mixture after each step with a single element. This is then fed into the second filter and updated again, see figure 6.2. The update for a particular camera c is then

$$\begin{aligned}
\mathbf{P}_c^* &= w(\mathbf{a}_c)\mathbf{P}_{c-1}^+ + (1 - w(\mathbf{a}_c))\mathbf{P}_{c-1}^- && (6.3)\\
&= w(\mathbf{a}_c)(\mathbf{I} - \mathbf{K}_c\mathbf{H}_c)\mathbf{P}_{c-1}^- + (1 - w(\mathbf{a}_c))\mathbf{P}_{c-1}^- && (6.4)\\
&= w(\mathbf{a}_c)(\mathbf{I} - \mathbf{K}_c\mathbf{H}_c)\mathbf{P}_{c-1}^- + (1 - w(\mathbf{a}_c))\mathbf{P}_{c-1}^- && (6.5)\\
&= (\mathbf{I} - w(\mathbf{a}_c)\mathbf{K}_c\mathbf{H}_c)\mathbf{P}_{c-1}^- && (6.6)
\end{aligned}$$

The input covariance matrix to the update for the first camera, $\mathbf{P}_0$, is the predicted covariance $\mathbf{P}^-$. The matrices $\mathbf{K}_c$ and $\mathbf{H}_c$ are the Kalman filter gains and observation Jacobians for the corresponding camera.

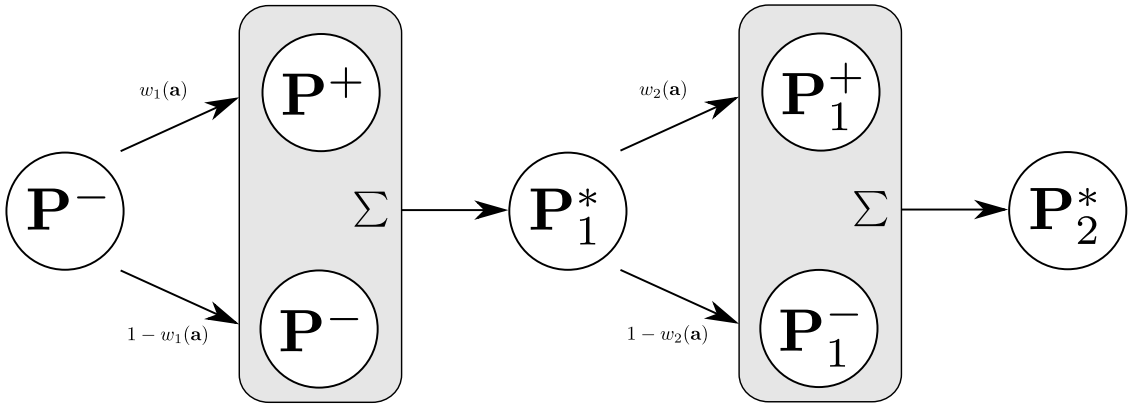


Figure 6.2: For each camera, the resulting updated covariance $\mathbf{P}_c^*$ is approximated by averaging over all possible outcomes.

The resulting covariance of a single target observed by a set $C = \{c_1, \ldots, c_n\}$ of cameras, which can be a subset of all cameras C^*, is thus the product of all Kalman filter gains:

$$\mathbf{P}_k^+ = \left(\prod_{c \in C} (\mathbf{I} - \mathbf{K}_c \mathbf{H}_c) \right) \mathbf{P}_k^- . \tag{6.7}$$

Using the closed form for the differential entropy of a Gaussian distributed state vector in equation 3.3.4, the conditional entropy of the sequential Kalman filter with covariance given in equation 6.1, reduces to

$$H_{\mathbf{a}}(\mathbf{x} | \{\mathbf{o}_1 \ldots \mathbf{o}_C\}) = \frac{n}{2}(1 + \log(2\pi) + \sum_{c \in C} \log |\mathbf{I} - w(\mathbf{a}_c)\mathbf{K}_c \mathbf{H}_c| + \log |\mathbf{P}_k^-|). \tag{6.8}$$

The mutual information for a single target is then

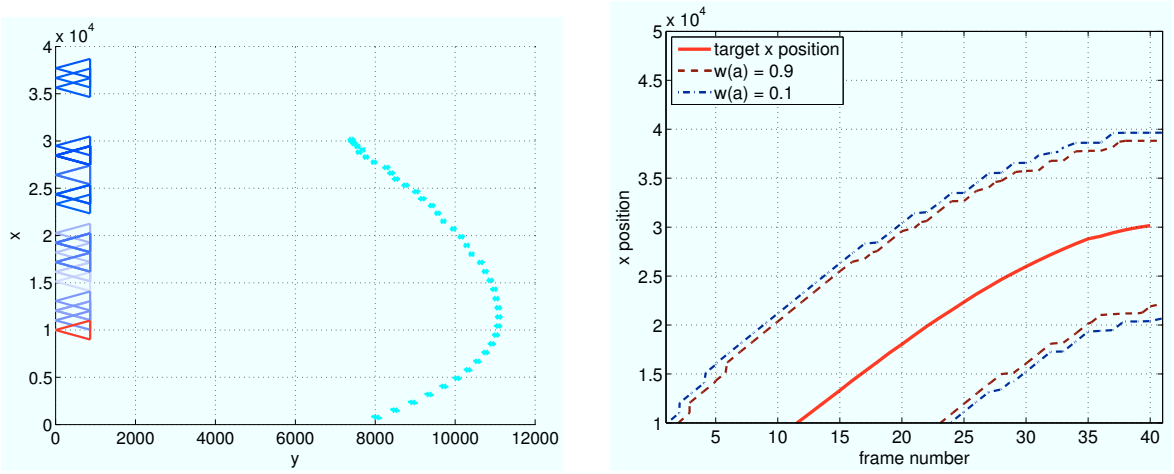
$$I_{\mathbf{a}}(\mathbf{x}; \{\mathbf{o}_1 \ldots \mathbf{o}_C\}) = H(\mathbf{x}) - H_{\mathbf{a}}(\mathbf{x} | \{\mathbf{o}_1 \ldots \mathbf{o}_C\}) = -n/2 \sum_{c \in C^*} \log |\mathbf{I} - w(\mathbf{a}_c)\mathbf{K}_c \mathbf{H}_c|. \tag{6.9}$$

6.2.2 Examples

To highlight the influence of this objective function on the parameter selection, we created some simple artificial scenes. Here, a single target is moving on a ground plane and has no spatial extent, i.e. the observation yields a point in the image plane. The target moves on a predefined B-spline curve of order 3, which exhibits second order dynamics. The

motion model of the Kalman filter is linear, i.e. only first order dynamics of the target are expected.

In the first scene setup, the target is observed by two cameras, both looking in the same direction. Neither has zoom capabilities. The first one is fixed, and the second one can move along a part of the x-axis to follow the target, but otherwise has the same observation parameters as the first. A top view of this scene is shown in figure 6.3(a). Given that the cameras are sufficiently far away from the target, the choice of the camera position will not have an influence on the linearised observation model nor the target's covariance matrix; instead, it will have an impact on the visibility term in equation 6.6 only. Thus the maximal mutual information for a single target will simply maximise the probability of making an observation, which is given in equation 6.6. The contour plot 6.3(b) shows this visibility term for varying positions at different time steps. It forms a plateau, where the chance of making an observation is 1, which falls off to 0 according to the predicted positional uncertainty and observation noise. The selected camera position is anywhere on this plateau, which follows the expected target movement. In the second artificial scene, the second camera is now also allowed to zoom in discrete steps between 1 and 12. This setup and the resulting camera positions are shown in figure 6.4(a). The zoom setting influences the Jacobian of the observation model, and therefore the expected covariance of the target, as in equation 6.8. The objective function for these parameters is shown as a contour plot for three frames in figure 6.4(b). The initial zoom value is limited because of the high uncertainty after initialisation of the Kalman filter. Mutual information first decreases as more observations at small zoom levels reduce the initial uncertainty, then increases due along with rising zoom values. Once the target is tracked as the highest zoom level the mutual information gradually decreases. Finally, the target leaves the



(a) Scene setup

(b) Contour plot of visibility

Figure 6.3: (a): Top view of the simple scene setup. The first camera (red) remains fixed. The second camera (blue) follows the target, maximising expected visibility. Lighter tones correspond to earlier frames. (b): The resulting visibility term for a number of camera positions at every step, given for a likelihood of 0.1 and 0.9. The predicted target position is drawn in red/bold.

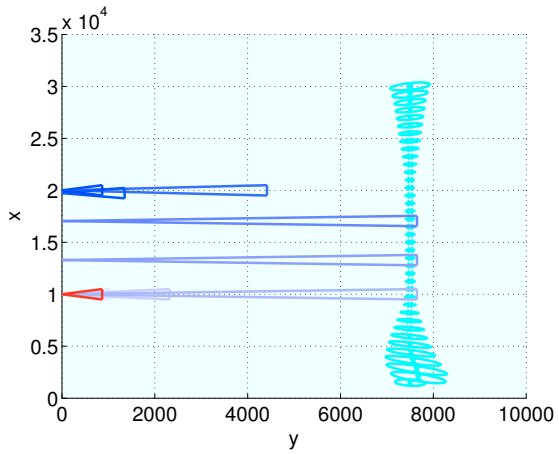
field of view of the first camera and the second one has to zoom out to maintain successful tracking.

6.2.3 Prior Knowledge

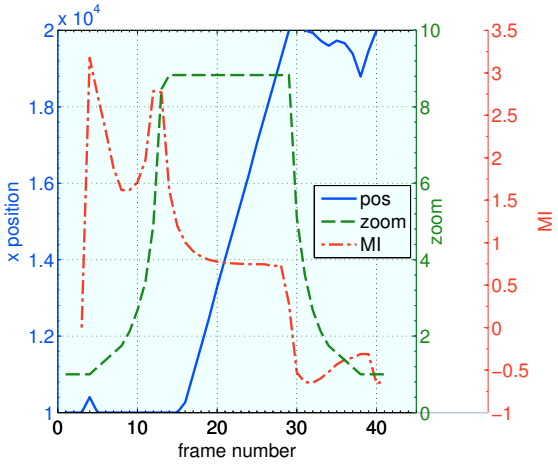
The visibility term in equation 6.6 is amenable to incorporation of prior knowledge about the scene. In this example we add a visibility factor to the expected visibility from the tracking process, to add information about the use of the specific camera for a given parameter setting. This information can come from a floor plan or from the current target states to take into account occlusion of one target by another.

To show collaborative tracking of cameras with mutual information from multi camera observations, we set up an artificial scene motivated by a real world data set with multiple cameras and manually annotated ground truth[1].

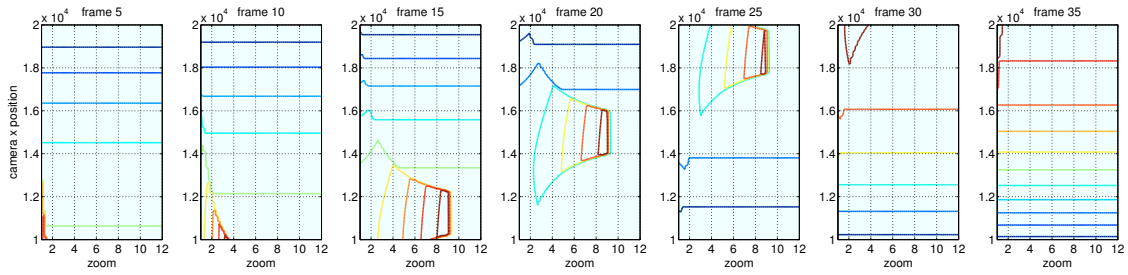
[1] PETS 2001 data set: http://pets2001.visualsurveillance.org



(a) Scene setup



(b) Entropy, zoom and translation



(c) Objective function

Figure 6.4: (a): Top view of the scene setup. The second (blue) camera follows the target and zooms onto the target, trading expected visibility for decreased entropy. Lighter tones correspond to earlier frames. (c): The resulting entropy term for all positions and zoom settings for three frames. (b): Contour plot of mutual information over camera parameters (position and zoom) for each of the highlighted positions in (a). Darker/warmer tones correspond to higher values.

We model the state of the targets as a 3-dimensional bounding box moving on a ground plane with linear dynamics. The observation model yields the two-dimensional bounding box of the projected vertices of the target. The only parameters varied here are the zoom settings for both cameras. Figure 6.5 shows the tracking of a pedestrian in said scenario. The rectangles mark the bounding box of the actor, whose path is occluded artificially in the centre. For this, the term $p(\mathbf{o}|\mathbf{x})$ in equation 4.13 (which yields the tar-

get's likelihood of being visible) is set to 0 at all pixels where the occluder is closer to the camera.

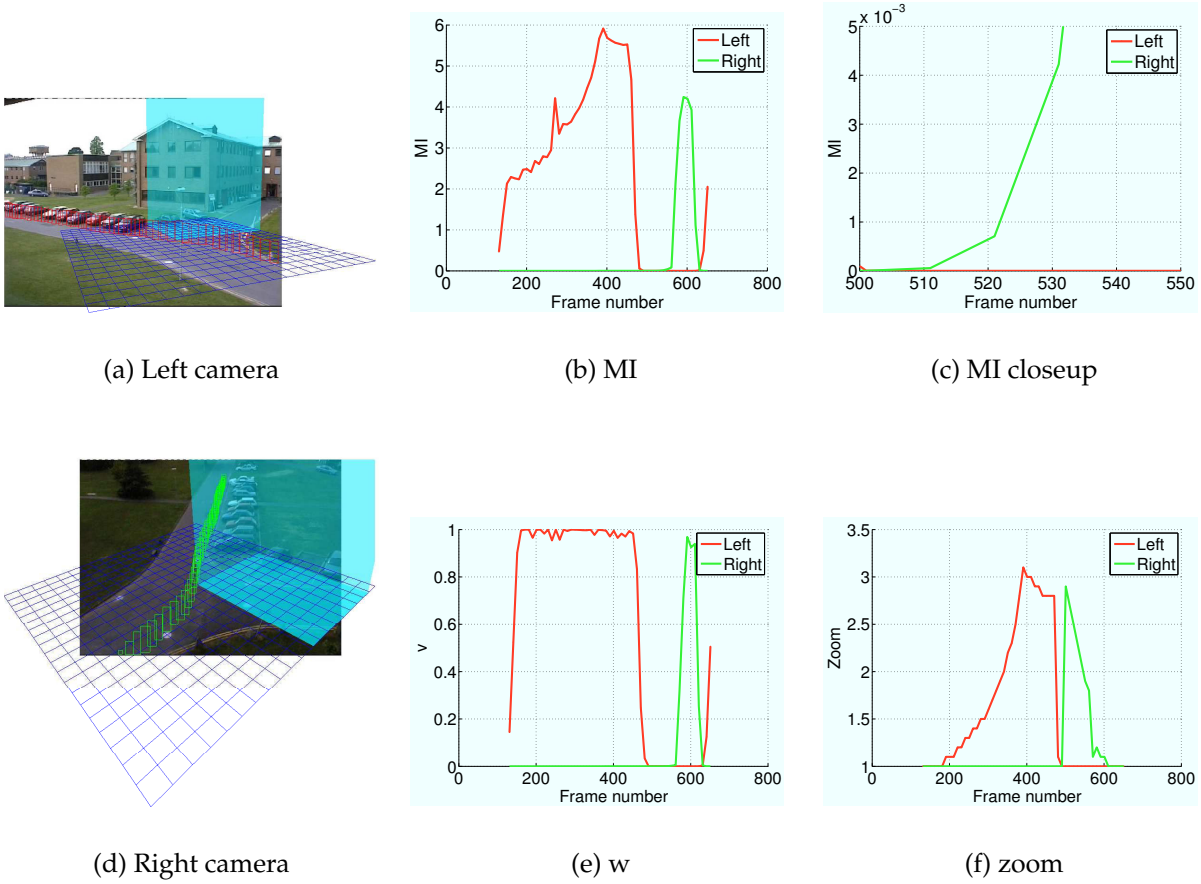
The plots 6.5(b) and 6.5(c) show the mutual information of a single moving target, and plots 6.5(e) and 6.5(f) show the visibility term $w(\mathbf{a})$ and the chosen zoom factor.

At the start of the sequence, the mutual information for the right camera is close to zero, because the view to the target is occluded. As long as the other camera observes the target, the position estimate is accurate enough to be sure that the target is still blocked from view. Once the target is lost by the left camera, the mutual information rises since the uncertainty of the target's position increases – hence an observation might be made in the area surrounding the blocked view. This behaviour is shown in the close-up of the development of the mutual information in figure 6.5(c). This behaviour is sensible in that there is no other objective for the first camera. As soon as another target of interest is available, the camera will focus on this.

6.2.4 Live system

To get an impression how the objective function influences the choice of targets, we implemented an initial version of the control method in a live system, where the targets are acquired and tracked by a supervisor camera, and two active cameras are slaved according to the output of the objective function 6.1.

The information sent to the objective function comprises the target positions and constant covariance matrices, as the supervisor camera has a near top-down view on the supervised area. With this method we test the initial assignment of targets to each of the cameras, as if for every location, targets were detected anew. Furthermore, we modify the visibility term $w_c(\mathbf{a})$ in the objective function according to a floor plan that specifies the visibility of a location for each of two active cameras. The top row of figure 6.6 shows



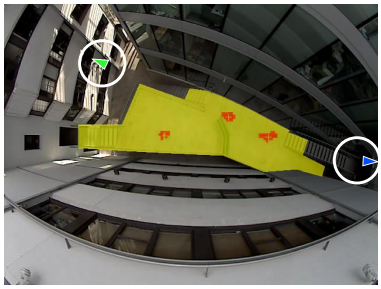
(a) Left camera (b) MI (c) MI closeup

(d) Right camera (e) w (f) zoom

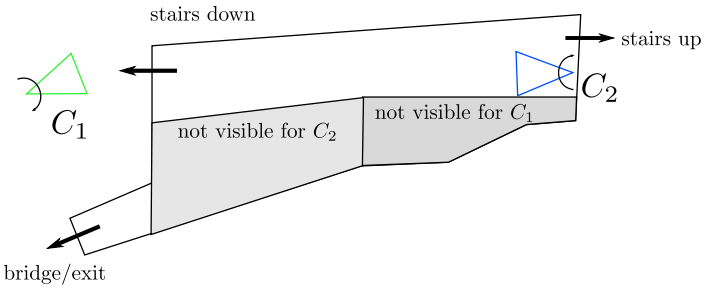
Figure 6.5: Trajectories of first actor of the PETS 2001 data set with superimposed ground-plane in left and right view. Only every 10th frame is shown. Plot 6.5(e) shows the likelihood of making an observation for both cameras (left camera:red, right camera: green). The resulting zoom setting for the two cameras is shown in plot 6.5(f). See text for details.

the view of the supervisor camera, and the floor plan including the position of the active cameras. The presence of a screen mid-way along the atrium creates areas of blocked visibility for each active camera.

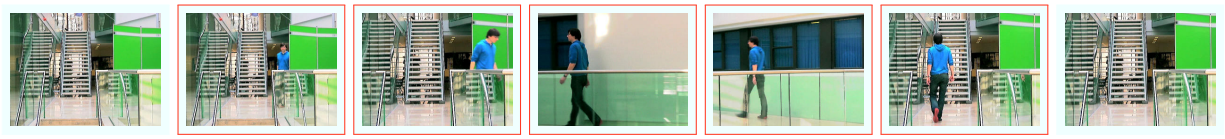
In the experiments performed, the cameras were following the targets (albeit with lag, as they are basically slaved to the output of the supervising tracker). The tracking of one of the cameras stopped once the target was about to enter the area that was not visible, and resumed once the target reappeared in this target's field of view.



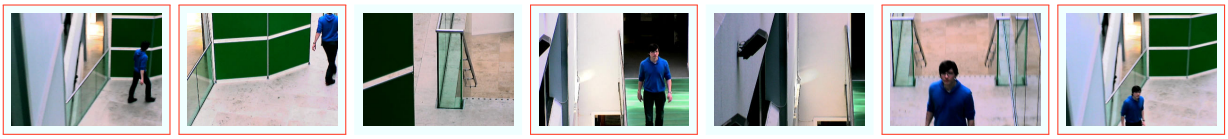
(a) Top view



(b) Floor plan



(c) View camera 1



(d) View camera 2

Figure 6.6: Live experiment with modified visibility term. The target is handed over upon appearance in the second camera's field of view. Boxed frames come from a camera that is "active", and has non-zero mutual information.

6.3 Scene exploration using Multiple Cameras

We now address the information gain from detecting new targets. The approach in section 5.4.1 introduced an extra term to the objective function which models the appearance of a single target. This approach fails to take into account two important factors: One is the integration of the likelihood of actually detecting the target, depending on its presence or absence. Before targets are tracked in a visual system, they need to be discriminated from the image background. The performance of these methods – also known as detectors – depends on the resolution the location is observed at, e.g. given sufficient noise in the imaging conditions, there is little use of investigating a single pixel to make a decision, or closing up on the tip of the nose of a person.

The second factor which the previous method fails to address is the observation of the same location with multiple cameras. The more cameras that observe the same location, the better the resulting estimate will be. But the underlying question is how much improvement over one fewer camera this will yield, as the extra camera might be serving better by observing a different place. Another point is how the information gain from detecting new targets weighs against the information from continued tracking.

This section shows how to use the performance of the detector to yield a method which gauges the expected increase in terms of the information gained, and that it is legitimate to compare it with the information from tracking targets. For this, we derive expressions for the mutual information resulting from a search for targets, and quantify the information gain arising from a particular detector algorithm used at a particular location. The information to be gained from detecting targets at a given parameter setting is then the sum over all visible locations.

The overall approach is similar to section 5.4.1, i.e. we discretise the supervised area into a disjoint set of locations, and affix a likelihood of target presence to each of these. These locations need not be confined to a certain geometry such as a ground plane, but have to be observable by at least one of the cameras. One crucial difference is that the information is directly compared against the information from tracking, without the detour through a target that has not yet appeared. This makes use of the approach to combine different objectives which we have addressed in section 3.6. We will explain this in detail in the next sections.

6.3.1 Prior: Birth-and-death process

In this section, we discuss the prior likelihood of target existence at a single location, and how it influences the uncertainty in the observation process.

