Linköping Studies in Science and Technology
Dissertation No. 1395

Shape Based Recognition

Cognitive Vision Systems in Traffic Safety Applications

Fredrik Larsson

Linköping University
INSTITUTE OF TECHNOLOGY

Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden

Linköping November 2011

Shape Based Recognition - Cognitive Vision Systems in Traffic Safety Applications

Department of Electrical Engineering
Linköping University
SE-581 83 Linköping
Sweden

ISBN: 978-91-7393-074-1 ISSN 0345-7524

Linköping Studies in Science and Technology
Dissertation No. 1395

Abstract

Traffic accidents are globally the number one cause of death for people 15-29 years old and is among the top three causes for all age groups 5-44 years. Much of the work within this thesis has been carried out in projects aiming for (cognitive) driver assistance systems and hopefully represents a step towards improving traffic safety.

The main contributions are within the area of Computer Vision, and more specifically, within the areas of *shape matching*, *Bayesian tracking*, and *visual servoing* with the main focus being on shape matching and applications thereof. The different methods have been demonstrated in traffic safety applications, such as bicycle tracking, car tracking, and traffic sign recognition, as well as for pose estimation and robot control.

One of the core contributions is a new method for recognizing closed contours, based on complex correlation of Fourier descriptors. It is shown that keeping the phase of Fourier descriptors is important. Neglecting the phase can result in perfect matches between intrinsically different shapes. Another benefit of keeping the phase is that rotation covariant or invariant matching is achieved in the same way. The only difference is to either consider the magnitude, for rotation invariant matching, or just the real value, for rotation covariant matching, of the complex valued correlation.

The shape matching method has further been used in combination with an implicit star-shaped object model for traffic sign recognition. The presented method works fully automatically on query images with no need for regions-of-interests. It is shown that the presented method performs well for traffic signs that contain multiple distinct contours, while some improvement still is needed for signs defined by a single contour. The presented methodology is general enough to be used for arbitrary objects, as long as they can be defined by a number of regions.

Another contribution has been the extension of a framework for learning based Bayesian tracking called *channel based tracking*. Compared to earlier work, the multi-dimensional case has been reformulated in a sound probabilistic way and the learning algorithm itself has been extended. The framework is evaluated in car tracking scenarios and is shown to give competitive tracking performance, compared to standard approaches, but with the advantage of being fully learnable.

The last contribution has been in the field of (cognitive) robot control. The presented method achieves sufficient accuracy for simple assembly tasks by combining autonomous recognition with visual servoing, based on a learned mapping between percepts and actions. The method demonstrates that limitations of inexpensive hardware, such as web cameras and low-cost robotic arms, can be overcome using powerful algorithms.

All in all, the methods developed and presented in this thesis can all be used for different components in a system guided by visual information, and hopefully represents a step towards improving traffic safety.

Populärvetenskaplig sammanfattning

Trafikolyckor är globalt sett den vanligaste dödsorsaken för människor i åldrarna 15-29 år och är bland de tre vanligaste dödsorsakerna för alla åldersgrupper 5-44 år. En stor del av arbetet, som har lett fram till denna avhandling, har skett inom projekt med fokus på system för att hjälpa bilförare. Förhoppningsvis bidrar de resultat som presenteras till förbättrad trafiksäkerhet och i förlängningen även till räddade människoliv.

Avhandlingen beskriver metoder och algoritmer inom ämnet datorseende. Datorseende är en ingenjörsvetenskap som har som mål att skapa seende maskiner, vilket i praktiken innebär utveckling av algoritmer och datorprogram som kan extrahera och använda information från bilder.

För att vara mer specifik så innehåller denna avhandling metoder inom delområdena formigenkänning, målföljning och visuellt återkopplad styrning. De olika metoderna har framförallt demonstrerats i tillämpningar med anknytning till trafiksäkerhet, så som trafikskyltsigenkänning och följning av bilar, men också inom andra områden, bland annat för att styra mekaniska robotarmar.

Tyngdpunkten hos avhandlingen ligger inom området formigenkänning. Formigenkänning syftar till att automatiskt kunna identifiera och känna igen olika geometriska former trots försvårande omständigheter, som rotation, skalning och deformation. Ett av huvudresultaten är en metod för att känna igen former genom att betrakta ytterkonturer. Denna metod är baserad på korrelation av så kallade Fourier-deskriptorer och har använts för detektion och igenkänning av trafikskyltar. Metoden bygger på att känna igen delregioner hos skyltar var för sig och sedan kombinera dessa med krav på inbördes geometriska förhållanden. Formigenkänning har tillsammans med målföljning även använts för att detektera och följa cyklister i videosekvenser, genom att känna igen cykelhjul vilka avbildas som ellipser i bildplanet.

Inom området målföljning presenteras en vidareutveckling av tidigare arbeten inom så kallad kanalbaserad målföljning. Målföljning handlar om att noggrant uppskatta tillstånd, till exempel position och hastighet, hos objekt. Detta görs genom att använda observationer från olika tidpunkter tillsammans med rörelse- och observationsmodeller. Den metod som presenteras har använts i en bil för att följa positionen hos andra bilister, vilket i slutändan används för att varna föraren vid potentiella faror.

Det sista delområde som berörs handlar om styrning av robotar med hjälp av återkopplad visuell information. Avhandlingen innehåller en metod inspirerad av hur vi människor lär oss att använda vår kroppar redan i fosterstadiet. Metoden bygger på att i ett första skede skicka slumpmässiga kontrollsignaler till roboten, vilket resulterar i slumpmässiga rörelser, och sedan observera resultatet. Genom att göra detta upprepade gånger kan den omvända relationen skapas, som kan användas för att välja de kontrollsignaler som krävs för att uppnå en önskad konfiguration

Tillsamman utgör de presenterade metoderna olika komponenter som kan användas i system som använder visuell information, ej begränsade till de tillämpningar som beskrivs ovan.

Acknowledgments

I would like to thank all current and former members of the Computer Vision Laboratory. You have all in one way or another contributed to this thesis, either scientifically or, equally important, by contributing to the friendly and inspiring atmosphere. Especially I would like to thank:

- Michael Felsberg for providing an excellent working environment, for being an excellent supervisor, and a never ending source of inspiration.
- Per-Erik Forssén for being an equally good co-supervisor and for sharing lots of knowledge regarding object recognition, conics, and local features.
- Gösta Granlund for initially allowing me to join the CVL group and for sharing knowledge and inspiration regarding biological vision systems.
- Johan Wiklund for keeping the computers reasonably happy most of the time and for acknowledging the usefulness of gaffer tape.
- Liam Ellis, Per-Erik Forssén, Klas Nordberg and Marcus Wallenberg for proofreading parts of this manuscript and giving much appreciated feedback.

Also I would like to thank all friends and my family for support with non-scientific issues, most notably:

- My parents Ingrid and Kjell for infinite love and for always being there, your love and support means the world to me.
- Marie Knutsson for lots of love and much needed distractions, your presence in my life makes it richer on all levels.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 215078 DIPLECS, from the European Community's Sixth Framework Programme (FP6/2003-2007) under grant agreement n° 004176 COSPAL and from the project Extended Target Tracking funded by the Swedish research council, all which are hereby gratefully acknowledged.

Fredrik Larsson November 2011

Contents

Chapter 1

Introduction

1.1 Motivation

Road and traffic safety is an ever important topic of concern. About 50 million people are injured and more than 1.2 million people die in traffic related accidents every year, which is more than one person dying every 30 seconds. Road traffic injuries are globally the number one cause of death for people 15-29 years old and is among the top three causes for all age groups 5-44 years [85].

The United Nations General Assembly has proclaimed the period 2011-2020 as the *Decade of Action for Road Safety* with a goal to first stabilize and then to reduce the number of traffic fatalities around the world [84]. The number of yearly fatalities is expected to raise to 1.9 million around 2020 and to 2.4 million around 2030 unless the trend is changed [85].