First, we need to point out that a Poisson process alone cannot be used as the basis of an information theoretic objective function. Unlike to the previous chapter, we are looking for a function that yields a higher uncertainty of target existence for longer periods without observation. This is not true for a Poisson process. Its entropy is not monotonously increasing with respect to time: after a certain time, the probability of a target's appearance increases above 0.5 and the entropy begins to *decrease*. This is natural, as it is more likely that an actor has appeared, which provides less information to gain from observing the location.

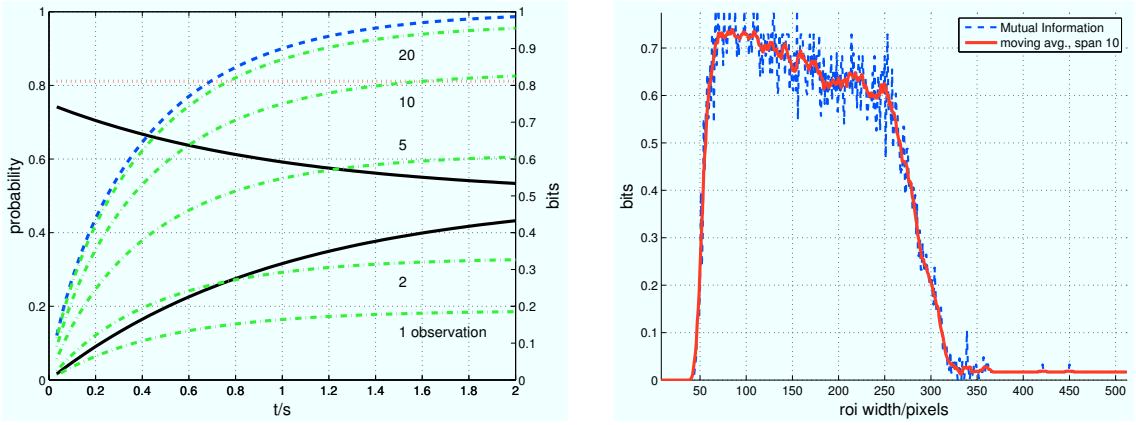
Instead, we propose a different model: just as the likelihood that a target appears rises, so does the likelihood that it moves on to another place, and thus disappears from its current location. The result is that the longer a place has not been observed, the less certain we are about its occupancy.

The mathematical model for this kind of behaviour is known as Birth and Death process [62], for which a small summary can be found in appendix B. The Birth and Death process has two competing rates: the birth rate λ, analogous to the appearance rate of the Poisson process, and the death rate ν, at which a target disappears. For each scene location we model the existence of a target with such a process, and by assuming that appearance of an object is equally likely as a disappearance, the birth rate λ is equal to the death rate ν.

When the time step of the last observation of a location is k_0, then the probability of existence e of a target at a given location at time step k is

$$p(e(k)) = (\alpha - 0.5)e^{-\lambda(k-k_0)} + 0.5, \tag{6.10}$$

where $\alpha = p(e(k_0))$ represents the initial uncertainty. A single time step thus triggers a "forgetting" of the current state. Whenever a location is observed, k_0 is set to the current



(a) (b)

Figure 6.7: (a): Birth-and-death process for fixed λ at a single scene location, at two different starting conditions $\alpha = 0$ and $\alpha = 0.75$ (black/solid lines), and entropy $\alpha = 0$ (blue/dashed). Green/dot-dashed: Mutual information gain for 1,2,5,10 and 20 observations of the same location, with detector performance $H(d) \approx 0.8$ bits (red/dotted) (b): Typical detector performance: The MI gain as a function of face size in the image degrades at higher resolution because false positives become more likely (red line: is the moving average over 10 pixels).

---

k, and α is set to an initial value based on the detector performance (e.g. for a perfect detector, α is set to 1 or 0 depending on the detector output).

The development of the probability of existence is portrayed in figure 6.7(a), for an initial probability of $\alpha = 0$ and for $\alpha = 0.75$. The probability of the existence of a target approaches the maximally entropic value of $p(e) = 0.5$ for $k \to \infty$.

6.3.2 Observations: Detector performance

The accurate detection of an object at a location depends on the actual method used, and the sensor parameters. In particular the zoom level will affect the resolution at which the target is imaged, and hence the performance of a detector. This can be characterised by two functions of zoom level z, $p_z(d|e = 0,1)$, (i.e. the chance of a detection given existence or not) representing the performance in terms of true and false positives. An example of such a curve, for the OpenCV implementation of face detection evaluated on the Point-

ing'04 dataset [66], is shown in figure 6.7(b). [2] The performance peaks at a favoured size of 50-100 pixels, corresponding to the size of the images in the training set.

The expected width of a location under observation is obtained from projecting a three-dimensional bounding box of the location onto the image plane, whereas the height over the ground plane is the average height of the targets that are to be detected. In this work, we use an average height of $1.77m$[3]. The resulting width is used to index the detector performance.

The final mutual information gain at a single location at time t is then a function of zoom

$$I_z(e;d) = H(e) - H_z(e|d) = H_z(d) - H_z(d|e) \tag{6.11}$$

dependent on the birth-and-death process, equation (6.10) (yielding the term $H(e)$). and the detector performance $p_z(d|e)$.

6.3.3 Multiple observations

Several cameras can observe the same scene location, but this is potentially a waste of sensing resource: certainly more observations will provide more information, but an observation of another location might be more sensible. Here we characterise the mutual information for detections at the same location from multiple cameras.

Assuming independence between the set of observations $\{d\}$, we have:

$$p(e|\{d\}) = p(\{d\}|e)p(e)/p(\{d\}) = p(e)\Pi_i p(d_i|e)/p(\{d\}). \tag{6.12}$$

[2] To count correct detections, a bounding box overlap criterion is used [4, 90], where positional errors must not exceed a percentage of the overall extent of the ground truth annotation in location and scale space. Out of multiple detections that meet this criterion, only one is counted as true positive, and the others as false positives. The negative examples for evaluation are images from the Caltech 101 object database [56], excluding face images. Face images were up- and downsampled to the test width.

[3] $177.6cm$ is the mean height for men aged $25-34$, NHS Health Survey for England, 2008

The resulting conditional entropy for C observations is then

$$\begin{aligned} H(e|\{d_c\}_C) &= -\sum_e p(e) \sum_{d_{1\ldots C}} p(d_C|e) \log(p(d_C|e)) \\ &= H(e) - H(d_C) + \sum_{c=1\ldots C} H(d_c|e) \end{aligned} \tag{6.13}$$

Figure 6.7(a) shows the mutual information for increasing numbers of observations: a single scene point is observed continuously by 1, 2, 5, 10 or 20 cameras, which all have the same detector performance (i.e. the cameras are arranged around the location in a circular fashion). While the mutual information nearly doubles when going from a single camera to two, it only triples when using 5 cameras, and in order to increase the information gain further by a similar amount, the number of observations need to be quadrupled. The mutual information does indeed increase for more observations, the improvements are getting smaller and smaller (this is the diminishing returns property of mutual information, see section 3.3.6). For better raw detector performance the effect is more pronounced (a perfect detector would have $H(d) = 0$ and no further observations would add information). This trade-off is important for the collaborative exploration of the scene by several cameras – extensive overlap of the supervised area does not necessarily yield more information than a disparate setting.

The information gain for C cameras and N locations is thus

$$I = \sum_{i=1\ldots N} H(\{d_{i,c}\}_C) - \sum_{c=1\ldots C} H(d_{i,c}|e_i) \tag{6.14}$$

The important term is $H(\{d_{i,c}\}_C)$, which is the joint entropy of all measurements for location i. For any given location, this rises to a maximum of one bit, with the rate of increase determined by the detector performance, the number of observations and the birth-and-death rates λ.

Note that the total information gain from observing the scene is a relatively simple formula calculated from the detector performance characterisations and the birth-and-

death process. For a number of camera parameter settings, the set of observable locations, as well as the expected detector performance can be pre-computed.

6.3.4 Implementation and behaviours

We quantise the pan and tilt values into M values (not necessarily evenly spaced) and zoom into N steps. The choice of parameters then reduces to an exhaustive search over the $(M^2N)^C$ parameters. For modest C (i.e. 2 or 3) the search space is not unreasonably large, but rapidly becomes unwieldy for four or more cameras.

We illustrate the performance of the mutual information objective function on prerecorded data from the PETS data set. We use one camera, with 36 different pan and tilt settings ($M = 6$) and $N = 4$ zoom settings. As the data is prerecorded, we use an affine mapping only, i.e. pan, tilt and zoom parameters select a different subsection of the original image, and resample the resulting section to a constant size.

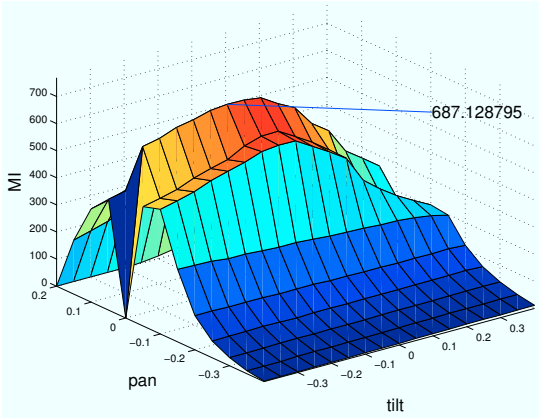
Figure 6.8 shows the evolution of expected mutual information over time for the selected zoom setting $z = 3$. At each time step, we choose the set of values that yield the maximum MI. Note how immediately after an observation at a particular location (pan-tilt setting), the gain in MI is significantly reduced.

6.4 Combining objectives

As explained in section 3.6, we can combine two information gains from detection and tracking via linear blending, which yields a combined utility for both goals – exploration and investigation – of the control:

$$U = \zeta I_{T,\mathbf{a}}(\mathbf{x};\mathbf{o}) / I_{T,max} + (1-\zeta)\hat{I}_{N,\mathbf{a}_t} / \hat{I}_{N,max} \tag{6.15}$$

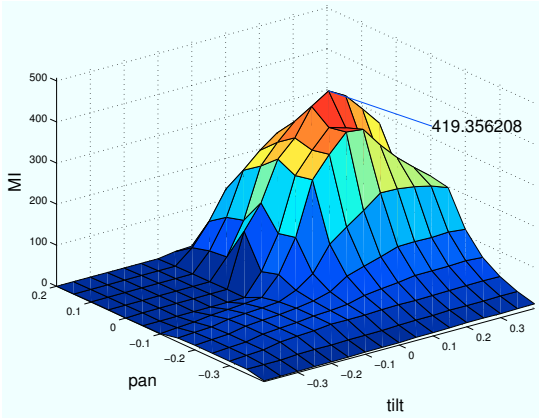
One problem is that the entropy of a continuous probability variable is theoretically unbounded, and can thus not easily be compared with the uncertainty of discrete state



(a) MI over pan, tilt, $z = 3$, $t = 10$



(b) corresponding image



(c) MI over pan, tilt, $z = 3$, $t = 20$



(d) corresponding image

Figure 6.8: Plots 6.8(a) and 6.8(c) show mutual information per pan,tilt-setting at constant zoom for a camera in the PETS 2001 sequence at time $t = 10$ and $t = 20$. After the first observation, the expected gain around the observed area is reduced and another location is chosen in the second step.

spaces. In practice, however, the conditional entropy in continuous state spaces is bounded: the upper limit depends on the maximum uncertainty that is tolerated in a tracker before it is deemed to have failed; e.g. when the uncertainty in the state space encompasses the size of the observation area. The lower limit depends on the observation model. In the Kalman filter case, this bound $H_{KF,\text{min}}$ can be obtained by the steady state solution for the state's covariance matrix at the highest zoom level. Thus both normalising constants

$I_{T,max}$ and $I_{N,max}$ can be obtained *a priori* from the contributing target distributions. The parameter ζ can be seen as the factor that balances between different objectives.

6.5 Experiments

For the experiments, we model the targets as 3 dimensional bounding boxes on a ground plane with process noise of 30cm per frame, and pixel noise of one pixel at smallest zoom. The scene locations are 1 by 1 square metre cells, and the appearance rate is one actor per second for the whole scene. For repeatability of experiments, we use ground truth data with artificial noise. As detector performance we used the one given in section 6.3.2, arguing that any other template or code-book oriented detector also has a fixed training size.

We evaluate the combined system using three metrics, each of which concerns the performance of one aspect of the system: (i) *Resolution:* the average increase in resolution over all targets, based on the observed area of every target in every frame compared to the unzoomed, ground truth case; (ii) *Latency:* measures the average time taken for the system to locate a new (or previously lost) actor in the scene; (iii) *Fragmentation*, which is the average number of splits in trajectories, i.e. how often the target was completely out of view of all cameras, and then reacquired.

Clearly not all of these metrics can be maximised simultaneously. In figure 6.9 we show the values of each metric as a function of the blending weight, ζ, computed for the PETS2001 dataset. When ζ is close to zero, the scene term dominates the objective function yielding low latency, but at the cost of low resolution and highly fragmented actor trajectories. In contrast, ζ close to one yields low fragmentation and high resolution

tracking (an average zoom of 3 from a maximum of 4), but at the cost of longer delays in detecting new actor arrivals.

We compared our method with standard rule based scheduling methods, i.e. random selection of targets and the first come, first serve rule (FCFS). Both methods follow a target for a given number of frames (10). In the random case, a new target is selected randomly, whereas in the FCFS case, targets are prioritised by age, i.e. the "oldest" chosen first. Once a target has been observed for 10 frames, a "younger" target is chosen, if one is available.

Furthermore we evaluated a simple scanning method, which is a standard method for surveillance systems[4]. Here each camera simply visits all pan, tilt positions in a boustrophedon manner: The camera pans first left-to-right until the maximum setting is reached, increases the tilt value and pans right-to-left, increases the tilt parameter once more and panning left-to-right again. Once the maximum pan, tilt settings have been reached, the camera moves back to the minimum settings. We change this method slightly by tracking a target for 10 frames, once a target enters the field of view.

As can be seen in table 6.1, our method clearly outperforms the other approaches due to the active nature of our method, i.e. direct reactions to the targets in the scene. Whereas the standard methods seem to have a smaller fragmentation, this is due to the fact that a target is usually observed for a short period, which is reflected in the small average resolution and high latency. The performance also degrades significantly once the cameras are treated independently, i.e. no information is propagated between the sensors in the maximisation step. Even though tracking still happens via the shared Kalman filters, there is no matching between the local aims of the sensors, which leads to smaller zoom levels than in the coupled variant.

[4] This is also called "guard tour", see e.g. Axis PTZ 214 Manual `http://www.axis.com/products/cam_214/index.htm`, last accessed 11.01.2011

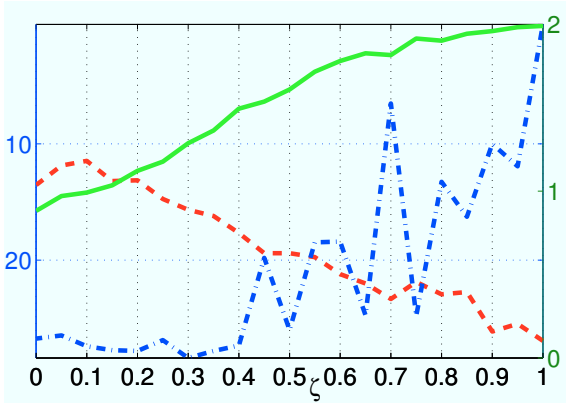


Figure 6.9: Performance metrics as a function of ζ. Red/dashed: Fragmentation; blue/dotted: Latency (both left ordinate); Green: resolution improvement. ζ mediates between scene mutual information ($\zeta = 0$) and tracking mutual information ($\zeta = 1$). The peak in latency at 0.70 is due to the system missing the longest trajectory in the scene.

Table 6.1: Performance comparison of mutual information based scheduling with standard methods. The table shows the metrics fragmentation, resolution and latency of the resulting trajectories. scan@N: Independent scanning at zoom level N. fcfs, random: scanning at zoom 2 and further zoom onto first or randomly chosen target. MI, ind: maximisation of MI for each camera independently, at $\zeta = 0.75$

method	fragmentation	resolution	latency
scan@4	1.06	0.41	138
scan@3	1.13	0.55	121
scan@2	1.06	0.64	96
fcfs@2	1.22	1.28	100
random@2	1.23	1.21	101
MI, ind.	16.4667	0.8821	3.2667

6.6 Conclusion

This chapter has addressed the problem of tracking targets at increased resolution, as well as detection of new targets without supervisor cameras. This poses conflicting objectives, as a camera can either follow a target, or investigate the environment for arrivals. Whereas some previous work on multi-objective planning exists in the robotics community, our contribution seems to be among the first that address this problem in the surveillance context.

The main contribution of this chapter are two approaches to address this problem. We use decision theoretic insights that information-theoretic utilities allow a linear combination to address any desired preference.

In both approaches, and completely novel to the field, is the underlying idea to explicitly model the uncertainty about target existence in the supervised region.

The first approach does so by anticipating a single target, where the likelihood of appearance has been learned from long term, low resolution scene observations. The second approach introduces a dependency on the detector performance. The detector can effectively be seen as a sensor, which has varying performance for different depths. The introduction of the detector performance into the exploration process is another contribution of this work. The result is coverage of the supervised area to maximise expected likelihood of making detections.

Furthermore, we addressed the simplification of mutual information from multiple expected observations using the sequential Kalman filter. In this formulation, it becomes straightforward to add external constraints, such as target inter-occlusion terms or other visibility factors that come from scene knowledge. In section 6.2.4, we have presented a real-time implementation of a control rule using this approach, confirming the expected behaviour of target hand-off between cameras.

Lastly, we have shown that our proposed method yields improved performance on standard data sets, when comparing it to standard methods.

6.6.1 Future Work

The main limitation of the current approach is that the performance is severely limited by the number of cameras. The information theoretic basis of the methods proposed in this chapter make work from the robotics community directly applicable. where distributed

data fusion and control methods have been pioneered[69, 98]. A game theoretic approach to camera control has recently been introduced by Song et al. [140], where a consensus among best actions is found among neighbouring devices.

Furthermore, the detector performance in section 6.3.2 currently addresses scale changes only. For cameras placed close to the targets, the influence of the tilt angle becomes relevant. A natural extension is thus to learn the detection performance for varying view directions, and to integrate this into the observation performance.

Another simplification in our method is the reacquisition likelihood of targets. Currently, it is assumed that once a target has been acquired and tracked, it will also be reacquired when the best observation parameter is a move away from this target. In general, there is no penalty for excess movement of the cameras. This can be addressed by simple addition of a quadratic cost function to the objective function. The natural way, though, is to introduce likelihoods that motion of cameras risks losing targets, i.e. that the tracking process can fail.

Chapter 7

Target Identification

In this chapter we address camera control in the context of identification systems, i.e. where the main purpose of the system is classification of the targets in the supervised area. Typically, these systems have been treated as an instance of network routing systems, and the uncertainty in the sensing process neglected. We address two types of uncertainty. The first uncertainty is the one intrinsic to the identification process, i.e. the classifier employed to actually identify the subjects. We show how this uncertainty influences the performance of the system when selecting targets for observation.

The second type of uncertainty we address lies in the gaze direction. We use the output of a gaze direction method to choose targets that are looking towards the camera, by taking into account the current mean direction and the uncertainty in the estimate. We show in simulation how using the gaze direction in the target selection process improves performance with respect to the number of successful identifications of targets. We furthermore present a live system implementation of the latter.

Parts of this chapter have been published in [139].

7.1 Introduction

The accuracy of face recognition in high resolution images has improved steadily over recent years [117], however in common with most other methods for biometric identification, cooperation from the subject is needed to acquire the required measurement. The ability to automatically capture high resolution face images from a remote camera would

allow face recognition systems to be deployed in surveillance situations where people are free to move without constraint.

When only a single person is being monitored, an active camera can simply follow them until the required image is captured, but when multiple persons are present, the camera controller must make a decision about which target to follow. Since a face image can only be captured when a surveillance subject looks towards the observing camera, advance knowledge of where the people in a scene are looking can be used to guide the choice of person for an active camera to follow.

In this chapter we show how to use the performance of the detection system to select the targets that one or more active cameras should observe, with the aim of maximising the likelihood that the targeted subjects will be correctly identified. We further show how gaze estimates from a static camera can be used. In particular we make four main contributions. First, we show how methods for tracking and control based on expected mutual information gain can be used to prioritise targets. Furthermore, this can be naturally extended to make use of gaze information. Third, we demonstrate that the use of the (noisy) gaze data is indeed beneficial, yielding improved performance over no knowledge of gaze, or the first-order approximation that a person's gaze direction and their direction of motion are always the same. Finally, we demonstrate a real-time implementation with a static camera and two pan-tilt-zoom devices, involving real-time tracking, processing and control.

7.2 Related Work

To date a significant body of research has taken the approach of using static cameras to guide the control of active cameras in distributed visual surveillance systems. March-

esotti et al. [99] used tracking in a single static camera to guide an active camera to capture face images. A similar approach was used by Hampapur et al. [72] but with multiple active cameras. Qureshi and Terzopoulos [121] and Costello et al. [35] looked into the planning and scheduling aspect of target acquisition, drawing from a router analogy. Bagdanov et al. [10] uses reinforcement learning to explore the action space of the control problem, and Del Bimbo and Pernici[45] approached this using the kinetic travelling salesman problem. However, mainly fixed rules are used to decide which cameras should follow which targets, including those which favoured targets walking towards observing cameras to maximise the chance of detecting a face, neglecting the uncertainty in the sensing process. Whereas above papers used geometric insight, others suggested cinematographic rules [50] and analogies to mechanical forces [1]. Isler et al. [78] address assignments of targets to cameras to minimise expected error in the estimation of target locations. Choosing the optimal assignments can be phrased as a general assignment problem if multiple observations of the same target are disallowed.

The full case, known as general multiple assignment problem, has been addressed for objective functions that are concave with respect to the number of assignments [159], i.e. for diminishing returns. This is true for mutual information, see section 3.3.6. The approach is a linear program building on consecutive evaluation of the Hungarian algorithm, but the complexity is $O(CM^3)$ for C cameras and M targets. In Operations Research literature, this particular assignment problem is also known as weapon-target-assignment problem (WTA), and even referred to as its "holy grail" [6].

Qureshi and Terzopolous [122] address the target selection for tracking as a planning problem with a centralised planner. The actions in this problem are simplified in that cameras are assumed to autonomously acquire and track targets, and a longer term plan

is found using greedy planning over 5 to 10 time steps into the future. The targets are assigned to cameras according to the expected sequence's quality. This quality is a heuristic which is a product of factors over the sequence length. Each factor at each time step in turn is a product of factors such as proximity of a camera to a target, visibility constraints, and probability of acquiring a target. The parameters for each of these factors are chosen empirically.

The emphasis of the work in this chapter is on the objective function, not the planning stage, as we are interested in the influence of gaze estimation, and the use of the classifier's uncertainty in the sensing process. In fact, the proposed objective function could be used in long term planning approaches as the one proposed by Qureshi.

The approach described in this chapter is unique in using coarse gaze estimates to optimise camera allocation. Previous work on coarse gaze estimation has covered only a few application areas such as attention measurement in surveillance situations [18, 91, 132], providing speaker feedback for presentations [61] and identifying the focus of attention of drivers [113] and meeting participants [128, 148].

7.3 Camera Control

In this chapter, we base the control method on the following observations. Ideally, identification systems return a likelihood of either having successfully identified the target, or a distribution over potential identities of the captured target. In the simplest case, a system returns a truth value whether the target has been identified, and the only probabilistic feedback is the identification performance of the system, which yields the percentage of false positives and false negatives. Furthermore, typical identification systems have a better performance when the target's face is captured frontally. For example, a recent

method [30] first looks for frontal face detections, and then tries to identify the captured frames.

An underlying goal of surveillance systems is to track objects and obtain images at a high resolution for the purpose of later identification. This means the targets are actually only tracked and observed at higher resolution to allow identification afterwards. We phrase this as an information gathering problem: we are interested in minimal uncertainty of the targets' identities.

In this chapter, we consider a setup of static supervisor cameras that suffice to track target locations and do data association. This means that the control approach does not require any minimisation of the target's positional uncertainty; however, we will show how our approach would nonetheless be able to take this into account.

7.3.1 Formal Problem Statement

Suppose a set of C active cameras can change zoom, pan and tilt to provide imagery that allows identification of a *single* individual. For all cameras at a given time, this defines a parameter vector with corresponding pan, tilt and zoom triples. Furthermore, there are N targets under surveillance, each with a class or identity t_k, and a position $\mathbf{x}_k$: $\mathcal{X}_k = \{\mathbf{x}_k^1, \ldots, \mathbf{x}_k^N\}$, $\mathcal{T} = \{t^1, \ldots, t^N\}$. These cameras need to be steered to some of these targets to facilitate their identification.

This evokes the problems of target assignment to each camera, and parameter selection for each camera. The first summarises which camera is to be used to track which target. The second problem addresses the choice of zoom and other parameters for each camera, which are bounded by the uncertainty of the object's motion, as well as its spatial extent.

In keeping with the theme of this thesis, we phrase this problem in a coherent information-theoretic manner. Apart from differing objective and observation functions, the overall

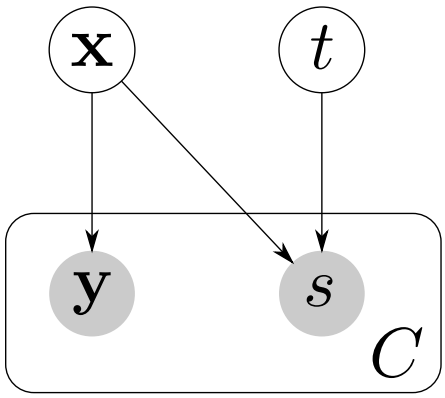


Figure 7.1: Graphical model showing the dependency of the random variables for each target: Position $\mathbf{x}$, observed positions $\mathbf{y}_c$, identity t, and observed identity s_c. The plate C shows that C observations are generated, one for each camera.

procedure remains the same: Before making an observation at time k, we choose the best parameter $\mathbf{a}_k$ for the observation. The parameter $\mathbf{a}_k$ summarises the different settings for the observation process, i.e. assignment of targets to cameras, and the respective pan, tilt and zoom settings. Among all choices, the parameter $\mathbf{a}^*$ will provide observations that maximise information

$$\mathbf{a}^* = \arg\max_{\mathbf{a}_k} I_{\mathbf{a}_k}(\{\mathcal{X}_k, \mathcal{T}\}; \mathcal{O}_k). \tag{7.1}$$

Applying the chosen parameter yields CN observations $\mathcal{O}_k = \{o_{1,1} \dots o_{C,N}\}$, which also counts observations that have not been successful. The observations of each target are finally used to update the distribution $p(\mathbf{x}_k^n, t^n)$ for each observed target n and the procedure repeats.

Using the data association provided by the supervisor camera, we can directly relate the measurements to the targets. This makes the measurements independent, and the mutual information gain turns into a sum over all targets currently being observed:

$$\mathbf{a}^* = \arg\max_{\mathbf{a}_k} \sum_{n=1\dots N} I^n_{\mathbf{a}_k}(\mathbf{x}_k, t; \{o_1 \dots o_C\}). \tag{7.2}$$

Each observation has a class-related and a positional part: $o_k^c = \{s_k^c, \mathbf{y}_k^c\}$.

The graphical model in figure 7.1 shows the dependency between the random variables for each target. For each target, we can thus phrase the joint probability of all random

variables involved as follows.

$$p(\mathcal{S},t,\mathbf{x},\mathcal{Y}) = p(t)p(\mathbf{x})p(\mathcal{Y}|\mathbf{x})p(\mathcal{S}|t,\mathbf{x}) \tag{7.3}$$

Here we summarized the group of positional measurements: $\mathcal{Y} = \{\mathbf{y}_1,\ldots,\mathbf{y}_C\}$, and likewise for the classification results: $\mathcal{S} = \{s_1,\ldots,s_C\}$. This assumes that the position of a target is independent of their identity, and the identification result depends on the position of the target. In the context of an identification system, this also ignores the fact that there should be only one instance of each "enrolled" target class visible.

We now consider a single target at a given time step. For clarity, we omit the temporal index k and use the target index n as a superscript on the information term $I^n(\mathbf{x},t;\mathcal{Y},\mathcal{S}) = I(\mathbf{x}^n,t^n;\mathcal{Y}^n,\mathcal{S}^n)$. For each target n, the information gain can be split into a positional $I^n(\mathbf{x};\mathcal{Y})$, and a class-related part $I^n(t;s,\mathcal{Y})$.

$$\begin{aligned}
I^n_{\mathbf{a}}(\mathbf{x},t;\mathcal{Y},\mathcal{S}) &= H^n(\mathbf{x},t) - H^n_{\mathbf{a}}(\mathbf{x},t|\mathcal{Y},\mathcal{S}) && (7.4)\\
&= H^n(\mathbf{x},t) - H^n_{\mathbf{a}}(\mathbf{x}|t,\mathcal{Y},\mathcal{S}) - H^n_{\mathbf{a}}(t|\mathcal{Y},\mathcal{S}) && (7.5)\\
&= H^n(\mathbf{x}) - H^n(\mathbf{x},\mathcal{Y})\\
&\quad + H^n(t) - (w^n_{\mathbf{a}}H^n(t|\mathcal{S}) + (1-w^n_{\mathbf{a}})H^n(t)) && (7.6)\\
&= I^n(\mathbf{x};\mathcal{Y}) + w^n_{\mathbf{a}}I(t,\mathcal{S}) && (7.7)
\end{aligned}$$

The first term, $I^n(\mathbf{x};\mathcal{Y})$, corresponds to the term in the previous chapter. In our setup we use a static camera that provides sufficient positional accuracy, and we are not interested in its optimisation. The important factor in the class-related part $I^n(t;\mathcal{S},\mathcal{Y})$ is the expected visibility $w^n_{\mathbf{a}}$, which adds the dependency on the expected target location to the information gain from classification.

For a single camera, this factor has the same form as in equation 4.13, repeated here as equation 7.8, and is the integral over the area of the sensor that can actually make the

measurement, e.g. field of view Ω:

$$w^n(\mathbf{a}) = \int_{\Omega(\mathbf{a})} p_{\mathbf{a}}(\mathbf{y}_k)\, d\mathbf{y}_k \tag{7.8}$$

The field of view $\Omega(\mathbf{a})$ of the camera is a function of the chosen camera pan, tilt angles α and its zoom value.

The objective function in equation 7.7 could easily handle optimisation with respect to both the positional accuracy and identity of the targets if the two terms are weighted appropriately, i.e. in the way described in section 3.6. For now, we confine ourselves to a study of the purely classification based element of the mutual information, and will investigate the influence of the class-related part without addressing visibility. Here the constraint is that a camera can only observe a single target at a time.

7.3.2 Mutual Information from Classification

We first derive the class-based objective function $I(t;\mathcal{S})$ from equation 7.7, which describes the information expected from observing a target with identity t. We show that information depends on the performance of the classification method used.

The output of a classification method can be a distance measure from each of the learnt classes to the data presented. This distance measure can also be a distribution over the enrolled identities. Furthermore, targets can be classified as "not enrolled" by the system if the returned features are too far from any of the previously learned classes. This can also be seen as a different class for the targets.

This output is turned into a decision by using the most likely label. An uncertain classification can be labelled as "unknown" if several features have similar likelihoods. This is close to David Lowe's ratio test [92], and is used in the classification system by Apostoloff and Zisserman [8]. Another method, and the one we employ here, relies on the entropy of

the distribution over target assignments. If there are N different target identities $t_1 \dots t_N$ and k measurements $\mathcal{S}_k = s_1, \dots s_k$ for a single target, we obtain a posterior distribution $p(t|\mathcal{S}_k)$ for a target identity t. The assigned class is the one with the maximum likelihood, but is rejected if the entropy of the distribution $p(t)$ is higher than a given threshold θ:

$$t^* = \begin{cases} \arg\max_t p(t|\mathcal{S}_k) & \text{if } H(t|\mathcal{S}_k) < \theta, \\ \text{unknown} & \text{if } H(t|\mathcal{S}_k) \geq \theta \end{cases} \tag{7.9}$$

This results in an extra label, U, as the output for classification systems.

Given these labels and the input classes, we describe an arbitrary classification method using its performance, that is the likelihood of assigning a certain label to a given class, and the prior likelihood for these classes to appear.

In general, we assume that the output of an identification system can disambiguate N targets, has one "not registered" class R, and one extra "unknown" label U: $\mathcal{S} = \{1 \dots N, R, U\}$. The system can be described by a confusion matrix which gives the likelihood for assigning a label s to a target t, $p(s|t)$, where $s \in \mathcal{S}$ and $t \in \{1 \dots N, R\}$:

$$p(s|t) = \begin{array}{c|ccccc} & 1 & 2 & \dots & N & R \\ \hline 1 & p_{11} & p_{12} & \cdots & p_{1N} & p_{1R} \\ 2 & p_{21} & p_{22} & \cdots & p_{2N} & p_{2R} \\ & & \ddots & & & \\ R & p_{R1} & p_{R2} & \cdots & p_{RN} & p_{RR} \\ U & p_{U1} & p_{U2} & \cdots & p_{UN} & p_{UR} \end{array}, \tag{7.10}$$

Furthermore, we are given a prior on actual targets, $p(t)$.

When M cameras are observing the same target, M observations are made, $\mathcal{S} = \{s_1 \dots s_M\}$, and the posterior $p(t|\mathcal{S})$ is obtained via Bayes' rule. These likelihoods yield entropy $H_{\mathbf{a}_k}(t|\mathcal{S})$ and mutual information $I_{\mathbf{a}_k}(t;\mathcal{S})$ as explained in section 3.3.1.

If all cameras can always observe all targets, and no positional information $I(\mathbf{x};\mathbf{y})$ is desired, then the final objective function in equation 7.7 is a sum of all mutual information

gained from each target:

$$\mathbf{a}^* = \arg\max_{\mathbf{a}_k} \sum_{n=1}^{N} I^n_{\mathbf{a}_k}(t;\mathcal{S}). \tag{7.11}$$

In the case of a single target per camera and C cameras, up to C targets can be observed. The maximisation of this function can be seen as a generalized assignment problem (or one-to-many matching) with varying profit function, i.e. depending on the number of cameras assigned to each target. Here we approximate this optimization assuming independence of the observations and simply select the best target per camera.

7.4 Performance Comparison

The objective function in equation 7.11 specifies the control law to always select the most informative target. We are interested to see how this specific control influences the number of identified targets in a system.

We now compare the performance of a surveillance system by using the objective functions in the following list. The functions are variants of the information-theoretic objective in equation 7.11, and standard approaches from literature.

- As an alternative to the mutual information based approach, we use the current entropy of the target's distribution over identities:

$$\mathbf{a}^* = \arg\min_{\mathbf{a}_k} \sum_{n=1}^{N} H^n_{\mathbf{a}_k}(t;\mathcal{S}). \tag{7.12}$$

  This chooses the target with the highest uncertainty and also tries to minimise the uncertainty about all targets' identities. However, a target might have a high uncertainty, but further observation might not be sensible as the expected gain in information is smaller than for other targets, due to a specific confusion matrix. Instead, mutual information encapsulates the expected information gain from sensing a target, taking into account the confusion in the classifier by expectation.

- We exploit the threshold for decision making in equation 7.9, and penalise targets in the selection process that have an entropy below this value. The idea is that, once the sensing processed crossed the entropy threshold in equation 7.9, further information about this target might not be needed at all. For each target, we modify the information as

$$I^{n*}(t;s) = \begin{cases} I^n(t;s) & \text{if } H(t) > \theta \\ I^n(t;s) - k & \text{otherwise} \end{cases} \tag{7.13}$$

  where k is a sufficiently large constant (greater than the maximum mutual information). Here we choose $k = 1000$.

- A standard approach in literature (e.g. [121]) is "first come, first served", which simply chooses the target that appeared earliest.

- An even simpler alternative is one that chooses the target closest to the camera – it minimises a "cost" of movement.

We further incorporate the cost of moving a camera into all objective functions by an addition of a quadratic term which penalises large movements:

$$\mathbf{a}^* = \arg\max_{a_k} \sum_{n=1}^{N} I^n_{\mathbf{a}_k}(t;\mathcal{S}) - \kappa \mathbf{a}_k^T \mathbf{Q} \mathbf{a}_k \tag{7.14}$$

This yields the following different objective functions: mutual information (also referred to as mi), entropy (ent), thresholded mutual information (mi thresh), "first come, first served" (fcfs), and minimum cost (cost).

7.4.1 Evaluation

We now investigate how the objective functions in section 7.3.2 perform for target selection in a simplified surveillance setting. As the aim is to classify as many targets as possible, we evaluate the accuracy of this method. We furthermore evaluate the ratio

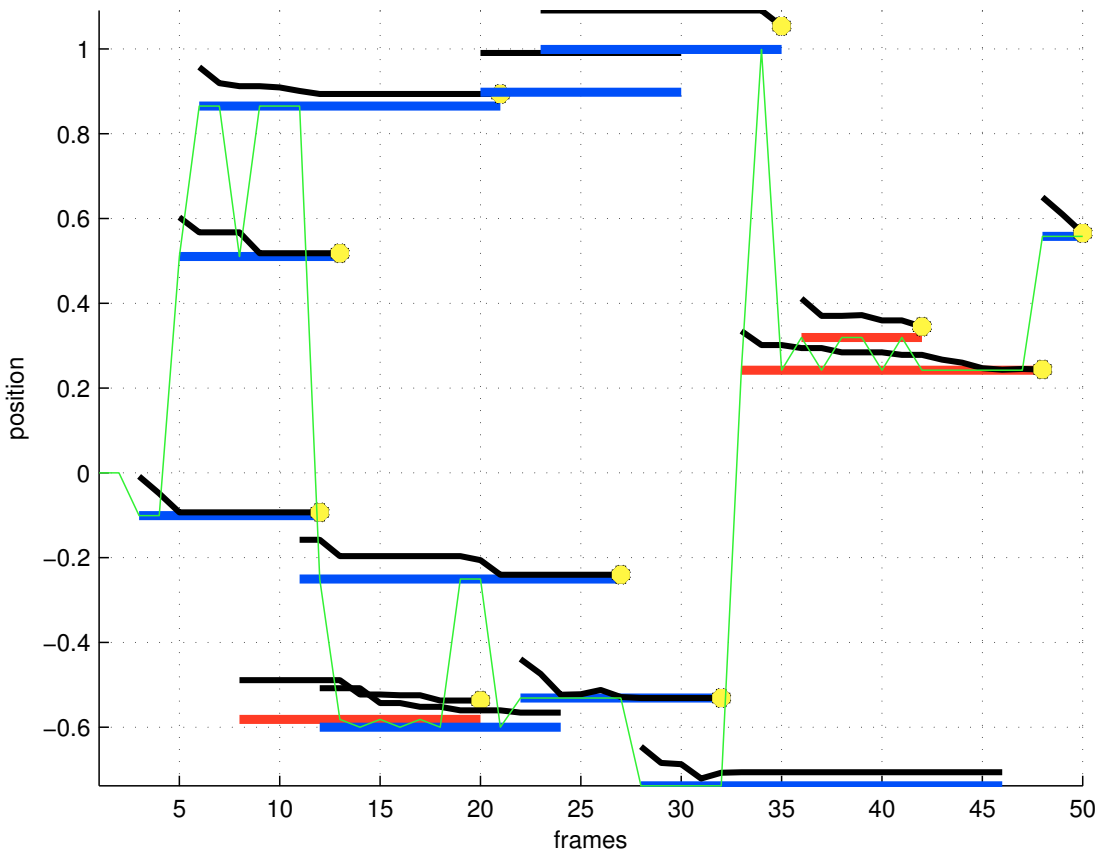


Figure 7.2: Example run of the simulation. The blue/red lines show the existence of a target at the position given on the y-axis. The camera motion is indicated by a green line. The entropy of each target is drawn on top of each target. Note how this reduces with each observation. Blue targets are enrolled in the system ($t \in \{1 \dots N\}$), whereas red ones are not ($t = R$). Yellow dots at the end of the target's depiction of entropy indicate a successful identification.

of average observation time of enrolled and not enrolled targets, as one of the ideal behaviours of the system is a focus on targets not seen before.

To better compare the influence of the target selection methods, we choose a simplified setting which largely ignores visibility constraints and simulate target arrivals and departures. For the purpose of comparison, this simulation ignores gaze directions or motion models of the targets. The targets all look towards a camera that can move orthogonally to the targets' view direction with a velocity v_c. We further assume a static camera that maintains an overview of the targets currently visible and signals arrival and departure to the active camera. The set of targets generated as such is then presented to a virtual

camera. The system has to decide which target to observe for classification, and adjusts the position of the active camera. If the chosen target is in the field of view of the camera, it can be observed and potentially identified. An example from the simulation is shown in figure 7.2. Optimally, the camera followed the path that yields the most identifications.

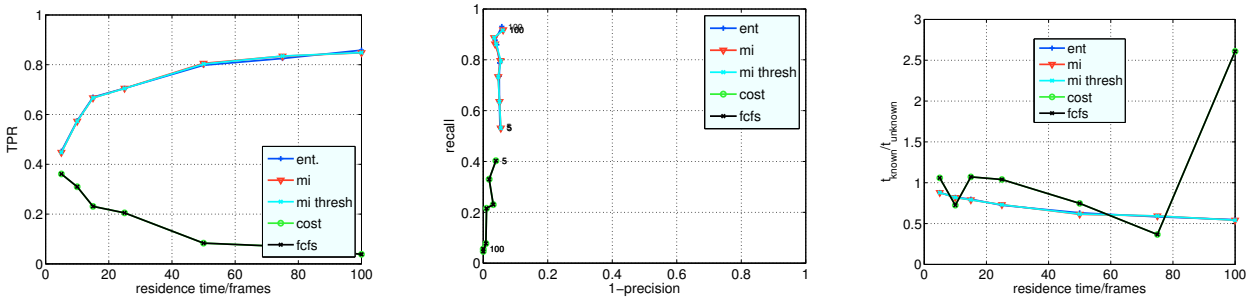
The details of the target generation are as follows. The system is interpreted as a single server (the whole area under observation), and the arrivals and departures are modelled by a birth-and-death process. This is also known as $M/M/1$ model in queuing theory [62]. For a birth-and-death process with parameters λ and μ, arrivals appear according to a Poisson process with rate λ, and their residence time has an exponential distribution with mean $1/\mu$. The target positions are uniformly distributed between -1 and 1 along a line. The identities of the targets are drawn using a prior on their identities, $p_0(t)$. In our experiments, we assumed an equal proportion of targets enrolled in the system to the ones not enrolled.

The classifier used has a performance $p(s|t)$ close to identity for $s \neq U$, and a uniform performance for $s = U$. This means that each target class is equally likely to result in an "unknown" label. Added to the classifier is a uniform random matrix with standard deviation of $\sigma = 0.05$, and the confusion matrix renormalised. In all of our experiments, we used $N = 20$ different target identities.

The following two sections detail the experiments performed. Each experiment has been run 10 times with the same settings, and we report the average results.

Ideal camera

The first experiment varies the residence time T_r between 5 and 100 frames, and sets the velocity of the camera to infinity, that is, it can reach the selected target in one time step. The appearance rate is set to $\lambda = 0.1$, and the residence time to $1/\mu = 10$ frames.



(a) (b) (c)

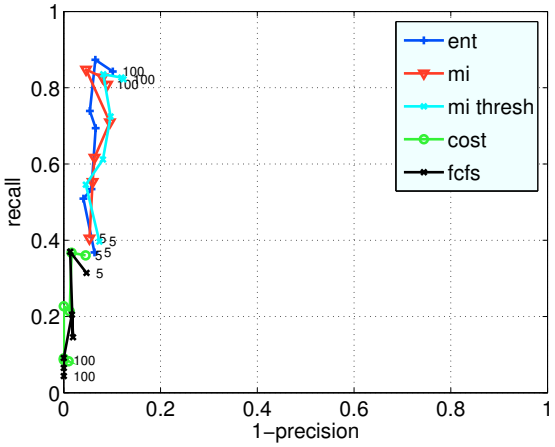
Figure 7.3: Left: percentage of correctly identified targets for a system with a single camera with infinite speed and for differing residence times of targets. Centre: recall/precision plot for the system. Right: ratio of observation times of enrolled and not enrolled ($t = R$) targets (less is better).

Figure 7.3 shows the that using the uncertainty in the target selection clearly outperforms the simpler approaches. All information-theoretic approaches perform equally well.