Among the actions stipulated are the tasks of designing safer roads, reducing drunk driving and speeding, and to improve driver training and licensing, also the responsibility of vehicle manufacturers to produce safe cars is mentioned [16].

Much of the work within this thesis has been performed in projects aiming for (cognitive) driver assistance systems and hopefully represents a step towards improving traffic safety.

The main technical contributions of this thesis are within the area of Computer Vision, and more specifically, within the areas of *shape matching*, *Bayesian tracking* and *visual servoing* with the main focus being on shape matching and applications thereof. The different methods have been demonstrated in traffic safety applications, such as bicycle tracking, car tracking, and traffic sign recognition, as well as for pose estimation and robot control.

Work leading to this thesis has mostly been carried out within three projects. The main parts originate from research within two European projects, COSPAL (COgnitive Systems using Perception-Action-Learning [1]) and DIPLECS (Dynamic-Interactive Perception-Action LEarning Systems [2]), while some of the latest contributions stem from the project ETT (Extended Target Tracking) funded by the Swedish research council, see Sec. 1.3 for more details on the projects.

1.2 Outline

This thesis is written as a collection of previously published papers and is divided into two main parts in addition to this introduction. The rest of this introductory chapter contains brief information about the included publications together with explicit statements of the contributions made by the author, followed by a section describing the different projects that the work was carried out within. Part I contains chapters on background theory and concepts needed for Part II, and a concluding chapter. Part II contains the six included papers which make up the core of this thesis.

1.2.1 Outline Part I: Background Theory

Each of the main topics of the thesis, *shape matching*, *Bayesian tracking* and *visual servoing* are given one introductory chapter, covering the basics within these fields. Part I ends with a concluding chapter that summarizes the main results of the thesis and briefly discusses possible areas of future research. Part of the material in Part I has previously been published in [55].

1.2.2 Outline Part II: Included Publications

Edited versions of six papers are included in Part II. The included papers are selected in order to reflect the different areas of research that was touched upon by the author during the years as a Ph.D. student at the Computer Vision Laboratory at Linköping University.

Paper A contains work on relative pose estimation using a torch light. The reprojection of the emitted light beam creates, under certain conditions, an ellipse in the image plane. We show that it is possible to use this ellipse in order to estimate the relative pose.

Paper B builds on the ideas presented in paper A and contains initial work on bicycle tracking, done jointly with the Automatic Control group at Linköping University. The relative pose estimates are based on ellipses originating from the projection of the bicycle wheels into the image. The different ellipses have to be associated to the correct ellipses in previous frames, i.e. front wheel to front wheel and rear wheel to rear wheel. This is combined with a particle filter framework in order to track the bicycle in 3D.

Paper C contains work on generic shape recognition using Fourier descriptors, while papers A and B only deal with ellipses. The paper presents theoretical justifications for using a correlation based matching scheme for Fourier descriptors and also presents initial work on traffic sign recognition.

Paper D extends the work on traffic sign recognition by introducing spatial constraints on the local shapes using an implicit star-shaped object model. The earlier paper C focus on recognizing individual shapes while this work takes the configuration of different shapes into consideration.

Paper E contains work on learning based object tracking. In Paper B the motion model of the tracked object is known beforehand. This is not always the

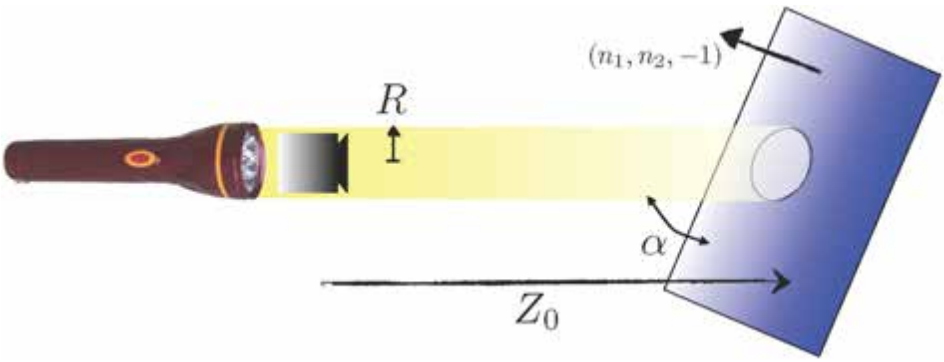
case and the method presented in paper E addresses this scenario. The approach is evaluated in car tracking experiments.