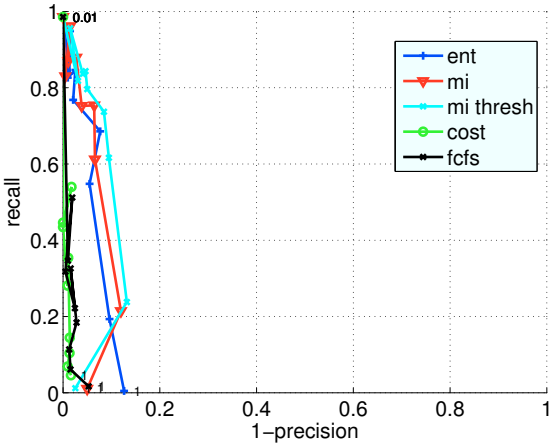
Realistic camera

Figure 7.4(a) shows that the influence of uncertainty in the classifier diminishes once the camera movement requires time to observe targets. In figure 7.4(b) the appearance rate of the targets is varied. The comparison shows that information-theoretic approaches are able to identify more targets at higher rates. The precision/recall curve seems to suggest that this performance gain comes from identifying fewer targets than enrolled. This is however natural as more targets appear and not all can be addressed by a single camera. The standard methods instead focus on the oldest or spatially closest target for a longer time, but correctly classify these ones.

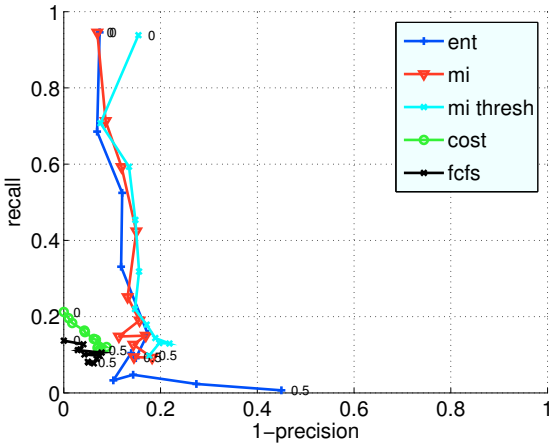
We furthermore evaluated the performance of the control methods by varying the parameters of noise in the classifier. Here, the noise matrix added to the classifier performance has an increasing variance from 0.05 to 0.5. The results of this experiment are



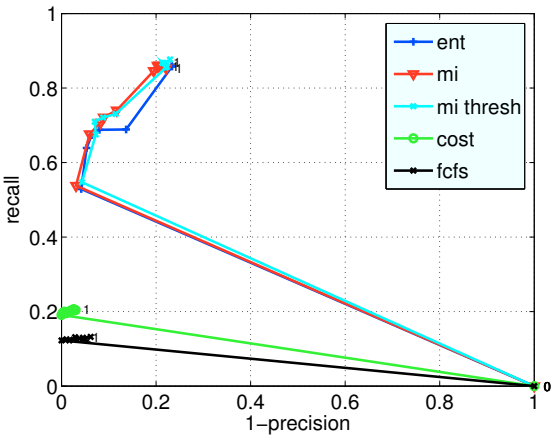
(a) Residence time



(b) Appearance rate



(c) Noise in classifier



(d) Classification threshold

Figure 7.4: Recall/precision plot for varying parameters.

shown in figure 7.4(c). As the plot shows, the information-theoretic approaches provide better results over a wider range of values than the standard methods.

Lastly we evaluated the influence of the classification threshold in equation 7.9, which is also the threshold in the modified mutual information, see equation 7.13. The threshold was varied from 0 to 1.0, i.e. on one end the targets were immediately identified after the first sensing, at the other end targets were only identified after they left. As shown in figure 7.4(d), the performance of the information-theoretic methods is better than the standard methods. In this comparison, the difference between the methods based on

mutual information and entropy becomes apparent. The methods based on mutual information are able to attain a better performance as they take into account the performance of the classifier. If a particular target is likely to be poorly classified, then it is less often observed – instead, the system chooses targets that can be classified. The performance of thresholded mutual information (mi thresh) is not much different than the variant without thresholding (mi). The additional sensing of some targets in the latter approach makes the state estimate more certain, at the expense of fewer targets classified overall.

7.4.2 Discussion

The evaluation has shown that for all but exact classifiers and very short residence times, target selection using an information-theoretic objective function yields better results than standard approaches. The difference in performance depends on the appearance rate, the actual time targets reside in the supervised area, the camera velocity and the classifier performance. All these properties can be learned offline (from the scene the cameras are placed in), and a decision can be made which scheduling method is beneficial.

7.5 Simplified approach

In the next sections we simplify the performance of the identification system to typical characteristics given for biometric systems [97] by abstracting the identification process in terms of its false positive and negative rates.

Performance for identification systems is usually given in related likelihoods – false acceptance rate (FAR) of potentially not registered targets wrongly identified as known, and the false rejection rate (FRR) of registered targets which are wrongly considered not registered. It is related to the true positive rate (TPR) by $\mathrm{TPR} = 1 - \mathrm{FRR}$. The performance

of not being able to classify the system is called the "failure to capture" rate (FCR). From all presented input data, this percentage is classified as "unknown".

Typically, biometric systems are designed to give a very low failure to capture rate, and rely on cooperation from the user for identification. In our system the targets can be moving, and might not be aware of the identification, which yields a much higher rate.

In our approach, we maintain a (latent) belief state, e representing whether the target is an enrolled subject or not enrolled, and we model the identification process via the conditional likelihood of successfully identifying a target, $p(d|e)$, which is specified in terms of the true positive and negative rates of the system; here the indicator variable d is the binary event of a successful identification measurement from the system. An advantage of this approach is that this formulation does not require any enumeration of target identities, only an indication of the confidence with which the system has (or has not) identified the target.

For the purpose of controlling cameras, we reduce the system to the set of measurements $\mathcal{D} = \{1, 0, U\}$, that discards the actual observation of the target's identity s, but indicates whether the target is registered ($d = 1$), not registered ($d = 0$), or unknown ($d = U$). This likelihood varies according to an enrolled target ($e = 1$) or not enrolled target ($e = 0$).

These measurements specify the labelling of the target with respect to one of the labels in $\mathcal{L} = \{1 \ldots N\}$ and R. The accuracy of this measurement is given by the joint distribution, $p(s,t)$:

$$p(d = 1|e = 1) = \sum_{i=1..N} p(s = i, t = i) / p(t \in \mathcal{L}) = (1 - FCR)TPR, \tag{7.15}$$

which is the sum of diagonal elements of the joint distribution. We furthermore obtain

$$p(d = 1|e = 0) = \sum_{i=1..N} p(s = i, t \neq i) / p(t = R) = (1 - FCR)FAR, \tag{7.16}$$

which is the sum of the off-diagonal entries except the last entry, and

$$p(d=0|e=1) = \sum_{i=1..N} p(s=R,t=i)/p(t\in\mathcal{L}) = (1-FCR)FRR. \tag{7.17}$$

The likelihood of correctly identifying the target as not registered is

$$p(d=0|e=0) = p(s=i,t=R)/p(t=R). \tag{7.18}$$

Note that the misidentifications among the registered targets are also added to the false acceptance rate, as each target can have different authorisations. So even in case there is a closed system prior where $p(t=R)=0$, $p(d=0)$ is not zero if there is the chance the system misidentifies the target among the known N.

In our models, we assume $p(d=U|e)=FCR$ as uniform, but it relates to the confusion matrix as

$$p(d=U|e=1) = \sum_{i=1..N} p(s=U,t=i)/p(t\in\mathcal{L}). \tag{7.19}$$

and

$$p(d=U|e=0) = p(s=U,t=R)/p(R). \tag{7.20}$$

We give a full example for this relation in appendix D.

7.6 Incorporating Gaze

We now return to the original form of the mutual information gain in equation 7.7. To facilitate recognition, it is necessary to prioritise assignments of cameras to targets that look into the camera. For example, a recent identification method [30] first looks for frontal face detections, and then tries to identify the regions labelled as such. The classification process will not be started if no face has been detected. Typically[85, 122], this prioritisation is made by using the current target's velocity as an estimate for the target's gaze

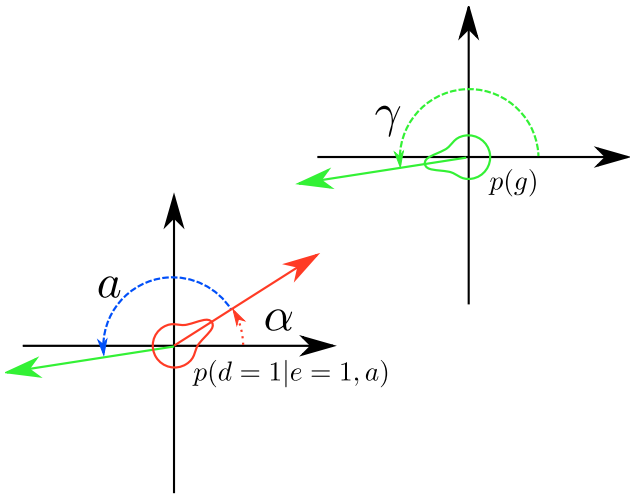


Figure 7.5: Top-down view of camera in direction α (red) facing a target with mean gaze γ, yielding an angle of incidence a.

direction. Recently, however, there has been made considerable progress in estimating head poses from images of limited resolution [5, 18, 110, 63, 163].

In this section we show how to integrate a gaze estimate and the target's positional uncertainty in the second mutual information term in equation 7.7. The gaze estimate can come from the target's velocity or from dedicated sensing. We finally evaluate the increase in performance when using a dedicated gaze estimate over the typical approach of using the target's velocity.

We model the identification process via an augmented conditional likelihood of successfully identifying a target, $p(d|e,a)$. Figure 7.5 shows a situation of a camera facing a single target. We make the assumption that the identification quality depends on the angle of incidence a of the target's gaze and the camera's orientation. The peak performance is specified in terms of the true positive and negative rates of the system; As in section 7.5, the indicator variable d is the binary event of a successful identification measurement from the system, e is the binary flag whether the target is enrolled or not, and g is the target's gaze.

More specifically, in our notation, the system's true positive rate with respect to the angle of incidence of the subject's gaze is denoted as $p(d = 1|e = 1, a)$. The event $e = 1$ describes whether a target is in the set of known targets. Consequently, the event $e = 0$ occurs if a target is not registered with the system. The distributions over e, $p(e), p(e|d)$ are then interpreted as the prior and posterior belief whether the target is known to the system or not. The resulting objective function follows unidentified targets – as long as they have not been classified, as intuition demands.

For any instant, our model requires a distribution $p(g)$ over the direction of the target's gaze. We obtain the maximum likelihood gaze direction from a gaze detection method [18] (also briefly summarised in section 7.7.2), and model the uncertainty over this using a von-Mises distribution:

$$p(g) = f_p(g; \gamma, \kappa_g) = e^{\kappa_g cos(g-\gamma)} / Z_g \tag{7.21}$$

which has a mean direction γ and angular covariance κ_g and a normalisation constant Z_g. With an orientation of the camera of α, the angle of incidence a is given as $a = g - \alpha$. Note that we assume perfect knowledge of the cameras' intrinsic and extrinsic parameters (which is realistic given accurate internal calibration and pan-tilt encoder feedback), otherwise the uncertainty in the angle α had to be included in the following derivation.

If a target is known ($e = 1$), we assume that the detection performance for all viewing angles is unimodal and can also be modelled by a similar function as in equation 7.21:

$$f(a) = \exp(\kappa_a \cos(a + \pi))) / Z_a \tag{7.22}$$

The spread κ_a is determined from the empirical performance of Benfold's algorithm [18]. The likelihood of successfully making a correct classification ($d = e$), given performance

$p_0(d|e)$ (for a frontal view) and incidence angle a is

$$p(d|e,a) = (p_0(d|e) - \epsilon)f(a) + \epsilon. \tag{7.23}$$

The likelihood of an incorrect classification is then $p(d = i|e = i,a) = 1 - p(d = i|e = i,a)\, i \in 0,1$. An example for the resulting identification performance can be seen in figure 7.6 for $\epsilon = {}^1/_2$.

We obtain the resulting expected detector performance – depending on the camera's angle α – by marginalising out the angle of incidence $a = g - \alpha$:

$$p_\alpha(d|e) \quad = \quad \int_{0..2\pi} p(d|e,a)p(a)da \tag{7.24}$$

This integral has no closed form solution.

Even though we are augmenting the state of the target with a gaze estimate g, we do not influence this estimate using the sensing on the active camera. The mutual information in equation 7.7 thus remains the same, but instead of the likelihood $p(d|e)$, the expectation $p_\alpha(d|e)$ over all angles is used.

This development does not take into account the limited field of view of each camera, which influences the expected visibility of each target. To address visibility, we consider the position $\mathbf{y}$ of the target, which influences the chance of actually making an observation o. In our formulation, this target position does not influence the actual identification capability $p(d|e)$, but only the chance of observing the target, i.e. how likely it is to actually capture a bounding box containing the face of the target for a given camera angle.

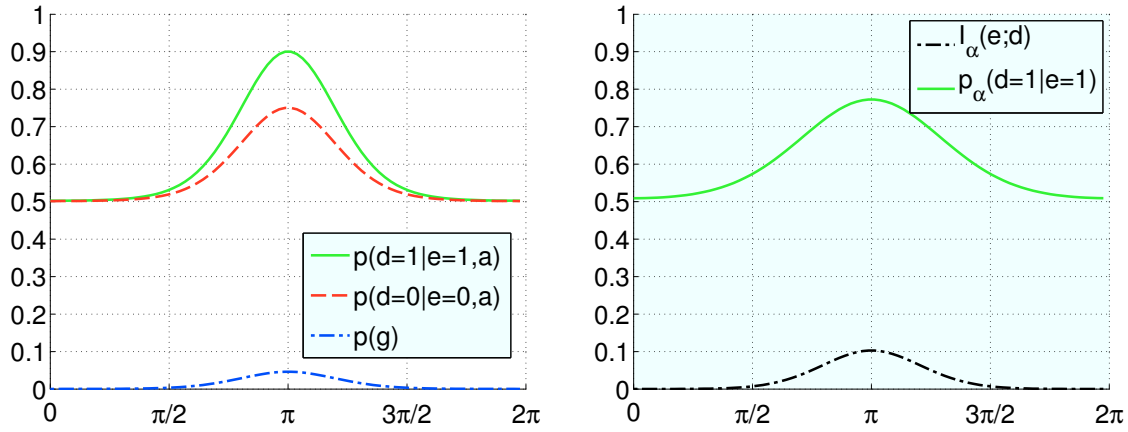


Figure 7.6: Left: Sample model of detector performance over angle of incidence a. Green: true positive identification likelihood, $\kappa_1 = 8/\pi$, true positive rate: 0.90. Red/dashed: true negative identification likelihood, true negative rate: 0.75. Blue/dash-dotted: sample distribution over gaze direction g. Right: Detector performance and mutual information for targets placed in a circle around the camera, each with the same distributions over gaze directions as in the left figure. Green: true positive identification likelihood. Black/dash-dotted: resulting information gain.

The conditional entropy information term for a target in equation 7.7, observed by a single camera, becomes

$$\begin{aligned} H_{\mathbf{a}}(e|\mathbf{y},d) &= -\int_{\mathbf{y}\in\Omega(\mathbf{a})} p(\mathbf{y}) \sum_{d,e} p_\alpha(e|d) \log p_\alpha(e|d)\, d\mathbf{y} \\ &\quad -\int_{\mathbf{y}\notin\Omega(\mathbf{a})} p(\mathbf{y}) H(e)\, d\mathbf{y} \qquad (7.25) \\ &= w_{\mathbf{a}} H_\alpha(e|d) + (1 - w_{\mathbf{a}}) H(e) \qquad (7.26) \end{aligned}$$

From this results the mutual information for a target:

$$\begin{aligned} I_{\mathbf{a}}(e;d,y) &= H(e) - H_{\mathbf{a}}(e|\mathbf{y},d) \qquad (7.27) \\ &= w_{\mathbf{a}} I_\alpha(e;d) \qquad (7.28) \end{aligned}$$

Both expected detection likelihood in equation 7.24 and the mutual information are shown in figure 7.6 for a set of targets surrounding the camera. The constant ϵ in equation 7.23 can be understood as the minimum confusion of the classifier. It addresses the fact that the system should be the least certain, or "blind", when the target is facing away. If the

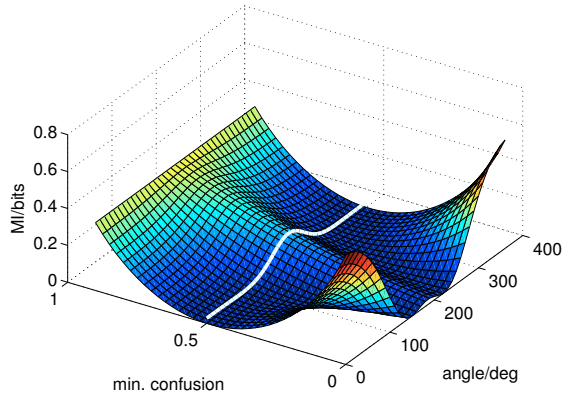
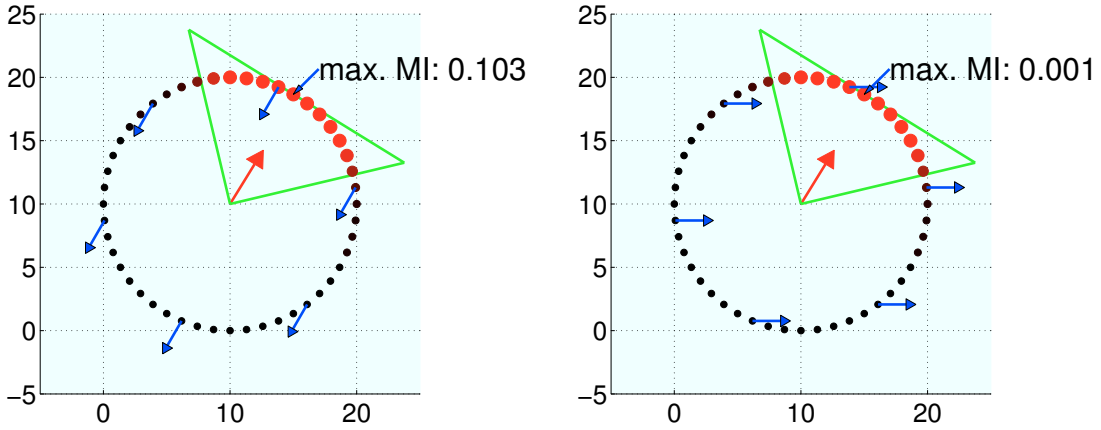


Figure 7.7: Surface of mutual information for varying angles and minimum performance setting ϵ. The mutual information for a minimum confusion of $\epsilon = {}^1/_2$ is highlighted in white.

likelihood of identifying the target as belonging to the system was zero at this angle, the classifier would be confident about the target class, and provide information. The influence of this threshold on the mutual information is shown in figure 7.7 for a performance of the classifier of $p(d = e|e) = 0.9$. When the target faces away ($a = 0$), The identifier yields a mutual information of 0 for $\epsilon = {}^1/_2$.

Finally, we address the visibility factor $w_{\mathbf{a}}$ in equation 7.28. We model the position of the target using a Gaussian, as would result from tracking using a Kalman filter, and linearise the projective transform so that the observation likelihood is a Gaussian, too. The effect of incorporating this term is – as one would expect – a drop of the mutual information towards the boundary of the camera's field of view. This is exactly the same process as in chapters 4 and 5. The visibility term $w_{\mathbf{a}}$ thus has the same form of an integral over the field of view of the camera as in equation 4.13. An example is given in figure 7.8. The cut-off at the edge of the field of view is more pronounced with a smaller covariance of the target position. After each report of a successful or unsuccessful identification, we update the hidden target state $p(e|d)$ using Bayes' rule and propagate this to the next observation period as the prior $p(e)$ (i.e. Markovian assumption). This results in a diminishing return for longer observation of the same target once it has been identified. Again, if the employed identification system had a perfect identification performance, the result-



(a) (b)

Figure 7.8: Different information gains for target positions and views. A camera (centre) surrounded by hypothetical targets, with orientations 7.8(a) towards the camera, and 7.8(b) away from the camera. The resulting information gain is proportional to the radius of the circle on each target. The targets facing away from the camera still yield an information gain due to the uncertainty in the gaze estimate. Note that the actual target position and its uncertainty is taken into account, i.e. even a target outside of the field of view of the camera yields an information gain, because our uncertainty in its position means there is a chance that it is in fact in the field of view and will be observed.

ing probability $p(e = 1|d)$ would be either 1 or 0. Such target state yields an information gain of zero, and would not be considered by the control method any more.

7.7 Results

To compare different methods in a fair manner, they should be exposed to the same input data. This is intrinsically difficult with live recordings of human actors, as no staged actor would ever move in the same manner, and repeated recordings to obtain an average control is tedious, as the same kind of movements need to be performed. A "virtual zoom" approach, where high resolution video data is subsampled and cropped according to the chosen parameters, is also not possible. Here the camera orientation has to remain constant, but the method presented here relies on exactly this degree of freedom.

We choose to use simulation to obtain quantitative and comparable results for different algorithms, and implemented one of the methods in a live system to show the feasibility and qualitative performance of the control method. The recorded images are currently not fed into an actual identification system, which removes any bias from this end. Instead, we posit that frontal images provide better identification performance, and a control method that captures more frontal views should benefit any identification system.

7.7.1 Quantitative Evaluation

To obtain quantitative results, we use simulation based on ground-truth data. The data set we use comprises manually tracked and labelled gaze directions from a typical high street, where pedestrians are expected to visually explore the shopping windows. It consists of 75 individual targets in 1500 frames (50s).

A virtual camera is placed at the periphery of all trajectories, such that all targets are visible from this position. We run the simulation for 18 different placements of the virtual camera and report the average performance for the methods. The setup is shown in figure 7.9.

We first show an example of the development of the uncertainty of the identification. Out of the data set, we select three targets which are on different trajectories; thus they yield different information gains for longer observations. The targets have very different gaze directions: one target moves away from the camera, the other one towards, and one starts looking towards the camera, but then turns away. These targets are observed from a constant camera position and view angle, and observations generated according to the performance of a given detector model (using the same parameters as for the example in figure 7.6). This setup is shown in figures 7.9 to 7.11.