Paper F describes a method for learning how to control a robotic arm without knowing beforehand what it looks like or how it is controlled. In order for the method presented in this paper to work, consistent estimates of the robot configuration/pose are needed. This is achieved by a heuristic approach based on template matching but could (preferably) be replaced using the tracking framework from papers B and E in combination with the shape and pose estimation ideas from papers A-D.

Bibliographic details for each of the included papers together with abstracts and statements of the contributions made by the author are given in this section.

Paper A: Torchlight Navigation

> M. Felsberg, F. Larsson, W. Han, A. Ynnerman, and T. Schön. Torchlight navigation. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, 2010. **This work received a paper award from the Swedish Society for Automated Image Analysis.**



Abstract: A common computer vision task is navigation and mapping. Many indoor navigation tasks require depth knowledge of flat, unstructured surfaces (walls, floor, ceiling). With passive illumination only, this is an ill-posed problem. Inspired by small children using a torchlight, we use a spotlight for active illumination. Using our torchlight approach, depth and orientation estimation of unstructured, flat surfaces boils down to estimation of ellipse parameters. The extraction of ellipses is very robust and requires little computational effort.

Contributions: The author was the main source for implementing the method, conducting the experiments and writing large parts of the paper. The original idea was developed by Felsberg, Han, Ynnerman and Schön.

Paper B: Bicycle Tracking Using Ellipse Extraction

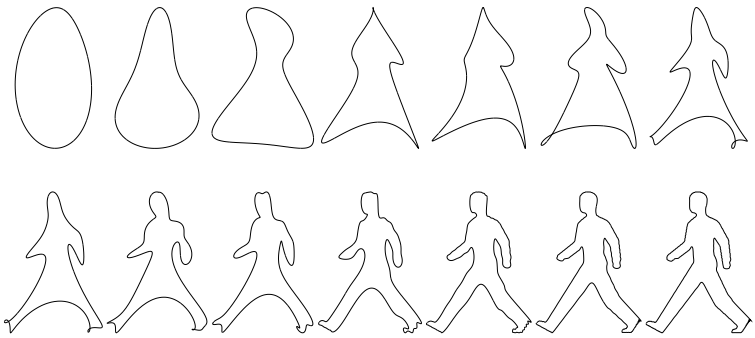
T. Ardeshiri, F. Larsson, F. Gustafsson, T. Schön, and M. Felsberg. Bicycle tracking using ellipse extraction. In *Proceedings of the 14th International Conference on Information Fusion*, 2011. **Honorable mention, nominated for the best student paper award**

Abstract: A new approach to track bicycles from imagery sensor data is proposed. It is based on detecting ellipsoids in the images, and treat these pair-wise using a dynamic bicycle model. One important application area is in automotive collision avoidance systems, where no dedicated systems for bicyclists yet exist and where very few theoretical studies have been published. Possible conflicts can be predicted from the position and velocity state in the model, but also from the steering wheel articulation and roll angle that indicate yaw changes before the velocity vector changes. An algorithm is proposed which consists of an ellipsoid detection and estimation algorithm and a particle filter. A simulation study of three critical single target scenarios is presented, and the algorithm is shown to produce excellent state estimates. An experiment using a stationary camera and the particle filter for state estimation is performed and has shown encouraging results.

Contributions: The author was the main source behind the computer vision related parts of this paper while Ardeshiri was the main source behind the parts related to control theory. The author implemented the method for ellipse estimation and wrote parts of the paper.

Paper C: Correlating Fourier Descriptors of Local Patches for Road Sign Recognition

F. Larsson, M. Felsberg, and P.-E. Forssén. Correlating Fourier descriptors of local patches for road sign recognition. *IET Computer Vision*, 5(4):244–254, 2011.



Abstract: The Fourier descriptors (FDs) is a classical but still popular method for contour matching. The key idea is to apply the Fourier transform to a periodic representation of the contour, which results in a shape descriptor in the frequency domain. Fourier descriptors are most commonly used to compare object silhouettes and object contours; we instead use this well established machinery to describe local regions to be used in an object recognition framework. Many approaches to matching FDs are based on the magnitude of each FD component, thus ignoring the information contained in the phase. Keeping the phase information requires us to take into account the global rotation of the contour and shifting of the contour samples. We show that the sum-of-squared differences of FDs can be computed without explicitly de-rotating the contours. We compare our correlation based matching against affine-invariant Fourier descriptors (AFDs) and WARP matched FDs and demonstrate that our correlation based approach outperforms AFDs and WARP on real data. As a practical application we demonstrate the proposed correlation based matching on a road sign recognition task.

Contributions: The author is the main source behind the research leading to this paper. The author developed and implemented the method and wrote the paper. Initial inspiration and ideas originated from Forssén and Felsberg, with Felsberg also contributing to the presented matching scheme.

Paper D: Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition

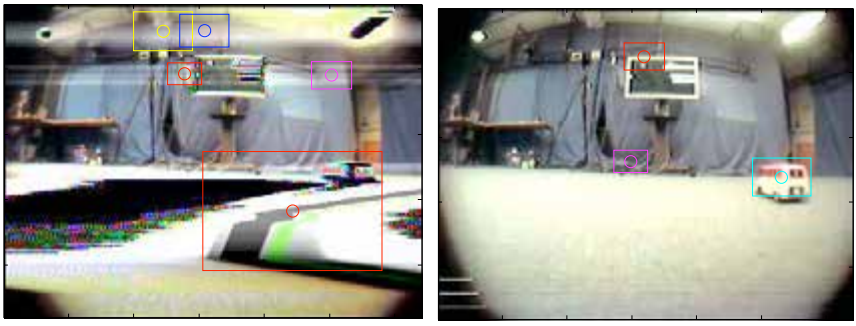
F. Larsson and M. Felsberg. Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition. In *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, volume 6688 of *Lecture Notes in Computer Science*, pages 238–249, 2011.

Abstract: Traffic sign recognition is important for the development of driver assistance systems and fully autonomous vehicles. Even though GPS navigator systems works well for most of the time, there will always be situations when they fail. In these cases, robust vision based systems are required. Traffic signs are designed to have distinct colored fields separated by sharp boundaries. We propose to use locally segmented contours combined with an implicit star-shaped object model as prototypes for the different sign classes. The contours are described by Fourier descriptors. Matching of a query image to the sign prototype database is done by exhaustive search. This is done efficiently by using the correlation based matching scheme for Fourier descriptors and a fast cascaded matching scheme for enforcing the spatial requirements. We demonstrated on a publicly available database state of the art performance.

Contributions: The author is the main source behind the research leading to this paper. The author developed and implemented the method and wrote the main part of the paper.

Paper E: Learning Higher-Order Markov Models for Object Tracking in Image Sequences

M. Felsberg and F. Larsson. Learning higher-order Markov models for object tracking in image sequences. In *Proceedings of the International Symposium on Visual Computing (ISVC)*, volume 5876 of *Lecture Notes in Computer Science*, pages 184–195. Springer-Verlag, 2009.