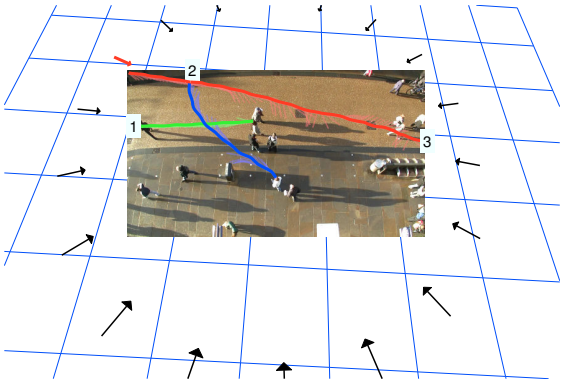
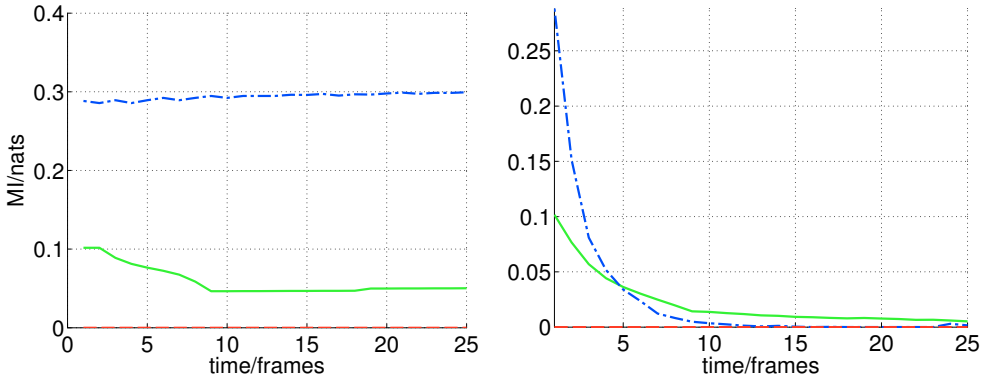


Figure 7.9: Example view of the data set and camera positions used for evaluation, including three trajectories used for visualisation in the next plots (best seen in colour), and a camera located in the top left corner. The figure also shows the position of other cameras used in future experiments. The camera observes targets with varying orientations: towards the camera (blue, 2), away from the camera (red, 3), and slowly turning away from the camera (green, 1).

For each target, we randomly sample d from $p_\alpha(d|e=1)$ for the length of the trajectory of the target. We repeat this observation process for 100 trials and report average performance. The graph 7.10(a) shows how only the target with a frontal view towards the camera is selected if the belief state is not updated. However, graph 7.10(b) shows how the second target provides more information after a number of observations, as the first target has been observed long enough. This amount of observations depends on the system's performance (e.g. for a perfect identification a single observation suffices), and the actual result of the observations. The curves 7.11(a) and 7.11(b) show the development of the successful identification of the targets, $p(e=1)$. Note how for target 2 this converges rapidly towards 1, whereas target 1 requires more observations. The state of the third target remains uncertain, as it is never facing the camera.

For each frame and each target visible, we evaluate the expected information gain in equation 7.28 for a camera setting that centres this target in its own field of view. The camera is then directed to fixate the target which maximises the (expected) information gain.

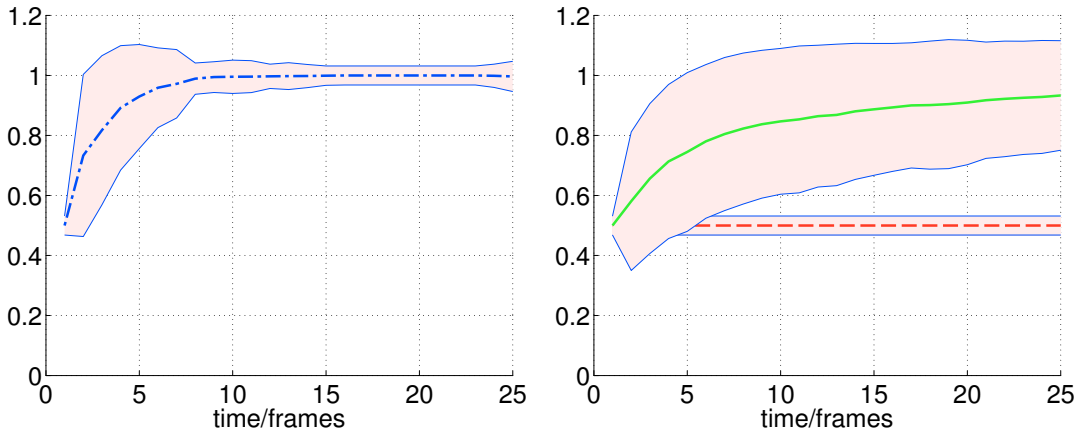


(a) $p(e)$ constant (b) $p(e)$ updated

Figure 7.10: The left graph shows the mutual information from each target for a constant, indiscriminate belief $p(e)$. Target 1 is green, target 2 blue/dash-dotted, and target 3 red/dashed. Plot 7.10(b) visualises the mutual information gain when constantly updating the belief according to simulated detections. Target 1 yields more information gain than target 3 from frame 5 onwards.

In order to evaluate the efficacy of our approach using information gain together with gaze, we compare against a variety of other possibilities: (i) the full system uses gaze estimates, together with an update of the latent belief state using Bayes' rule at every frame (denoted mi+gaze+prior); (ii) without a recursive update of the latent belief state, but still using gaze estimates (mi+gaze); (iii) recursive update of the belief state but using target motion direction as an approximation of gaze direction (mi+prior); (iv) no recursive update of the belief state, and using target motion direction as an approximation of gaze direction only; and (v) as a baseline method, we also choose a target at random.

As a metric of the performance we count how often the chosen target looks into the camera in the next frame. We determine this by thresholding the difference between the target's actual gaze and the camera's view direction. We also count the number of targets that have been looking into the camera at least once. This would be the metric for a perfect identification system, as each target could be identified from this single observation.



(a) Target 2 (blue)

(b) Target 1 and 3 (green, red/dashed)

Figure 7.11: Development of the belief state $p(e = 1)$ for each of the targets.

We performed two experiments. The first experiment measures how much the control methods influence the number of frontal faces captured. For this, we vary the threshold for the observation angle and keep the other parameters fixed ($\kappa_2 = 40/\pi$, detector parameters as in figure 7.6, delay 1 frame). Figure 7.12 shows the results. Whereas the random selection method observes roughly as many unique targets as the proposed method, the overall number of frontal observations is far smaller. The performance of the random method is only better if all of the captured images are sufficient to identify the targets – i.e. for a perfect identification system. For such a system, the mutual information gain would be 0 after the first observation, and our proposed method would trigger a faster change of targets, gathering more unique observations.

Note that updating the identification belief state recursively over time increases the number of uniquely observed targets. This is because the mutual information gain decreases with the number of observations of the same target, thus favouring a change of target even though the current target might still face the camera frontally.

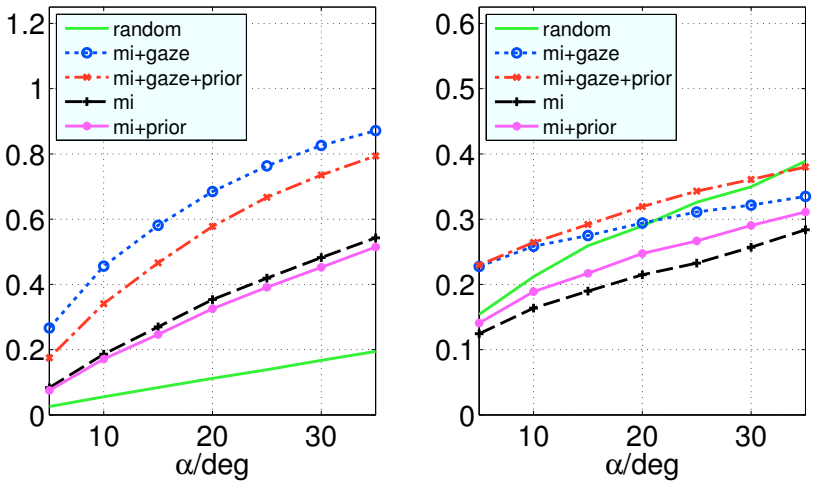


Figure 7.12: Performance comparison for camera control using gaze estimates. The detection performance is imperfect, and modelled as in figure 7.6. Left: fraction of frames where a target has been observed at an angle less than α. Right: fraction of targets observed at least once.

The second experiment shows the dependency on the temporal delay in the control. After a target has been selected by one of the objective functions, we assume that the positioning of the camera takes a number of frames, fixed for every chosen movement, and during this time no target can be sensed. We vary the delay between one and fifteen frames, which corresponds to $1/30s$ to $1/2s$. Figure 7.13 shows the average number of persons observed over all viewing angles. The standard deviation for the mutual information based approaches is about $\sigma = 0.16$, and for the random selection $\sigma = 0.05$, which shows that there is a statistically significant performance improvement when using the gaze based performance, even for a delay of up to half a second in the control method. The shaded area depicts the standard deviation of the random selection and the method of mutual information gain from gaze estimates plus updated prior (red/dash-dotted).

The right graph in figure 7.13 shows the relative number of unique observations for each of the methods. As the random selection method does not prioritise targets, it is more likely that targets are selected that do not face the camera or disappear before they can be observed. However, all mutual information based methods perform worse for a longer delay, as it is more likely that the target turns away.

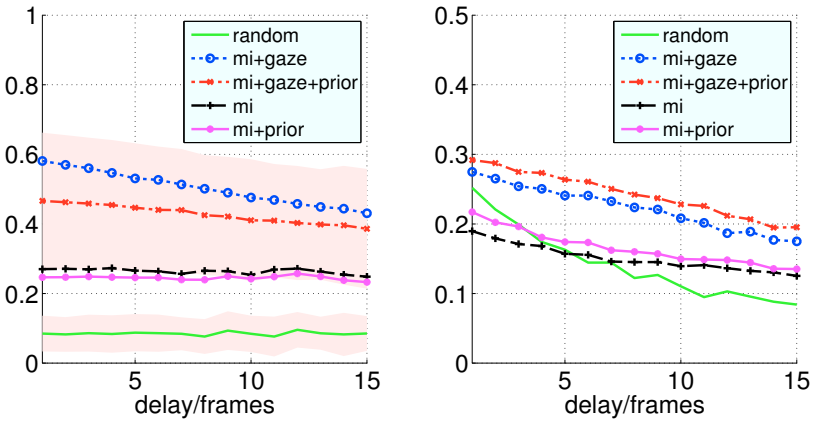


Figure 7.13: Left: fraction of frames where a target has been observed at an angle less than 15°, for varying delay between target selection and evaluation. Right: fraction of targets that have been observed at least once at an angle less than 15°.

7.7.2 Live System

To test the feasibility of our approach, we implemented a live system that uses a static camera to obtain the gaze directions. The tracker employed in this static camera uses the approach of Benfold and Reid [18] who tracked the heads of pedestrians using a combination of sparse optical flow measurements from KLT feature tracking [93] and head detections using Dalal and Trigg's HOG detection algorithm [38] trained on cropped head images.

In each frame of video, the sparse optical flow from the previous frame predicts the head location and the head detections provide absolute observations which are combined with the predictions using a Kalman filter. The resulting location estimates provide stable

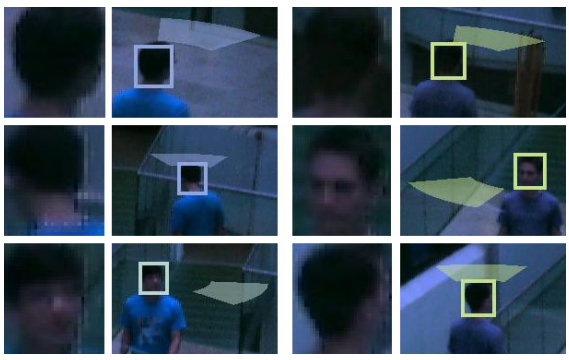


Figure 7.14: Sample head images with the corresponding gaze estimations

head images which are used for gaze estimation (see figure 7.14). The 2D image locations are converted into a 3D location estimates by assuming a mean human height of 1.7 metres using the camera calibration with a ground plane assumption.

Coarse gaze estimates are made using randomised ferns, a simplified form of randomised trees. The ferns use binary tests based on gradient directions and colour comparisons to index a leaf node containing a probability distribution over eight direction classes, of 45^o each, thus covering the full range of 360^o. To avoid having discrete direction estimates, a Gaussian Parzen window density estimate is used to interpolate the most likely angle. This has a width of $360^\circ/8 = 45^\circ$, corresponding to one class.

A supervisor camera runs the gaze tracker, and two active cameras are controlled according to the information gained without update of priors. The parameters of the active cameras are pan, tilt and zoom; we exploit the zoom capabilities by allowing the camera to look at several targets at the same time. Instead of selecting a single target with maximum information gain as in the previous section, we vary the parameters for each camera, and sum the information gain from each target observed. A full search of a discretised parameter space runs at two frames per second per camera. Sample frames from the live system are shown in figure 7.15. The cameras follow the targets, keeping those targets centred that look into the camera.

7.8 Conclusion

This chapter presented an extension of mutual information based tracking to classification systems, where the positional accuracy is of less importance than the correct identification of targets.

In particular, contributions of this chapter are

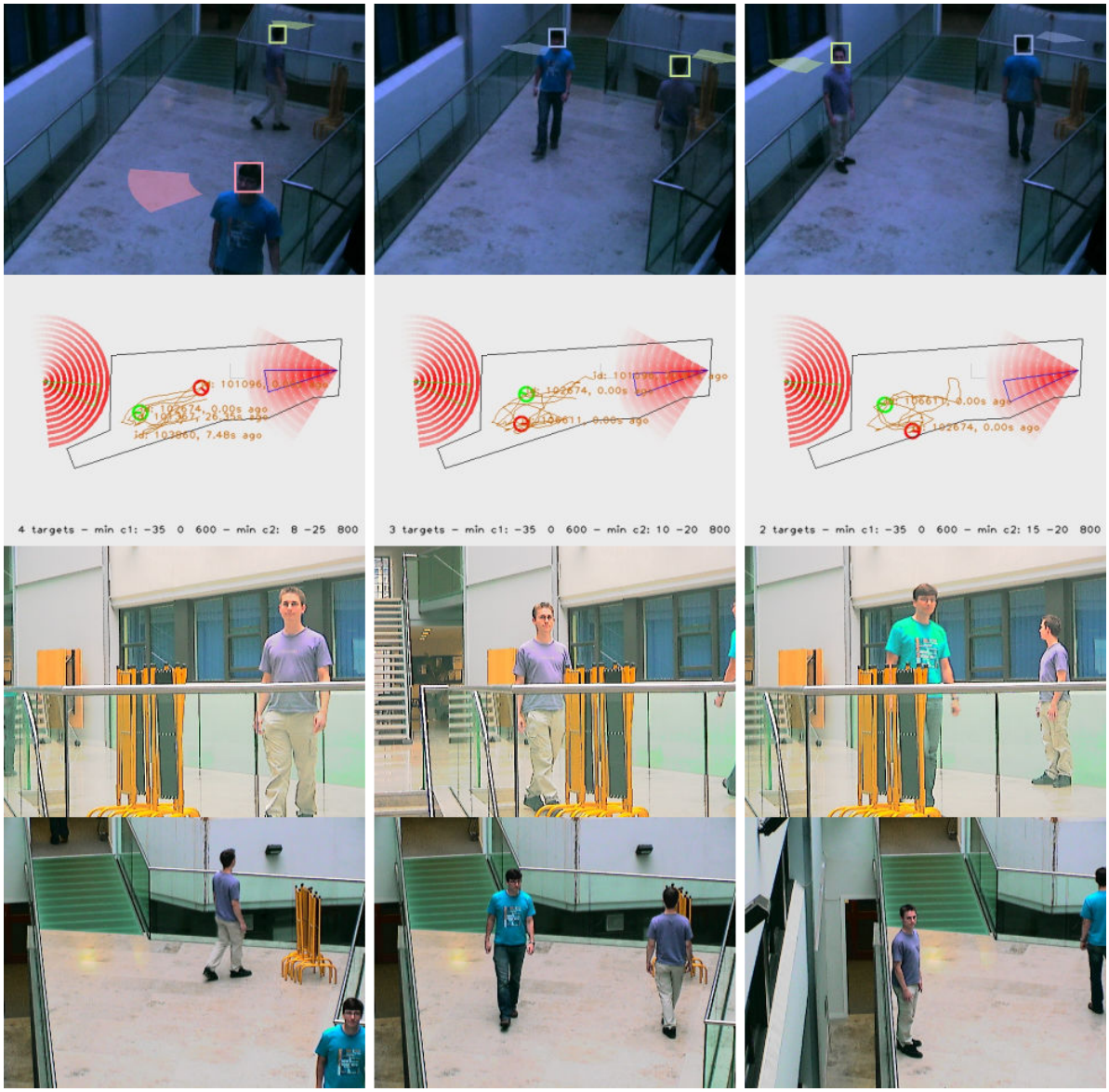


Figure 7.15: Sample images showing the operation of the live system at three different times (left to right). The top row shows images from the static camera. The tracked heads are annotated with boxes and their gaze direction represented by a circular section. The second row shows a schematic of the camera control method. The active cameras are depicted as cones. The objective function for pan, tilt and zoom settings for each camera is marked in red. Active targets are circled, and the target with the most information gain for the left camera is green. The trajectories of all targets from the last 30 seconds are drawn. The third row shows the images recorded by the left active camera, and the last row contains images from the right camera.

- a derivation of an information-theoretic objective function that addresses the uncertainty in the classifier
- experimental validation of the performance gain due to target selection using this objective function
- extension of the objective function to incorporate gaze or head pose
- quantitative evaluation of the use of gaze estimates when observing targets for the purpose of identification: Evaluation based on real data shows that the inclusion of gaze estimates results in more targets captured frontally, which will benefit identification systems.
- a real-time implementation of previous method.

7.8.1 Future work

In the current implementation of the live system, there is no incentive in the objective function to increase the zoom of the camera. Not only does this reduce the likelihood of making an observation (the field of view in equation 7.8 is smaller), but as mutual information is never negative, the inclusion of any target in the field of view will result in a higher mutual information gain, unless other targets become less likely to be observed successfully. This, of course, neglects the intuition that higher resolution faces should yield greater identification success. Within the present framework this could be modelled naturally by introducing a dependence of the identification process success rate on the zoom value. As the zoom increases, the false positive and negative rates would drop, resulting greater potential information gain, thereby providing pressure for a camera to zoom in.

Chapter 8

Conclusion

This chapter provides a summary of the thesis, and highlights the contributions made. Lastly, we will point out possible directions for future work.

8.1 Summary

In this thesis, we have presented a method to control a visual surveillance system in a way that addresses different, potentially conflicting objectives. The objectives typically use standard methods in computer vision, such as object detectors, Kalman filters for tracking, or classification methods.

As described in chapter 3, the approach taken in this thesis is to provide an information theoretic utility function for each separate objective. This puts the output of each of these methods on the same scale, i.e. each objective is fulfilled better if the underlying control provides more information.

These utility functions are then linearly combined into a single objective function, and a preference specified by adjusting the blending parameters between the objectives. Utility theory provides the insight that the control will correspond to the specified preference.

Particular to this thesis, we addressed the three objectives of tracking, detection and classification of targets. For each of these objectives, we formulated new information theoretic objective functions.

In chapter 4, we investigated methods for adjusting zoom on a camera while tracking a single target. We analysed an existing information theoretic objective function, which chooses the zoom setting that best reduces the expected uncertainty. As in this method the uncertainty is directly derived from the underlying Kalman filter, we are able to show that the method is not robust with respect to tracking errors. These can stem from an invalid assumption on the target's motion, or unmatched noise in the observation process. We then extend the method by keeping an on-line estimate of the innovation and add this to the originally modelled observation noise covariance. This increases the usable zoom range while still maintaining track.

Chapter 5 addressed the use of a single, active camera for two competing objectives in the surveillance process. We first extended the objective function from chapter 4 to several targets, and introduced a competing term that measures the information about the appearance of a single target in the area that is currently not under surveillance. We showed that this extension yields sensible behaviour, as the performance for both tracking and detection is higher than other, standard methods can provide.

In chapter 6, we extended the concepts of the previous two chapters to multiple cameras. For this, we derived an information theoretic observation model for the detection process that takes into account the performance of detectors for varying observation parameters. Whereas we concentrated on the zoom parameter, others are straightforward to include. We showed how scene knowledge is easily included in the visibility term of each target, and gave an example from a live system that confirmed our approach.

Lastly, we have developed an information theoretic objective function for the choice of targets in a classification system in chapter 7, which uniquely addresses the classification

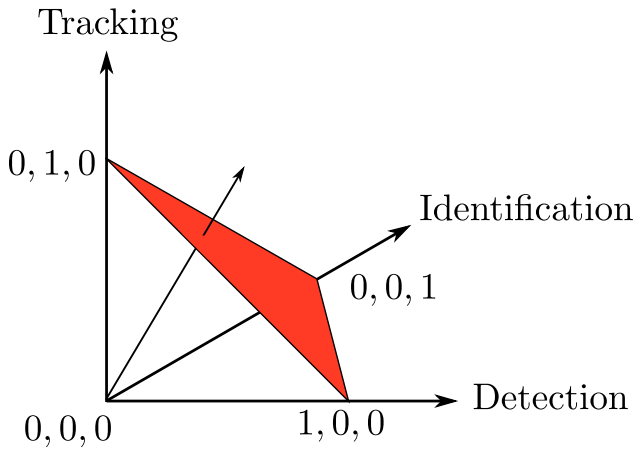


Figure 8.1: A preference for a certain surveillance objective can be specified by a point within the simplex spanned by the information theoretic objective functions.

performance of any method used. We extend this objective function to take into account the gaze or head pose.

We evaluated the proposed methods in two ways. A real-time implementation was used to show the qualitative aspects of the resulting control, and simulation on data sets provided a repeatable experimental setup and could thus be used for a quantitative evaluation.

8.1.1 Conclusion

The objective of this thesis was to provide a facility to control a visual surveillance system in a way that addresses different, potentially conflicting objectives, such as detection and tracking of targets.

In chapter 6, we showed how the objective functions for tracking and detection are combined, and how a single, scalar parameter ζ influences the performance of the controlled surveillance system. This parameter specified the preference for a given objective. With all three objectives of tracking, detection and identification in information theoretic form, the preference for one of these objectives can now be set by barycentric coordinates that uniquely define a position in the space defined by the objectives, see figure 8.1. The ap-

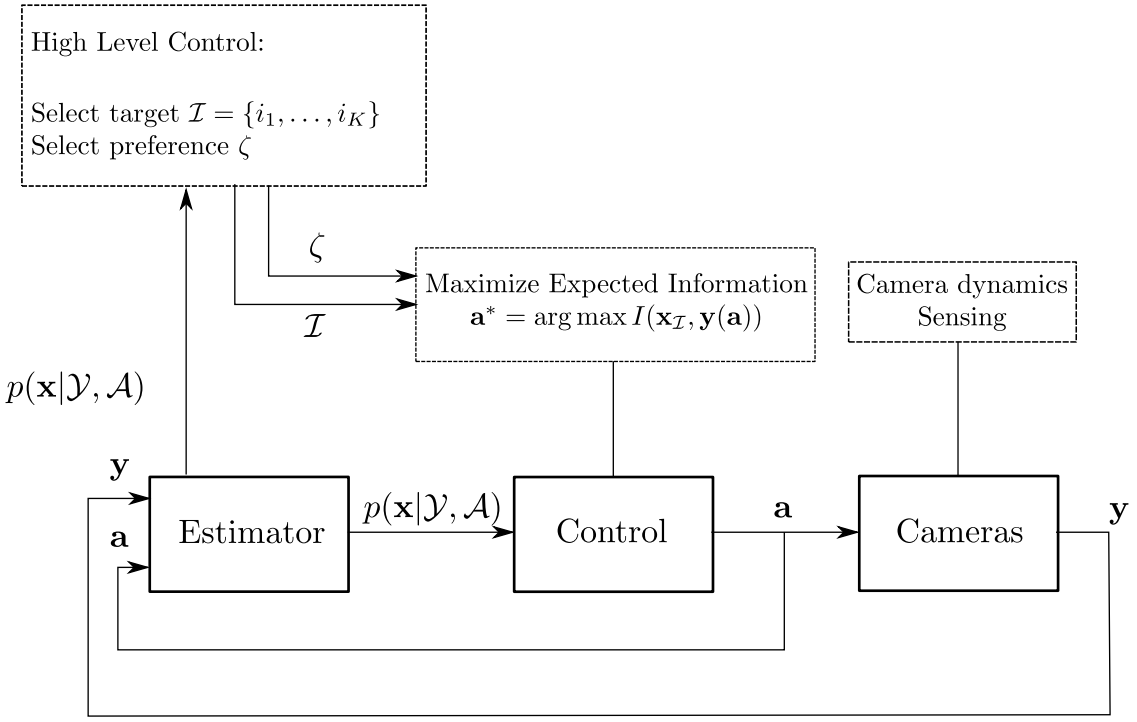


Figure 8.2: The proposed control method can act as a versatile middle layer between actual device control and high-level inferences and decision making. The latter module – either human or machine – decides upon appropriate parameters for the preference over objectives with ζ, and relevant targets by an index set $\mathcal{I}$.

proach taken here is not limited to three objectives. Other objectives, such as uncertainty reduction for action classification and target motion patterns, are conceivable, and in the form of information theoretic objective functions they can be included in the system in the same fashion.

Our proposed solution thus provides a simple interface to a surveillance system. Higher level control mechanisms can decide about the right choice of parameters. For example, consider a surveillance system in a shopping mall. During night time, a reasonable choice might be to prioritise early detection of targets in the supervised area to raise an alarm as rapidly as possible. During day time, it might be more interesting to follow certain targets. As shown in chapter 6, the geometric constraints on visibility can easily be included in the objective functions, and cameras will hand off targets as soon as they are about to leave the area each can supervise. These two very different preferences – de-

tection at night and tracking during the day – can be addressed easily by the preference parameter ζ.

Furthermore, higher level information, for example that a target needs to be tracked as it is suspected of criminal activity, can be included in the objective function for targets in a similar manner as the visibility information. This makes our method appropriate as a middle layer between actual camera control and high level reasoning, as shown in figure 8.2.

8.2 Contributions

We have presented the following main contributions in this work:

- We present a novel method to address several, conflicting objectives in control of surveillance systems. While this method has its roots in the robotics community, this method is completely novel to the field of visual surveillance (Chapter 3).

- We show how the overall control performance can be steered according to a single parameter that expresses a preference for detection or tracking (Chapter 6).

- We extend an information theoretic objective function based on the covariance of a tracked target's Kalman filter. We add a running average of the innovation to the covariance of the observation noise, and show how this effectively increases available zoom range and robustness (Chapter 4).

- We show how this objective function can be used to control camera parameters for multiple targets (Chapter 5).

- We introduce a novel term into the objective functions that addresses the detection of new targets for further tracking. This term is put into competition with the

objective function for tracking, resulting in a camera behaviour that alternates between tracking and exploration (Chapter 5), and provides higher tracking accuracy of more detected targets.

- We show how to include scene knowledge in the objective function by modifying the visibility term (Chapter 6).
- We turn the output of an object detector into a sensor model, which can then be used as the basis for an information theoretic objective function that addresses collaborative detection of several cameras (Chapter 6).
- We integrate classification performance into the scheduling of targets for identification. (Chapter 7).
- We perform a quantitative evaluation of the proposed methods, based on simulation.
- We provide a qualitative evaluation on a live system, based on a database based architecture, and a client implementation that performs tracking as well as detection in parallel (Section 3.9).

8.3 Future Work

This work addressed several areas of interest which justify further investigation. We first discuss questions that can be approached directly, and then propose a long term direction for future research.

8.3.1 Data association

Chapter 5 briefly addressed the issue of data association in the objective function, which involves the association of each of the observations from the cameras with the currently

tracked targets, and decisions regarding instantiation of new targets, as well as cessation of tracks with have not been updated for a certain time. To integrate data association in the information gain, the expectation step in equation 3.26 – where the predicted state estimates are updated according to all measurements that can be expected – needs to be evaluated for all possible data associations.

Standard approaches, such as the JPDAF [13] and multi-hypothesis tracking [123], maintain the current estimate as a set of Kalman filters for each target, and model the interdependencies of targets through an association matrix, which is filled according to a distance measure between expected and observed locations. The expected updated covariance of each filter will be a Gaussian mixture model. The information theoretic measure for the resulting distributions can thus not be obtained in an explicit form as in the single Gaussian case.

Similarly, more recent approaches to data association use particle filters to approximate the distribution of the targets' states [83]. We thus propose a straightforward extension to employ numerical quadrature [74, 125, 146] or approximations methods [76] for the calculation of these.

8.3.2 Target reacquisition

As explained in the future work section of chapter 6, the problem of target reacquisition is currently not addressed in this work. Instead, we rely on the development of the target's covariance to remain bounded, such that a new observation in the region of the target's last observation can be uniquely associated to the previous target's filter. This puts a limit on the proximity of targets. A better method would be to address explicitly the risk of misassignment when targets are sufficiently close or approaching, or an extension of the tracking process to include features that allow re-identification of previously tracked

targets. As in the approach to camera control to benefit data association, this could trigger a purposive zoom onto the target to improve discriminability of a set of targets.

8.3.3 Live System

Due to constraints on time, the experiments with a full implementation of a live system were limited. Although the approaches presented already show an performance increase when compared to standard approaches on synthetic data, it would be a strong argument in favour of this work if the results from the quantitative evaluation can be confirmed in a complete, live system implementation. The drawback in the approach of dynamically using active cameras is the increased difficulty of tracking and detecting targets robustly. Despite advances in object detection methods, single object detections are not fully reliable. We have addressed this fact in joint work with Breitenstein [28], where we tried to learn the reliability of an object detector with respect to the detection's location in the environment. Static cameras provide a (more or less) static background, which can be exploited to make a system far less susceptible to false detections.

8.3.4 Inclusion of scene knowledge

We have already started to include prior knowledge about the scene in chapter 6 by appropriate modification of the visibility term according to occlusions. In work conducted with collaborators, we have addressed the automatic extraction of scene knowledge for future use in control [11, 28, 53]. In the case of simple motion patterns, the inclusion and use of the learnt models should be straightforward. For example, if instead of a fixed motion model of targets, particles are propagated along the learnt motion pattern, and mutual information and entropy can then be estimated numerically.

The use of camera control to enhance scene knowledge, however, might not be as straightforward. For example, typical paths, actions and intentions of targets can be extracted from a supervised area [39, 42, 89]. Two approaches can be taken: an information theoretic objective function is phrased that rewards reduction of uncertainty in the estimation of the current target's class of activity or its path. This is conceptually very similar to the approach in chapter 7. Take for example a target that can move on one out of several parallel paths. For optimal reduction in uncertainty about the actual path the target takes, the camera needs to find a position that allows a discernible view of the paths the target could be moving on. This could be a view of both paths, which allows a more certain assignment of the target to the path. Alternatively, one of the paths the target might move on can be selected. Information is still gained if the target is not discovered on the path chosen.

A step further would be an objective function that rewards a better estimate of the different classes. That is, instead of continuously classifying targets with respect to a previously learnt set of classes, the objective function needs to address reduction of uncertainty about the number of different classes. As an example, consider Damen and Hogg's bicycle stand scenario [39]. Here the events of dropping and picking up a bike are fixed events that have been manually specified. If there is uncertainty about the types of events that can occur, then an information theoretic objective function can serve to reduce this. In this concrete example, a third type of event, bike repair, could exist that is indiscernible at low zoom levels.

8.3.5 Planning Horizon

Among the work that this thesis touched upon, planning over multiple steps has been addressed only in the context of linearisable optimal control and for small state spaces.

Learning of scene knowledge could be employed to provide a simplification of the latter, sufficient to address longer planning horizons. This could provide the camera a mid-term objective, e.g. acquisition of a target at a meeting point in the supervised area which has been learnt from long term observations. A greedy optimisation then adjusts the path from mid-term control in a reactive fashion upon scene activity. Some mid-term goal abstractions have been used by Qureshi and Terzopolous [122], but there seems to be no work towards automatic learning of these concepts.

8.3.6 Reinforcement Learning

The methods proposed in this thesis rely on a good model for the observation, target motion and control processes. Reinforcement learning is a model free approach that can learn the mapping from current state and action to future state and an appropriate reward function. It has been applied in the context of visual surveillance for relatively small action spaces, e.g. tracking and acquisition of single targets only [9, 45, 47, 105], as the state space of multiple, interacting targets is too large for a complete coverage by standard reinforcement learning methods. Principled inclusion of prior knowledge, such as target motion patterns [53, 80], as well as constraints on the action space [84] might be an approach to alleviate this shortcoming of this method.

8.3.7 Higher Level Control

As described in section 8.1.1, our method provides a simple interface to express preferences for certain objectives and targets that are to be followed. These can be set by human operators. A future line of research is to learn this higher level control using scene knowledge, e.g. night and daytime patterns, or to incorporate more higher order logic in the form of spatio-temporal rules [103].

Appendix A

Information-theoretic Identities

A.1 Derivation of Mutual Information

Mutual information measures the dependence between two random variables. For two sets of symbols, $\mathcal{S} := s_1, ..., s_n$ and $\mathcal{T} := t_1, ..., t_n$, let the probability of encountering symbol t_i be $p(t_1)...p(t_n)$, and the joint likelihood ot two events s_i, t_j is denoted as $p(s_i, t_j)$

$$
\begin{aligned}
I(\mathcal{S};\mathcal{T}) &= \mathrm{E}_{\mathcal{S},\mathcal{T}}\left\{\log\frac{p(s_i,t_j)}{p(s_i)t(s_j)}\right\} && \text{(A.1)}\\
&= \sum_{\mathcal{S},\mathcal{T}} p(s_i,t_j)(\log(p(s_i,t_j)) - \log(p(s_i)p(t_j))) && \text{(A.2)}\\
&= \sum_{\mathcal{S}} p(s_i) \sum_{\mathcal{T}} p(t_j|s_i)(\log(p(t_j|s_i)) - \log(p(t_j))) && \text{(A.3)}\\
&= H(\mathcal{T}) - \sum_{\mathcal{S}} p(s_i) H(\mathcal{T}|s_i) && \text{(A.4)}\\
&= H(\mathcal{T}) - H(\mathcal{T}|\mathcal{S}) && \text{(A.5)}
\end{aligned}
$$

The difference between entropy and mutual information in equation A.5 is the conditional entropy $H(\mathcal{T}|\mathcal{S})$, which can be interpreted as the amount of uncertainty remaining in $\mathcal{T}$ once $\mathcal{S}$ is known.

A.2 Mutual Information always increases with more data

To show that on average, mutual information does not decrease if more evidence is added,

$$I(x;a,b) \geq I(x;a) \tag{A.6}$$

we have to show that

$$H(x) - H(x|a,b) \geq H(x) - H(x|a) \tag{A.7}$$

We proceed by reducing the expression to a form of mutual information between the variables, which is always positive. For this, we subtract $H(x)$ and use the definition of entropy

$$\begin{aligned} H(x|a,b) - H(x|a) &= \mathrm{E}_{x,a,b}\{-\log p(x|a,b)\} - \mathrm{E}_{x,a}\{-\log(p(x|a)\} && \text{(A.8)} \\ &= \mathrm{E}_{x,a,b}\{\log(p(x|a)/(p(b|x,a)p(x|a)/p(b))\} && \text{(A.9)} \\ &= \mathrm{E}_{x,a,b}\{\log(p(b)/(p(b|x))\} && \text{(A.10)} \\ &= -I(b;x) \leq 0 && \text{(A.11)} \end{aligned}$$

In equation A.9 we used the substitution $p(x|a,b) = p(x|a)p(b|x,a)/p(b)$, and made use of the independence assumption of the observations, $p(b,a,x) = p(b|x)p(a|x)p(x)$ resulting in $p(b|x,a) = p(b,a,x)/p(x,a) = p(b|x)$.

This can also be shown using the "information never hurts" bound, i.e. for two sets $\mathcal{A}, \mathcal{B}$ of observations, $H(x|\mathcal{A}) \geq H(x|\mathcal{A} \cup \mathcal{B})$ [36, 87].

A.3 Mutual Information yields diminishing returns

The relative mutual information gained from N to $N+1$ observations is less than the relative mutual information gained from $N-1$ to N observations. This can be expressed as

$$I(x;a,b) - I(x;a) \geq I(x;a,b,c) - I(x;a,b) \tag{A.12}$$

We proceed similar to section A.2 and subtract $H(x)$, then use the definition of entropy

$$
\begin{aligned}
& I(X;A,B) - I(X;A) - (I(X;A,B,C) - I(X;A,B)) && \text{(A.13)}\\
&= H(X) - H(X|A,B) - H(X) + H(X|A) && \text{(A.14)}\\
&\quad - (H(X) - H(X|A,B,C) - H(X) + H(X|A,B)) && \text{(A.15)}\\
&= -H(X|A,B) + H(X|A) - (-H(X|A,B,C) + H(X|A,B)) && \text{(A.16)}\\
&= \mathrm{E}_{a,b,c,x}\log(p(x|a)) - \log(p(x|a,b)) + \log(p(x|a,b,c)) - \log(p(x|a,b)) && \text{(A.17)}\\
&= \mathrm{E}_{a,b,c,x}\log((p(x|a)p(x|a,b,c))/(p(x|a,b)p(x|a,b))) && \text{(A.18)}\\
&= \mathrm{E}_{a,b,c,x}\log((p(x|a)p(x|a,b)p(c|x,a,b)/p(c))/(p(x|a,b)p(x|a,b))) && \text{(A.19)}\\
&= \mathrm{E}_{a,b,c,x}\log((p(x|a)p(c|x,a,b))/(p(x|a,b)p(c))) && \text{(A.20)}\\
&= \mathrm{E}_{a,b,c,x}\log((p(x|a)p(c|x,a,b))/((p(x|a)p(b|x,a)/p(b))p(c))) && \text{(A.21)}\\
&= \mathrm{E}_{a,b,c,x}\log((p(c|x,a,b)p(b))/(p(b|x)p(c))) && \text{(A.22)}\\
&= \mathrm{E}_{a,b,c,x}\log((p(c,b|x,a))/(p(b|x,a)p(c))) && \text{(A.23)}\\
&= I(c;b|x,a) \geq 0 && \text{(A.24)}
\end{aligned}
$$

In equation A.18 we used $p(x|a,b,c) = p(x|a,b)p(c|x,a,b)/p(c)$, similarly in equation A.21.

A.4 Difference between Differential and Discrete Entropy

Assume the probability density function $p(\mathbf{x})$ is sufficiently smooth that we can approximate the cumulative density function (using the mean value theorem) for sufficiently small Δ.

$$
\begin{aligned}
p(\mathbf{x} - \Delta/2 < \mathbf{x} < \mathbf{x} + \Delta/2) &= \int_{\mathbf{x}-\Delta/2}^{\mathbf{x}+\Delta} p(\mathbf{x})d\mathbf{x} && \text{(A.25)}\\
&= \Delta p(\mathbf{x}) && \text{(A.26)}
\end{aligned}
$$

We now discretise the function $p(x)$ and enumerate the bins $p(\Delta i) = \Delta p(\mathbf{x})$.

Using the discrete version of entropy, we obtain

$$H(\Delta i) = \sum \Delta p(\Delta i) \log(\Delta p(\Delta i)) \tag{A.27}$$

When now considering the limit of $\Delta \to 0$ (and thus the exact representation of $p(x)$):

$$\begin{aligned} H(\mathbf{x}) &= -\lim_{\Delta \to 0} \sum \Delta p(\Delta i) \log(\Delta p(\Delta i)) && \text{(A.28)} \\ &= -\lim_{\Delta \to 0} \sum \Delta p(\Delta i) \log p(\Delta i) && \text{(A.29)} \\ &+ -\lim_{\Delta \to 0} \sum \Delta p(\Delta i) \log \Delta && \text{(A.30)} \\ &= -\int \log p(\mathbf{x})\, d\mathbf{x} - \lim_{\Delta \to 0} \log \Delta && \text{(A.31)} \end{aligned}$$

The latter term is not defined, and Shannon proposed to simply ignore this term.

A.5 Entropy of Multivariate, Normal-distributed Random Variables

Using the definition of continuous entropy in section 3.3.3, we can derive a closed form solution for the entropy of a random variable with Gaussian distribution.

For a k-dimensional vector $\mathbf{x}$ with covariance $\Sigma = \mathrm{E}\{\mathbf{x}^T\mathbf{x}\}$ and mean $\boldsymbol{\mu} = \mathrm{E}\{\mathbf{x}\}$, the normal distribution is

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp(-{}^1\!/_2 (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})) \tag{A.32}$$

The entropy is

$$
\begin{aligned}
H(\mathbf{x}) =& \mathrm{E}\{-\log p(\mathbf{x})\} && \text{(A.33)}\\
=& \mathrm{E}\left\{ {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + {}^1\!/_2 (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\} && \text{(A.34)}\\
=& {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + {}^1\!/_2 \mathrm{E}\left\{ (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\} && \text{(A.35)}\\
=& {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + {}^1\!/_2 \mathrm{E}\left\{ \sum_i ((\mathbf{x}-\boldsymbol{\mu})^T)_i \sum_j (\Sigma^{-1})_{ij} (\mathbf{x}-\boldsymbol{\mu})_j \right\} && \text{(A.36)}\\
=& {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + {}^1\!/_2 \mathrm{E}\left\{ \sum_{ij} ((\mathbf{x}-\boldsymbol{\mu})^T)_i (\mathbf{x}-\boldsymbol{\mu})_j (\Sigma^{-1})_{ij} \right\} && \text{(A.37)}\\
=& {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + {}^1\!/_2 \sum_{ij} \mathrm{E}\left\{ ((\mathbf{x}-\boldsymbol{\mu})^T)_i (\mathbf{x}-\boldsymbol{\mu})_j \right\} (\Sigma^{-1})_{ij} && \text{(A.38)}\\
=& {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + {}^1\!/_2 \sum_{ji} \Sigma_{ji} (\Sigma^{-1})_{ij} && \text{(A.39)}\\
=& {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + {}^1\!/_2 \sum_j \sum_i \Sigma_{ji} (\Sigma^{-1})_{ij} && \text{(A.40)}\\
=& {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + {}^1\!/_2 \sum_j (\Sigma\Sigma^{-1})_{jj} && \text{(A.41)}\\
=& {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + {}^1\!/_2 \sum_j \mathbf{I}_{jj} && \text{(A.42)}\\
=& {}^1\!/_2 \log((2\pi)^k \det(\Sigma)) + \frac{k}{2} && \text{(A.43)}\\
=& {}^1\!/_2 \log((2\pi e)^k \det(\Sigma)) && \text{(A.44)}
\end{aligned}
$$

Appendix B

Birth and Death Process

A system is in two states, empty or occupied by one or more targets. The event that it is empty after a small duration h occurs if it was empty beforehand and no nobody arrives

$$p(A) = \underbrace{p_0(t)e^{-\lambda h}}_{\text{empty, nothing new}} = p_0(t)(1 - \lambda h - \underbrace{o(h)}_{\text{higher order terms}}) \tag{B.1}$$

or occupied but vacated during the given time frame

$$p(B) = \underbrace{p_1(t)(1 - e^{-\nu h})}_{\text{occupied, vacated}} = p_1(t)(\underbrace{\nu h}_{\text{limit behaviour}} - o(h)) \tag{B.2}$$

The resulting probability of being empty is the sum of those mutually exclusive terms

$$p_0(t + h) = p(A) + p(B) = p_0(t)(1 - \lambda h) + \nu h p_1(t) + o(h) \tag{B.3}$$

When taking $\lim_{h \to 0} p_0$ this leads to

$$\dot{p_0} = \begin{bmatrix} -\lambda & \nu \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \end{bmatrix} \tag{B.4}$$

Which can be solved as a ODE knowing that $p_0 = 1 - p_1$, and using any kind of boundary condition $p_0(0) = \alpha$:

$$p_0(t) = \left(\alpha - \frac{\nu}{\lambda + \nu}\right) e^{-(\lambda+\nu)t} + \frac{\nu}{\lambda + \nu} \tag{B.5}$$

Appendix C

Kalman Filter

C.1 Kalman filter

The derivation and resulting equations of the Kalman filter can be found in textbooks, e.g. [13]. We give only a short overview over the main equations. We want to estimate the state of a random variable $\mathbf{x}_k \in \mathbb{R}^n$ from a noisy observation $\mathbf{o}_k \in \mathbb{R}^m$. We are given a prior estimate of the state from previous time step $k-1$, which has a normal distribution according to prior covariance $\mathbf{P}_{k-1}^+$. The state of the system develops according to a model

$$\mathbf{x}_k^- = \mathbf{f}(\mathbf{x}_{k-1}^+, \mathbf{a}_k) + \mathbf{v}_k \tag{C.1}$$

where $\mathbf{a}_k$ is the instantaneous control input to the system, and $\mathbf{v}_k$ is additive Gaussian noise according to covariance $\mathbf{Q}_k^{n\times n}$.