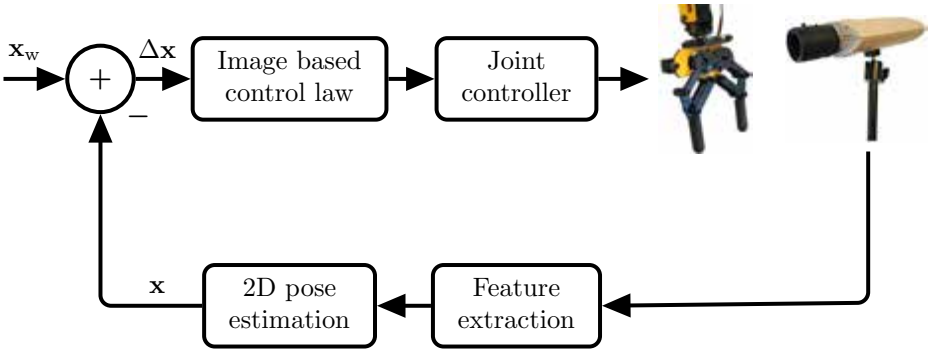


Abstract: This work presents a novel object tracking approach, where the motion model is learned from sets of frame-wise detections with unknown associations. We employ a higher-order Markov model on position space instead of a first-order Markov model on a high-dimensional state-space of object dynamics. Compared to the latter, our approach allows the use of marginal rather than joint distributions, which results in a significant reduction of computation complexity. Densities are represented using a grid-based approach, where the rectangular windows are replaced with estimated smooth Parzen windows sampled at the grid points. This method performs as accurately as particle filter methods with the additional advantage that the prediction and update steps can be learned from empirical data. Our method is compared against standard techniques on image sequences obtained from an RC car following scenario. We show that our approach performs best in most of the sequences. Other potential applications are surveillance from cheap or uncalibrated cameras and image sequence analysis.

Contributions: The core ideas behind this paper originates from Felsberg. The author wrote parts of the paper and was the main source for implementing the theoretical findings and for conducting experiments validating the tracking framework.

Paper F: Simultaneously Learning to Recognize and Control a Low-Cost Robotic Arm

F. Larsson, E. Jonsson, and M. Felsberg. Simultaneously learning to recognize and control a low-cost robotic arm. *Image and Vision Computing*, 27(11):1729–1739, 2009



Abstract: In this paper, we present a visual servoing method based on a learned mapping between feature space and control space. Using a suitable recognition algorithm, we present and evaluate a complete method that simultaneously learns the appearance and control of a low-cost robotic arm. The recognition part is trained using an action precedes perception approach. The novelty of this paper, apart from the visual servoing method per se, is the combination of visual servoing with gripper recognition. We show that we can achieve high precision positioning without knowing in advance what the robotic arm looks like or how it is controlled.

Contributions: The author is the main source behind the research leading to this paper. The author developed and implemented the method and wrote the main part of the paper.

1.3 Projects

Most of the research leading to this thesis was conducted within the two European projects COSPAL and DIPLECS. Both projects were within the European Framework Programme calls for cognitive systems and thus had a strong focus on learning based methods able to adapt to the environment. DIPLECS can be seen as the follow up project to COSPAL and was closer to real applications, exemplified by driver assistance, than the previous project. Some of the latest contributions stem from the project ETT funded by the Swedish research council. ETT shares some similarities with DIPLECS such as applications within the traffic safety domain and the use of shape recognition techniques. Additional details about the three projects can be found below.

1.3.1 COSPAL

COSPAL (COgnitive Systems using Perception-Action-Learning[1]) was a European Community's Sixth Framework Programme project carried out between 2004 and 2007 [1]. The main goal of the COSPAL project was to conduct research leading towards systems that learn from experience, rather than using predefined models of the world.

The key concept, as stated in the project name, was to use perception-action-learning. This was achieved by applying the idea of *action-precedes-perception* during the learning phase [39]. Meaning that, the system learns by first performing an action (random or goal directed) and then observing the outcome. By doing so, it is possible to learn the inverse mapping between percept and action. The motivation behind this reversed causal direction is that the action space tends to be of much lower dimensionality than the percept space [39]. This approach was successfully demonstrated in the context of robot control described in the included publication [65].