The observation model is

$$\mathbf{o} = \mathbf{h}(\mathbf{x}_k, \mathbf{a}_k) + \mathbf{w}_k, \tag{C.2}$$

and we know the noise $\mathbf{w}_k$ is also normal distributed according covariance $\mathbf{R}_k^{m\times m}$.

This yields the predictions

$$\mathbf{x}_k^- = \mathbf{f}(\mathbf{x}_{k-1}^+, \mathbf{a}_k) \tag{C.3}$$

$$\mathbf{P}_k^- = \nabla\mathbf{f}_\mathbf{x}\mathbf{P}_{k-1}^+\nabla\mathbf{f}_\mathbf{x}^T + \mathbf{Q}_k \tag{C.4}$$

and the updates with observation $\mathbf{o}_k$

$$\mathbf{x}_k^+ = \mathbf{x}_k^- + \mathbf{W}_k \boldsymbol{\nu}_k \tag{C.5}$$

$$\mathbf{P}_k^+ = \mathbf{P}_k^- - \mathbf{W}_k \mathbf{S}_k \mathbf{W}_k^T \tag{C.6}$$

The innovation is

$$\boldsymbol{\nu}_k = \mathbf{o}_k - \mathbf{h}(\mathbf{x}_k^-, \mathbf{a}_k) \tag{C.7}$$

and the innovation covariance is

$$\mathbf{S}_k = \nabla \mathbf{h}_\mathbf{x} \mathbf{P}_k^- \nabla \mathbf{h}_\mathbf{x}^T + \mathbf{R}_k. \tag{C.8}$$

Finally the Kalman gain is

$$\mathbf{W}_k = \mathbf{P}_k^- \nabla \mathbf{h}_\mathbf{x}^T S_k^{-1} \tag{C.9}$$

The matrices $\nabla \mathbf{f}_\mathbf{x}$ and $\nabla \mathbf{h}_\mathbf{x}$ are the Jacobians of $\mathbf{f}$ and $\mathbf{h}$, respectively, evaluated at the latest available estimates. The observation matrix $\nabla \mathbf{h}_\mathbf{x} = \mathbf{H}(\mathbf{a})$ reflects the dependency on the current zoom or other parameter $\mathbf{a}$.

The noise covariances for a constant velocity target can be derived from the continuous-time state equation and continuous-time white noise process by integrating over the sampling interval T [13, p84]. The resulting covariances per axis x y of the observed target are then

$$\mathbf{Q}_{xy} = \begin{bmatrix} {}^1\!/_3\, T^3 & {}^1\!/_2\, T^2 \\ {}^1\!/_2\, T^2 & T \end{bmatrix} \tag{C.10}$$

and similarly for the covariance of the state dynamics.

C.2 Jacobians for Various Observation Functions

C.2.1 Pin-hole Projection

A point at world coordinates $\mathbf{x}$ is transformed into camera coordinates by

$$\mathbf{g}(\mathbf{x}) = \mathbf{R}\mathbf{x} + \mathbf{t}, \tag{C.11}$$

for a camera at with rotation $\mathbf{R}^T$ and position $\mathbf{R}^T\mathbf{t}$. Its Jacobian is

$$\frac{d}{d\mathbf{x}}\mathbf{g}(\mathbf{x}) = \mathbf{R}. \tag{C.12}$$

The planar projection of a point $\mathbf{x}_c = [x\ y\ z]^T$ in camera coordinates with a camera of focal length f is

$$\mathbf{h}(\mathbf{x}_c) = [\hat{x}\ \hat{y}]^T = \left[\frac{fx}{z}\ \frac{fy}{z}\right]^T. \tag{C.13}$$

The Jacobian is

$$\frac{d}{d\mathbf{x}_c}\mathbf{h}(\mathbf{x}_c) = f/z \begin{bmatrix} 1 & 0 & -x/z \\ 0 & 1 & -y/z \end{bmatrix}. \tag{C.14}$$

The Jacobian of the projection of a point in world camera coordinates is thus

$$\frac{d}{d\mathbf{x}}s(\mathbf{x}) = \frac{d}{d\mathbf{x}}h(g(\mathbf{x})) = \frac{d}{d\mathbf{x}_c}h(\mathbf{x}_c)\Big|_{\mathbf{R}\mathbf{x}+\mathbf{t}}\mathbf{R} \tag{C.15}$$

C.2.2 Spherical Projection

The spherical projection of a point $\mathbf{x}_c = [x\ y\ z]^T$ in camera coordinates is

$$\mathbf{h}(\mathbf{x}_c) = \left[\frac{x}{\sqrt{x^2+y^2+z^2}}\ \frac{y}{\sqrt{x^2+y^2+z^2}}\right]^T. \tag{C.16}$$

With $r = \sqrt{x^2+y^2+z^2}$, the Jacobian is

$$\frac{d}{d\mathbf{x}_c}\mathbf{f}(\mathbf{x}_c) = 1/r^3 \begin{bmatrix} r^2-x^2 & -xy & -xz \\ -xy & r^2-y^2 & -yz \end{bmatrix} \tag{C.17}$$

C.2.3 Pin-hole Projection in Spherical Coordinates

With $r = \sqrt{x^2+y^2+z^2}$, a parameterisation of a point $\mathbf{x}_c$ in camera coordinates is

$$\mathbf{s}(\mathbf{x}) = \begin{bmatrix} \rho \\ \theta \end{bmatrix} = \begin{bmatrix} \arctan x/z \\ \arcsin y/r \end{bmatrix}. \tag{C.18}$$

The same angles can be obtained from image coordinates

$$\begin{bmatrix} \rho \\ \theta \end{bmatrix} = \begin{bmatrix} \arctan \hat{x}/f \\ \arcsin \hat{y}/\sqrt{(\hat{x}^2+\hat{y}^2+f^2)} \end{bmatrix}. \tag{C.19}$$

The Jacobian w.r.t. camera coordinates is

$$\frac{d}{d\mathbf{x}_c} s(\mathbf{x}_c) = \begin{bmatrix} \frac{a}{z} & 0 & \frac{-ax}{z^2} \\ -bxy & b(x^2+z^2) & -byz \end{bmatrix} \tag{C.20}$$

with

$$a = (1 + (x/z)^2)^{-1} \tag{C.21}$$

and

$$b = \frac{1}{r^3\sqrt{1-(y/r)^2}}. \tag{C.22}$$

Appendix D

Classification Systems

D.1 Biometric performance indicators

Here we relate the performance indicators of biometric systems to the confusion matrix of a classifier. The false acceptance rate (FAR) is the percentage of potentially not registered targets wrongly identified as known; The false reject rate (FRR) is the percentage of registered targets which are wrongly labelled as not registered. The "failure to capture" rate (FCR) is the likelihood of not acquiring the required data for the classification, and the output can be classified as "unknown".

These indicators are related to the confusion matrix in the following manner.

$$p(a = t|T) = (1 - FCR)(1 - FRR) \quad \text{(D.1)}$$

$$p(a = R|T) = (1 - FCR)FRR \quad \text{(D.2)}$$

$$p(T|t = R) = (1 - FCR)FAR \quad \text{(D.3)}$$

$$p(a = R|t = R) = (1 - FCR)(1 - FAR) \quad \text{(D.4)}$$

$$p(a = U|T) = FCR \quad \text{(D.5)}$$

$$p(a = U|t = R) = FCR \quad \text{(D.6)}$$

D.2 Worked example of classification

Here we give a full example of the continuous application of the classifier to a set of targets.

In our control approach, we are interested to pursue targets as long as their identity has not yet been established, i.e. the target was found in the database or confirmed as an unknown target. Once this has happened, further action should be taken, e.g. an alarm rung or the next target selected for identification.

We now show how the entropy of the target's state decreases for continuous measurements. For every identification on a set of images, the prior belief about the target state is updated into $p(e|d)$ according to the likelihood $p(d|e)$. In the next time step, this updated posterior is used as current prior. Without prior knowledge about the target, we take the approach of maximum uncertainty, and initialise the belief with a non-informative prior $p(e=1)=0.5$.

The development of $p(e|d)$ is shown in figure D.1 for 1000 runs of 100 continuous classifications for an enrolled and not enrolled target[1]. The detections are generated according to a performance matrix for a poor classification method and where classifications frequently are impossible:

$$p(d|e) = \begin{array}{c|c|c} & e=1 & e=0 \\ \hline d=1 & 0.425 & 0.2 \\ d=0 & 0.075 & 0.3 \\ d=U & 0.5 & 0.5 \end{array}, \tag{D.7}$$

[1] Note that the standard deviation that is indicated in these plots assumes an unbounded domain and thus simple addition to the mean can yield values beyond the range $\{0\ldots1\}$. A better way to display this uncertainty would be a box and whiskers plot of the values $p(e|d)$ per time step.

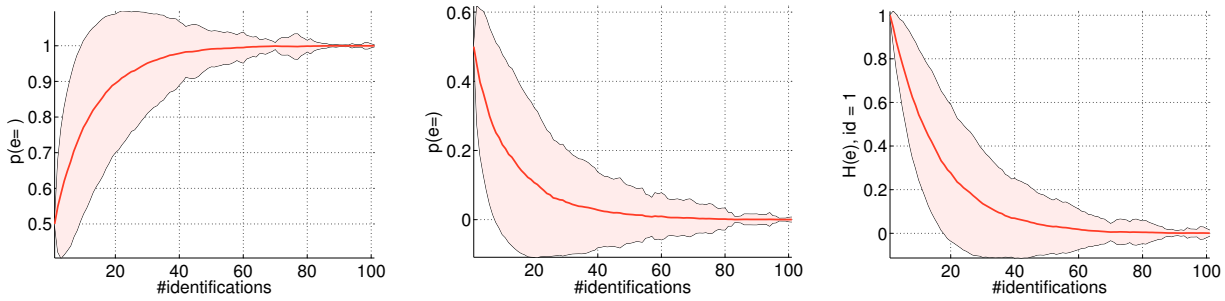


Figure D.1: Left: Development of classification for a known target (e=1), averaged over 1000 trials. Centre: same for unknown target (e=0). Right: Entropy in the state estimate. Results are visually indistinguishable for $e = 1$ and $e = 0$.

which was obtained from an initial confusion matrix for two different registered identities and an "unregistered" class:

$$p(a|t) = \begin{array}{c|ccc} & t=1 & t=2 & t=R \\ \hline a=1 & 0.4 & 0.1 & 0.1 \\ a=2 & 0.05 & 0.3 & 0.1 \\ a=R & 0.05 & 0.1 & 0.3 \\ a=U & 0.5 & 0.5 & 0.5 \end{array} , \tag{D.8}$$

These classes used the prior $p(t) = [0.4\ 0.4\ 0.2]$.

Bibliography

[1] A. E. Abdel-Hakim and A. A. Farag. Robust virtual forces-based camera positioning using a fusion of image content and intrinsic parameters. In *8th International Conference on Information Fusion*, volume 1, pages 522–529, 2005. 2.2, 7.2

[2] Besma R. Abidi, Nash R. Aragam, Yi Yao, and Mongi A. Abidi. Survey and analysis of multimodal sensor planning and integration for wide area surveillance. *ACM Comput. Surv.*, 41:7:1–7:36, January 2009. 2.4

[3] S. Abrams, P.K. Allen, and K.A. Tarabanis. Dynamic sensor planning. In *DARPA 1993 Image Understanding Workshop*, pages 599–607, 1993. 2.4

[4] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(11):1475 –1490, November 2004. 2

[5] Jania Aghajanian and Simon J.D. Prince. Face pose estimation in uncontrolled environments. In *Proceedings of the British Machine Vision Conference*, London, 2009. 7.6

[6] Ravindra K. Ahuja, Arvind Kumar, Krishna C. Jha, and James B. Orlin. Exact and heuristic algorithms for the weapon-target assignment problem. *Operations Research*, 55:1136–1146, 2007. 7.2

[7] Murad Al Haj, Andrew D. Bagdanov, Jordi Gonzàlez, and Xavier F. Roca. Reactive object tracking with a single PTZ camera. In *20th International Conference on Pattern Recognition (ICPR2010)*, Istanbul, Turkey, August 2010. 4.2

[8] N. E. Apostoloff and A. Zisserman. Who are you? Real-time person identification. In *Proceedings of the British Machine Vision Conference*, 2007. 7.3.2

[9] Andrew D. Bagdanov, Alberto Del Bimbo, Walter Nunziati, and Federico Pernici. A reinforcement learning approach to active camera foveation. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, VSSN '06, pages 179–186, New York, NY, USA, 2006. ACM. 5.2, 8.3.6

[10] Andrew D. Bagdanov, Alberto Del Bimbo, and Federico Pernici. Acquisition of high-resolution images through on-line saccade sequence planning. In *VSSN '05: Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, pages 121–130, New York, NY, USA, 2005. ACM. 2.2, 5.2, 7.2

[11] Pau Baiget, Eric Sommerlade, Ian Reid, and Jordi Gonzàlez. Finding prototypes to estimate trajectory development in outdoor scenarios. In *Proceedings of the First International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS2008)*, September 2008. 1.6, 8.3.4

[12] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1006, August 1988. 1.2, 3.5.1

[13] Yaakov Bar-Shalom and Thomas E. Fortmann. *Tracking and data association*, volume 179 of *Mathematics in Science and Engineering*. Academic Press Professional, Inc., San Diego, CA, USA, 1987. 3.7, 4.4, 5.3, 6.2.1, 8.3.1, C.1, C.1

[14] Faisal Bashir and Fatih Porikli. Collaborative tracking of objects in EPTZ cameras. In Chang W. Chen, Dan Schonfeld, and Jiebo Luo, editors, *Visual Communications and Image Processing 2007*, volume 6508. SPIE, 2007. 5.4.6

[15] Jorge P. Batista. Tracking pedestrians under occlusion using multiple cameras. In *Image Analysis and Recognition*, pages 552–562. Springer, 2004. 2.2

[16] N. Bellotto and H. Hu. People Tracking with a Mobile Robot: a Comparison of Kalman and Particle Filters. In *Proc. of the 13th IASTED Int. Conf. on Robotics and Applications*, pages 388–393, 2007. 3.9.2

[17] Nicola Bellotto, Eric Sommerlade, Ben Benfold, Charles Bibby, Ian Reid, Daniel Roth, Luc Van Gool, Carles Fernández, and Jordi Gonzàlez. A distributed camera system for multi-resolution surveillance. In *Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2009)*, 2009. 1.5, 1.6, 3.9, 3.9, 3.10

[18] Ben Benfold and Ian Reid. Guiding visual surveillance by tracking human attention. In *Proceedings of the 20th British Machine Vision Conference*, September 2009. 7.2, 7.6, 7.6, 7.6, 7.7.2

[19] Charles Bibby and Ian Reid. Robust real-time visual tracking using pixel-wise posteriors. In *Proceedings of the 2008 European Conference on Computer Vision*, 2008. 3.9.2, 3.2

[20] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 3.1, 3.2, 3.3.3

[21] P. Blauensteiner and M. Kampel. Visual surveillance of an airport's apron - an overview of the AVITRACK project. In W. Burger and J. Scharinger, editors, *Digital Imaging in Media and Education, Proc. of the 28th Workshop of the Austrian Association for Pattern Recognition (OAGM/AAPR)*, volume 179 of *Schriftenreihe der OCG*, pages 213–220, 2004. 1.1

[22] Robert Bodor, Andrew Drenner, Paul Schrater, and Nikolaos Papanikolopoulos. Optimal camera placement for automated surveillance tasks. *J. Intell. Robotics Syst.*, 50(3):257–295, 2007. 5.2

[23] Yvo Boers, Hans Driessen, Arunabha Bagchi, and Pranab Mandal. Particle filter based entropy. In *Proceedings of the 13th International Conference on Information Fusion, FUSION 2010*. The Institution of Engineering and Technology, July 2010. 3.7

[24] T. E. Boult, R. J. Micheals, X. Gao, and M. Eckmann. Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings. In *Proc. of the IEEE, 89(10)*, pages 1382–1402, October 2001. 3.2

[25] Frédéric Bourgault, Tomonari Furukawa, and Hugh F. Durrant-Whyte. Optimal search for a lost target in a bayesian world. In S. Yuta and H. Asama, editors, *Field and Service Robots IV*, volume 24 of *Springer Tracts in Advanced Robotics (STAR)*, pages 209–222. Springer-Verlag, 2006. 5.2

[26] Frédéric Bourgault, Alexei A. Makarenko, Stefan B. Williams, Ben Grocholsky, and Hugh F. Durrant-Whyte. Information based adaptive robotic exploration. In *Proc. IEEE/RSJ Intl Conf. on Intelligent Robots and Systems*, 2002. 3.6, 5.2

[27] Kevin J Bradshaw, Philip F McLauchlan, Ian D Reid, and David W Murray. Saccade and pursuit on an active head/eye platform. *Image and Vision Computing*, 12(3):155 – 163, 1994. BMVC 1993. 2.3

[28] Michael D. Breitenstein, Eric Sommerlade, Bastian Leibe, Luc van Gool, and Ian Reid. Probabilistic parameter selection for learning scene structure from video. In *British Machine Vision Conference (BMVC'08)*, September 2008. 1.6, 8.3.3, 8.3.4

[29] Derek W. Bunn. *Analysis for Optimal Decisions*. John Wiley & Sons, Inc, 1982. 3.6

[30] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 7.3, 7.6

[31] F. Chaumette and S. Hutchinson. Visual servo control, part I: Basic approaches. *IEEE Robotics and Automation Magazine*, 13(4):82–90, December 2006. 2.3

[32] Amit Singh Chhetri. *Sensor Scheduling and Efficient Algorithm Implementation For Target Tracking*. PhD thesis, Arizona State University, 2006. 2.5

[33] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(5):564–577, 2003. 4.5.1

[34] Diane J. Cook, Piotr Gmytrasiewicz, and Lawrence B. Holder. Decision-theoretic cooperative sensor planning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):1013–1023, 1996. 3.6

[35] Cash J. Costello, Christopher P. Diehl, Amit Banerjee, and Hesky Fisher. Scheduling an active camera to observe people. In *VSSN '04: Proceedings of the ACM 2nd international workshop on Video surveillance & sensor networks*, pages 39–45, New York, NY, USA, 2004. ACM Press. 2.2, 2.5, 7.2

[36] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc, July 2006. A.2

[37] Carla Currin, Toby Mitchell, Max Morris, and Don Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):pp. 953–963, 1991. 2.4

[38] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 886–893, Los Alamitos, CA, USA, June 2005. IEEE Computer Society. 7.7.2

[39] Dima Damen and David Hogg. Recognizing linked events: Searching the space of feasible explanations. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009. 8.3.4

[40] K. Daniilidis, C. Krauss, M. Hansen, and G. Sommer. Real-time tracking of moving objects with an active camera. *Real-Time Imaging*, 4(1):3 – 20, 1998. 2.3

[41] J. Davis, A. Morison, and D. Woods. An adaptive focus-of-attention model for video surveillance and monitoring. *Machine Vision and Applications*, 18(1):41–64, February 2007. 5.2

[42] Hannah Dee and David Hogg. Detecting inexplicable behaviour. In *Proceedings of the BMVC*, 2004. 8.3.4

[43] Hannah Dee, David Hogg, and Anthony Cohn. Scene modelling and classification using learned spatial relations. In Kathleen Hornsby, Christophe Claramunt, Michel Denis, and Gérard Ligozat, editors, *Spatial Information Theory*, volume 5756 of *Lecture Notes in Computer Science*, pages 295–311. Springer Berlin / Heidelberg, 2009. 5.4.3

[44] Alberto Del Bimbo, Fabrizio Dini, Giuseppe Lisanti, and Federico Pernici. Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks. *Computer Vision and Image Understanding (CVIU)*, 114(6):611–623, 2010. 5.2

[45] Alberto Del Bimbo and Federico Pernici. Towards on-line saccade planning for high-resolution image sensing. *Pattern Recognition Letters*, 27:1826–1834, November 2006. 2.2, 2.5, 3.8, 5.2, 7.2, 8.3.6

[46] Joachim Denzler, Matthias Zobel, and Heinrich Niemann. Information theoretic focal length selection for real-time active 3-d object tracking. In *9th IEEE International Conference on Computer Vision*, pages 400–407. IEEE Computer Society, 2003. 1.4, 4.2, 4.3, 4.3.1, 4.3.1, 4.5.1, 5.5

[47] Christian Derichs, Frank Deinzer, and Heinrich Niemann. Cost integration in multi-step viewpoint selection for object recognition. In Petra Perner and Atsushi Imiya, editors, *Machine Learning and Data Mining in Pattern Recognition*, volume 3587 of *Lecture Notes in Computer Science*, pages 415–425. Springer Berlin / Heidelberg, 2005. DAGM. 5.2, 8.3.6

[48] Benjamin Deutsch, Stefan Wenhardt, and Heinrich Niemann. Multi-step multi-camera view planning for real-time visual object tracking. In Katrin Franke, Klaus R. Müller, Bertram Nickolay, and Ralf Schäfer, editors, *DAGM-Symposium*, volume 4174 of *Lecture Notes in Computer Science*, pages 536–545. Springer, 2006. 6.2, 6.2.1

[49] T. Dinh, Q. Yu, and G. Medioni. Real time tracking using an active pan-tilt-zoom network camera. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2009. 2.3

[50] Petr Doubek, Indra Geys, Tomás Svoboda, and Luc Van Gool. Cinematographic rules applied to a camera network. In Peter Sturm, Tomás Svoboda, and Seth Teller, editors, *Omnivis2004: The fifth Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras*, pages 17–29, Prague, Czech Republic, 2004. Czech Technical University. 2.2, 7.2

[51] Alberto Elfes. *Occupancy Grids; A Probabilistic Framework for Robot Perception and Navigation*. PhD thesis, Carnegie Mellon University, 1989. 5.2

[52] Alberto Elfes. Robot navigation: Integrating perception, environmental constraints and task execution within a probabilistic framework. In Leo Dorst, Michiel van Lambalgen, and Frans Voorbraak, editors, *Reasoning with Uncertainty in Robotics*,

volume 1093 of *Lecture Notes in Computer Science*, pages 91–130. Springer Berlin / Heidelberg, 1996. 10.1007/BFb0013955. 5.2

[53] David Ellis, Eric Sommerlade, and Ian Reid. Modelling pedestrian trajectories with gaussian processes. In *Ninth International Workshop on Visual Surveillance*, 2009. 1.6, 5.4.3, 8.3.4, 8.3.6

[54] Ugur Murat Erdem and Stan Sclaroff. Automated camera layout to satisfy task-specific and floorplan-specific coverage requirements. *Computer Vision and Image Understanding*, 103(3):156–169, September 2006. 5.2

[55] Jeffrey A. Fayman, Oded Sudarsky, Ehud Rivlin, and Michael Rudzsky. Zoom tracking and its applications. *Machine Vision and Applications*, 13:25–37, 2001. 10.1007/PL00013268. 4.2

[56] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Proceedings of IEEE CVPR Workshop on Generative-Model Based Vision*, 2004. 2

[57] Jerome A. Feldman and Robert F. Sproull. Decision theory and artificial intelligence II: The hungry monkey. *Cognitive Science: A Multidisciplinary Journal*, 1(2):158–192, 1977. 3.4, 3.4.2

[58] Thomas S. Ferguson. *Mathematical Statistics A Decision Theoretic Approach*. Academic Press, Inc, 111 Fifth Av, New York, NY 10003, 1967. 3.4

[59] Tomonari Furukawa, Hugh F. Durrant-Whyte, and Benjamin Lavis. The element-based method - theory and its application to bayesian search and tracking. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007. 5.2

[60] Nicholas R. Gans, G. Hu, and W.E. Dixon. Keeping multiple objects in the field of view of a single ptz camera. In *Proceedings of the 2009 American Control Conference*, pages 5259–5264, St. Louis, MO, USA, June 2009. IEEE Press. 2.3

[61] Tianshi Gao, Chen Wu, and H. Aghajan. User-centric speaker report: Ranking-based effectiveness evaluation and feedback. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1004 –1011, sept. 2009. 7.2

[62] Boris Gnedenko. *Theory of Probability*. Mir Publishers, Moscow, 3rd edition, 1976. 6.3.1, 7.4.1

[63] Shaogang Gong, Tao Xiang, and Somboon Hongeng. Learning human pose in crowd. In *Proceedings of the 1st ACM international workshop on Multimodal pervasive video analysis*, MPVA '10, pages 47–52, New York, NY, USA, 2010. ACM. 7.6

[64] H. González-Banos. A randomized art-gallery algorithm for sensor placement. In *Proceedings of the seventeenth annual symposium on Computational geometry*, SCG '01, pages 232–240, New York, NY, USA, 2001. ACM. 2.4

[65] Stephen Gould, Joakim Arfvidsson, Adrian Kaehler, Benjamin Sapp, Marius Meissner, Gary Bradski, Paul Baumstarck, Sukwon Chung, and Andrew Y. Ng. Peripheral-foveal vision for real-time object recognition and tracking in video. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*, 2007. 5.2

[66] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*, 2004. 6.3.2

[67] Michael Greiffenhagen, Visvanathan Ramesh, Dorin Comaniciu, and Heinrich Niemann. Statistical modeling and performance characterization of a real-time dual camera surveillance system. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2335–2342. IEEE Computer Society, 2000. 2.2, 2.5

[68] Ben Grocholsky. *Information-theoretic control of multiple sensor platforms*. PhD thesis, The University of Sydney, 2002. 2.5, 3.4.2, 3.5

[69] Ben Grocholsky, James Keller, R. Vijay Kumar, and George J. Pappas. Cooperative air and ground surveillance. *IEEE Robotics and Automation Magazine*, 13(3):16–25, September 2006. 1.2, 3.4.2, 6.6.1

[70] Erdan Gu, Jingbin Wang, and Norman I. Badler. Generating sequence of eye fixations using decision-theoretic attention model. In *Proceedings of WAPCV 2007*, volume 4840 of *LNAI*, pages 277–292. Springer, 2007. 5.2

[71] D. Hall, J. Nascimento, P. Ribeiro, E. Andrade, P. Moreno, S. Pesnel, T. List, R. Emonet, R. B. Fisher, Santos J. Victor, and J. L. Crowley. Comparison of target detection algorithms using adaptive background models. In *International workshop on Performance evaluation of Tracking and Surveillance*, 2005. 3.9.1

[72] Arun Hampapur, Sharat Pankanti, Andrew Senior, Ying-Li Tian, Lisa Brown, and Ruud Bolle. Face cataloger: Multi-scale imaging for relating identity to location. In *AVSS '03: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 13 – 20, Washington, DC, USA, 2003. IEEE Computer Society. 2.2, 2.3, 2.5, 5.2, 7.2

[73] M. Hernández, J. Cabrera, M. Castrillón, A. Domínguez, C. Guerra, D. Hernández, and J. Isern. Deseo: An active vision system for detection tracking and recognition. In *Computer Vision Systems*, volume 1542 of *Lecture Notes in Computer Science*, pages 376–391. Springer Berlin / Heidelberg, 1999. 10.1007/3-540-49256-9_23. 1.2

[74] Gabriel M. Hoffmann and Claire J. Tomlin. Mobile sensor network control using mutual information methods and particle filters. *IEEE Transactions on Automatic Control*, 55:32–47, 2010. 1.2, 2.5, 8.3.1

[75] Geoffrey Hollinger, Sanjiv Singh, and Joseph Djugash. Efficient multi-robot search for a moving target. *International Journal of Robotics Research*, 28(2):201–219, February 2009. Graph-based approach to sensor placement, based on Krause and Singh's joint work. 5.2

[76] Marco Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe Hanebeck. On entropy approximation for gaussian mixture random vectors. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2008. 8.3.1

[77] Intel. OpenCV - open source computer vision library 1.0. URL: http://www.intel.com/research/mrl/research/opencv/, 2006. 4.5.1

[78] V. Isler, S. Khanna, J. Spletzer, and C.J. Taylor. Target tracking with distributed sensors: The focus of attention problem. *Computer Vision and Image Understanding Journal*, 100(1-2):225–247, October-November 2005. Special Issue on Attention and Performance in Computer Vision. 5.2, 7.2

[79] O. L. R. Jacobs. *Introduction to Control Theory, Second Edition*. Oxford University Press, Oxford, UK, 1996. 1.3, 3.4.2, 4.5

[80] Neil Johnson and David Hogg. Learning the distribution of object trajectories for event recognition. In *BMVC '95: Proceedings of the 6th British conference on Machine vision (Vol. 2)*, pages 583–592, Surrey, UK, UK, 1995. BMVA Press. 5.4.3, 8.3.6

[81] Edward K. Kao, Matthew P. Daggett, and Michael B. Hurley. An information theoretic approach for tracker performance evaluation. In *Proceedings of the 12th International Conference on Computer Vision (ICCV)*, pages 1523–1529, Kyoto, September 2009. 3.8

[82] Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc, New York, 1976. 3.6

[83] Zia Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1805–1819, 2005. 8.3.1

[84] Thomas Kollar and Nicholas Roy. Trajectory optimization using reinforcement learning for map exploration. *Int. J. Rob. Res.*, 27:175–196, February 2008. 8.3.6

[85] Nils Krahnstoever, Ting Yu, Ser-Nam Lima, and Kedar Patwardhana. Collaborative control of active cameras in large-scale surveillance. In Hamid Aghajan and Andrea Cavallaro, editors, *Multi-Camera Networks: Principles and Applications*, pages 165–188. Elsevier Science Inc., 2009. 1.2, 2.2, 2.5, 5.1, 7.6

[86] Andreas Krause and Carlos Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 324–331, Arlington, Virginia, 2005. AUAI Press. 3.3.6

[87] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. *Machine Learning Research*, 9:235–284, 2008. 2.4, 5.2, 5.2, A.2

[88] V. Krishnamurthy and D. V. Djonin. Structured threshold policies for dynamic sensor scheduling – a partially observed markov decision process approach. *Signal Processing, IEEE Transactions on*, 55(10):4938–4957, 2007. 3.4.2

[89] Daniel Kuettel, Michael D. Breitenstein, Luc Van Gool, and Vittorio Ferrari. What's going on? Discovering spatio-temporal dependencies in dynamic scenes. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2010. 8.3.4

[90] Bastian Leibe, Aleš Leonardis, and Bernt Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision Special Issue on Learning for Recognition and Recognition for Learning*, 77(1-3):259–289, 2007. 2

[91] Xiaoming Liu, Nils Krahnstoever, Ting Yu, and Peter Tu. What are customers looking at? In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 405 –410, 5-7 2007. 7.2

[92] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 10.1023/B:VISI.0000029664.99615.94. 7.3.2

[93] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In Patrick J. Hayes, editor, *IJCAI*, pages 674–679. William Kaufmann, 1981. 7.7.2

[94] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. 3.1, 3.3.3

[95] D. Makris and T. Ellis. Automatic learning of an activity-based semantic scene model. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 183–188, 2003. 5.4.3

[96] Dimitrios Makris and Tim Ellis. Spatial and probabilistic modelling of pedestrian behaviour. In *Proceedings of the 13th British Machine Vision Conference*, pages 557–566, Cardiff, September 2002. 5.4.3

[97] Davide Maltoni, Dario Maio, Anil K. Jain, and Salil Prabhakar. *Handbook of Fingerprint Recognition*. Springer Publishing Company, Incorporated, 2nd edition, 2009. 7.5

[98] James Manyika and Hugh Durrant-Whyte. *Data Fusion and Sensor Management a decentralized information-theoretic approach*. Electrical and Electronic Engineering. Ellis Horwood, Chichester, UK, 1994. 1.2, 3.4, 3.5, 3.6, 6.6.1

[99] L. Marchesotti, L. Marcenaro, and C. Regazzoni. Dual camera system for face detection in unconstrained environments. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I – 681–4 vol.1, sep. 2003. 7.2

[100] Stephen J. McKenna and Hammadi Nait-Charif. Summarising contextual activity and detecting unusual inactivity in a supportive home environment. *Pattern Analysis & Applications*, 7:386–401, 2004. 10.1007/s10044-004-0233-2. 5.4.3

[101] R. Mehra. On-line identification of linear dynamic systems with applications to Kalman filtering. *IEEE Transactions on Automatic Control*, 16(1):12–21, February 1971. 4.4.2

[102] Hans P. Moravec and Alberto Elfes. High resolution maps from wide angle sonar. In *Proceedings of the 1985 IEEE International Conference on Robotics and Automation*, pages 116–121, March 1985. 5.2

[103] H.-H. Nagel and I. Reid. Exploring situation-graph-trees for conceptual control of a pan-tilt-zoom-camera. Technical report, Institut für Algorithmen und Kognitive Systeme (IAKS), Fakultät für Informatik der Universität Karlsruhe (TH), Karlruhe, Germany and Active Vision Laboratory (AVL), Department of Engineering Science, University of Oxford, UK, 2008. 8.3.7

[104] Chandra Nair, Balaji Prabhakar, and Devavrat Shah. On entropy for mixtures of discrete and continuous variables. *CoRR*, abs/cs/0607075:1–8, 2006. 3.3.5

[105] Michael D. Naish, Elizabeth A. Croft, and Beno Benhabib. Object surveillance using reinforcement learning based sensor dispatching. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation (ICRA)*, 2004. 5.2, 8.3.6

[106] G. W. Ng and K. H. Ng. Sensor management - what, why and how. *Information Fusion*, 1(2):67 – 75, 2000. 1.2, 2.2

[107] D. W. North. A tutorial introduction to decision theory. *Systems Science and Cybernetics, IEEE Transactions on*, 4(3):200–210, 1968. 3.4

[108] Juhyun Oh, Seungjin Nam, and Kwanghoon Sohn. Practical pan-tilt-zoom-focus camera calibration for augmented reality. In Mario Fritz, Bernt Schiele, and Justus Piater, editors, *Computer Vision Systems*, volume 5815 of *Lecture Notes in Computer Science*, pages 225–234. Springer Berlin / Heidelberg, 2009. 4.5.1

[109] Joseph O'Rourke. *Computational Geometry in C (Cambridge Tracts in Theoretical Computer Science)*. Cambridge University Press, October 1998. 2.4

[110] Javier Orozco, Shaogang Gong, and Tao Xiang. Head pose classification in crowded scenes. In *Proceedings of the British Machine Vision Conference*, pages 1–11, 2009. 7.6

[111] Michael A. Osborne. *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. PhD thesis, University of Oxford, 2010. 2.4

[112] Lucas Paletta and Gerald Fritz. Reinforcement learning for decision making in sequential visual attention. In *Proceedings of WAPCV 2007*, volume 4840 of *LNAI*, pages 277–292. Springer, 2007. 5.2

[113] R. Pappu and P.A. Beardsley. A qualitative approach to classifying gaze direction. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 160 –165, 14-16 1998. 7.2

[114] Johnny Park, Priya C. Bhat, and Avinash C. Kak. A look-up table based approach for solving the camera selection problem in large camera networks. In *Workshop on Distributed Smart Cameras, in conjunction with ACM SenSys'06*, 2006. 5.2

[115] Alonso Patron, Eric Sommerlade, and Ian Reid. Action recognition using shared motion parts. In *Proceedings of the Eighth International Workshop on Visual Surveillance 2008*, October 2008. 1.6

[116] P. Peixoto, J. Batista, and H. Araujo. A surveillance system combining peripheral and foveated motion tracking. *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, 1:574–577 vol.1, Aug 1998. 2.2

[117] P.J. Phillips, W.T. Scruggs, A.J. O'Toole, P.J. Flynn, K.W. Bowyer, C.L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):831 –846, May 2010. 1.1, 3.8, 5.1, 7.1

[118] R. Pito. A solution to the next best view problem for automated surface acquisition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(10):1016 –1030, October 1999. 2.4

[119] Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2007. 3.4.2

[120] Victor Prisacariu and Ian Reid. fastHOG - a real-time GPU implementation of HOG. Technical Report 2310/09, Department of Engineering Science, University of Oxford, 2009. 3.9.2, 3.2

[121] Faisal Z. Qureshi and Demetri Terzopoulos. Surveillance in virtual reality: System design and multi-camera control. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2.2, 2.5, 5.2, 5.4.6, 7.2, 7.4

[122] Faizal Z. Qureshi and Demetri Terzopoulos. Planning ahead for PTZ camera assignment and control. In *Proc. Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 09)*, pages 1–8, Como, Italy, August 2009. 7.2, 7.6, 8.3.5

[123] Donald B. Reid. An algorithm for tracking multiple targets. *Transactions on Automatic Control*, 24(6):843–854, December 1979. 8.3.1

[124] Daniel Roth, Petr Doubek, and Luc Van Gool. Bayesian pixel classification for human tracking. In *WACV/MOTION 2005*, January 2005. 4.5.2

[125] Allison Ryan and J. Karl Hedrick. Particle filter based information-theoretic active sensing. *Robotics and Autonomous Systems*, 58(5):574 – 584, 2010. 2.5, 5.2, 8.3.1

[126] A.W. Senior, A. Hampapur, and M. Lu. Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration. In *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on*, volume 1, pages 433–438, January 2005. 2.3, 2.5

[127] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, July and October 1948. 3.3.1

[128] Jamie Sherrah, Shaogang Gong, A. Jonathan Howell, and Hilary Buxton. Interpretation of group behaviour in visually mediated interaction. In *In International Conference on Pattern Recognition, ICPR'2000*, pages 266–269, 2000. 7.2

[129] G. Sibley, G. Sukhatme, and L. Matthies. The Iterated Sigma Point Kalman Filter with Applications to Long Range Stereo. In *Proceedings of Robotics: Science and Systems*, Philadelphia, USA, August 2006. 3.7

[130] Robert Sim. Stable Exploration for Bearings-only SLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2422–2427, Barcelona, Spain, April 2005. IEEE, IEEE Press. 3.7

[131] Amarjeet Singh, Andreas Krause, Carlos Guestrin, and William J. Kaiser. Efficient informative sensing using multiple robots. *J. Artif. Int. Res.*, 34:707–755, April 2009. 2.5

[132] Kevin Smith, Sileye O. Ba, Daniel Gatica-Perez, and Jean-Marc Odobez. Tracking the multi person wandering visual focus of attention. In *ICMI '06: Proceedings of the 8th International Conference on Multimodal interfaces*, pages 265–272, New York, NY, USA, 2006. ACM Press. 7.2

[133] Paul Smith, Mubarak Shah, and Niels da Vitoria Lobo. Integrating multiple levels of zoom to enable activity analysis. *Computer Vision and Image Understanding*, 103(1):33–51, July 2006. 2.3

[134] Eric Sommerlade and Ian Reid. Cooperative surveillance of multiple targets using mutual information. In *Proceedings of the ECCV Workshop on Multi-camera and Multimodal Sensor Fusion Algorithms and Applications (M2SFA2)*, October 2008. 1.5, 1.6, 6

[135] Eric Sommerlade and Ian Reid. Influence of zoom selection on a kalman filter. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, September 2008. 1.5, 1.6, 4

[136] Eric Sommerlade and Ian Reid. Information theoretic active scene exploration. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2008. 1.5, 1.6, 5

[137] Eric Sommerlade and Ian Reid. Information-theoretic decision making for exploration of dynamic scenes. In *Proceedings of the 5th International Workshop on Attention in Cognitive Systems (WAPCV)*, May 2008. 1.5, 1.6, 5

[138] Eric Sommerlade and Ian Reid. Probabilistic surveillance with multiple active cameras. In *IEEE International Conference on Robotics and Automation*, pages 440–445, May 2010. 1.5, 1.6, 6

[139] Eric Sommerlade and Ian Reid. Gaze directed camera control for face image acquisition. In *Proceedings of the International Conference on Robotics and Automation*, 2011. 1.5, 1.6, 7

[140] Bi Song, Ahmed T. Kamal, Cristian Soto, Chong Ding, Jay A. Farrell, and Amit K. Roy-Chowdhury. Tracking and activity recognition through consensus in distributed camera networks. *Trans. Img. Proc.*, 19:2564–2579, October 2010. 2.2, 6.6.1

[141] Bi Song and A.K. Roy-Chowdhury. Robust tracking in a camera network: A multi-objective optimization framework. *Selected Topics in Signal Processing, IEEE Journal of*, 2(4):582 –596, August 2008. 2.2

[142] Stanislava Soro and Wendi Heinzelman. A survey of visual sensor networks. *Advances in Multimedia Systems*, 2009:1–21, 2009. 2.2

[143] Cristian Soto, Bi Song, and Amit K. Roy Chowdhury. Distributed multi-target tracking in a self-configuring camera network. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1486–1493. IEEE, 2009. 1.2, 2.2, 2.3, 2.5

[144] Matthijs T. J. Spaan, Nelson Gonçalves, and João Sequeira. Multirobot coordination by auctioning POMDPs. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1446–1451, 2010. 3.4.2

[145] John R. Spletzer and Camillo J. Taylor. Dynamic sensor planning and control for optimally tracking targets. *International Journal of Robotics Research*, 22:7–20, 2003. 2.5

[146] C. Stachniss, G. Grisetti, and W. Burgard. Information gain-based exploration using rao-blackwellized particle filters. In *Proc. of Robotics: Science and Systems (RSS)*, Cambridge, MA, USA, 2005. 8.3.1

[147] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000. 5.4.3

[148] Rainer Stiefelhagen. Tracking focus of attention in meetings. In *ICMI '02: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, page 273, Washington, DC, USA, 2002. IEEE Computer Society. 7.2

[149] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998. 3.4.2

[150] Noriko Takemura and Jun Miura. View planning of multiple active cameras for wide area surveillance. In *Proceedings of the 2007 IEEE Int. Conf. on Robotics and Automation*, 2007. 5.2

[151] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. Intelligent robotics and autonomous agents. MIT Press, September 2005. 1.2, 3.4.2

[152] John Tisdale, Allison Ryan, Zu Kim Kim, D. Tornqvist, and Karl Hedrick. A multiple uav system for vision-based search and localization. In *Proceedings of the IEEE American Controls Conference*, 2008. 5.2

[153] B. Tordoff and D. Murray. Reactive control of zoom while fixating using perspective and affine cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):98–112, January 2004. 4.2

[154] B.J. Tordoff and D.W. Murray. A method of reactive zoom control from uncertainty in tracking. *Computer Vision and Image Understanding*, 105:131–144, 2007. 4.2, 4.3, 4.3.2, 4.4

[155] O. Trémois and J.-P. Le Cadre. Optimal observer trajectory in bearings-only tracking for maneuvering sources. *IEE Radar, Sonar and Navigation*, 146(1):31–39, February 1999. 3.4.2

[156] R. van der Merwe. *Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models*. PhD thesis, OGI School of Science & Engineering, Oregon Health & Science University, Portland, OR, USA, April 2004. 3.7

[157] R. van der Merwe and E. Wan. Sigma-point kalman filters for probabilistic inference in dynamic state-space models. In *Proceedings of the Workshop on Advances in Machine Learning*, Montreal, Canada, June 2003. 3.7

[158] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*, 2001. 3.9.2, 3.2

[159] David W. Walkup and M. Donald MacLaren. A multiple-assignment problem. Technical Report D1-82-0346, Boeing Scientific Research Laboratories, http://handle.dtic.mil/100.2/AD602672, April 1964. Mathematical Note. 7.2

[160] Yi Yao, Chung-Hao Chen, David Page, Besma Abidi, Andreas Koschan, and Mongi Abidi. Sensor planning for automated and persistent object tracking with multiple cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2008. 5.2

[161] Yi Yao, Chung hao Chen, Besma Abidi, Andreas Koschan, and Mongi Abidi. Sensor planning for ptz cameras using the probability of camera overload. In *Proceedings of the 19th International Conference on Pattern Recognition (ICPR)*, 2008. 5.2, 5.2

[162] Fei Yin, D. Makris, and S. A. Velastin. Performance evaluation of object tracking algorithms. In *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2007), Rio de Janeiro, Brazil*, October 2007. 3.8

[163] Xenophon Zabulis, Thomas Sarmis, and Antonis Argyros. 3d head pose estimation from multiple distant views. In *Proceedings of the British Machine Vision Conference*, 2009. 7.6

[164] Yongmian Zhang and Qiang Ji. Active and dynamic information fusion for multisensor systems with dynamic bayesian networks. *Systems, Man, and Cybernetics, Part B, IEEE Transactions on*, 36(2):467–472, April 2006. 5.2

[165] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 4.5.1

[166] Chuan Zhao. *Towards Efficient Surveillance Using Visual Tracking and Zoom Control*. PhD thesis, University of Oxford, 2009. 3.10