The main demonstrator scenario of the COSPAL project involved a robotic arm and a shape sorting puzzle, see Fig. 1.1, but the system architecture and algorithms implemented were all designed to be as generic as possible. This was demonstrated in [20] when part of the main COSPAL system successfully was used for two different tasks, solving a shape sorting puzzle and driving a radio

[1] FP6/2003-2007, grant agreement n° 004176

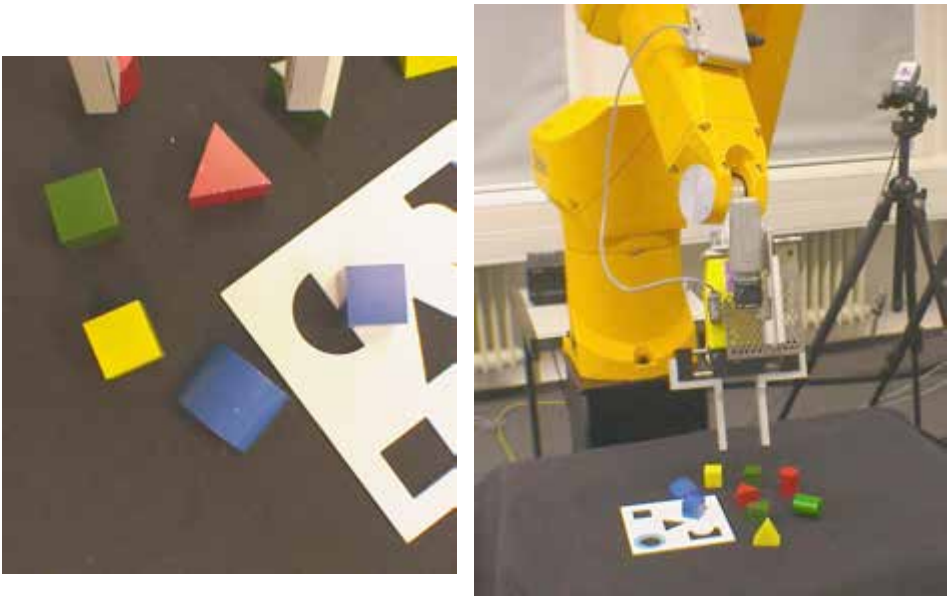


Figure 1.1: Images from the COSPAL main demonstrator. Left: A view captured by the camera mounted on the gripper. Right: Side view of the robotic arm and shape sorting puzzle.

controlled car. The results presented by the author in [62, 63, 64, 65], originate from the COSPAL project.

1.3.2 DIPLECS

DIPLECS (Dynamic-Interactive Perception-Action LEarning Systems[2]) was a European Community's Seventh Framework Programme project carried out between 2007 and 2010 [2]. DIPLECS continued the work of COSPAL and extended the results from COSPAL to incorporate dynamic and interaction with other agents.

The scenarios considered during the COSPAL project involved a single system operating in a static world. This was extended in DIPLECS to allow for a changing world and multiple systems acting simultaneously within the world. The main scenario of the DIPLECS project was driver assistance and one of the core ideas was to learn by observing human drivers, i.e. perception-action learning. The

[2]FP7/2007-2013, grant agreement n° 215078

following project overview is quoted from the DIPLECS webpage.

> 'The DIPLECS project aims to design an Artificial Cognitive System capable of learning and adapting to respond in the everyday situations humans take for granted. The primary demonstration of its capability will be providing assistance and advice to the driver of a car. The system will learn by watching humans, how they act and react while driving, building models of their behaviour and predicting what a driver would do when presented with a specific driving scenario. The end goal of which is to provide a flexible cognitive system architecture demonstrated within the domain of a driver assistance system, thus potentially increasing future road safety.'[2]

The DIPLECS integrated system was demonstrated in a number of different traffic scenarios using a RC-car, see Fig. 1.2, and a real vehicle, see Fig. 1.3. The RC-car allowed for the system to actively control the actions of the vehicle, for tasks such as automatic obstacle avoidance and path following [21, 41, 75], something that due to safety protocols was not done on the real car. The real car

Figure 1.2: The RC-car setup used for active control by the system.

was instrumented with multiple cameras mounted on the roof, on the dashboard facing out and also cameras facing the driver used for eye-tracking. Multiple additional sensors such as gas and break pedal proximity senors, differential GPS were also mounted in the car.

The images from the three roof mounted cameras were stitched into one wide field of view image, see Fig. 1.3. The observed paths of object in the world take on nontrivial properties due to the nonlinear distortions occurring on the stitching boundaries as well as the potential movement of both vehicle and observed object. Methods developed in the included publication on learning tracking models [26] were integrated in the instrumented vehicle in order to address these challenges.

The main demonstrator showed the systems ability to adapt to the behavior of the driver, [30]. One example was the grounding of visual percepts to semantic meaning based on driver actions, demonstrated with traffic signs, see Fig. 1.4 and videos at `www.diplecs.eu`. Originally the system is not aware of the semantic meaning of the detection corresponding to a stop sign. The system is aware that the reported detection is a a sign, just not of what type. After a few runs of stopping at a junction with the sign present, the system deduces that the sign

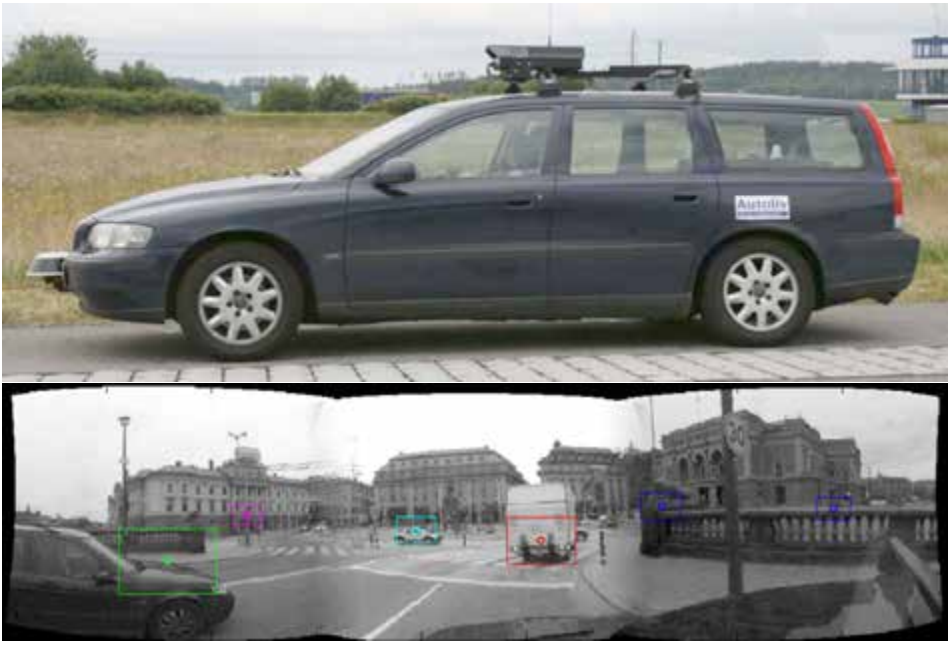


Figure 1.3: Top: The instrumented vehicle used in the DIPLECS project. Bottom: The combined view given by stitching the views given by the three individual cameras mounted on the roof of the vehicle.

might be a stop sign or a give way sign. After additional runs when the driver makes a full stop even though no other cars were present, the system correctly deduces that the sign type is in fact a stop sign.

Research leading to the included publications on shape matching and traffic sign recognition [58, 59] and of learning tracking models [26] was conducted within this project. Other publications by the author that originates from the time in the DIPLECS project are [25, 60, 61]. The author was to a large extent involved in implementing the required functionalities from CVL in the main demonstrator and was the main source behind implementing the functionalities needed for multi target tracking based on the *channel based tracking framework*, see paper E.

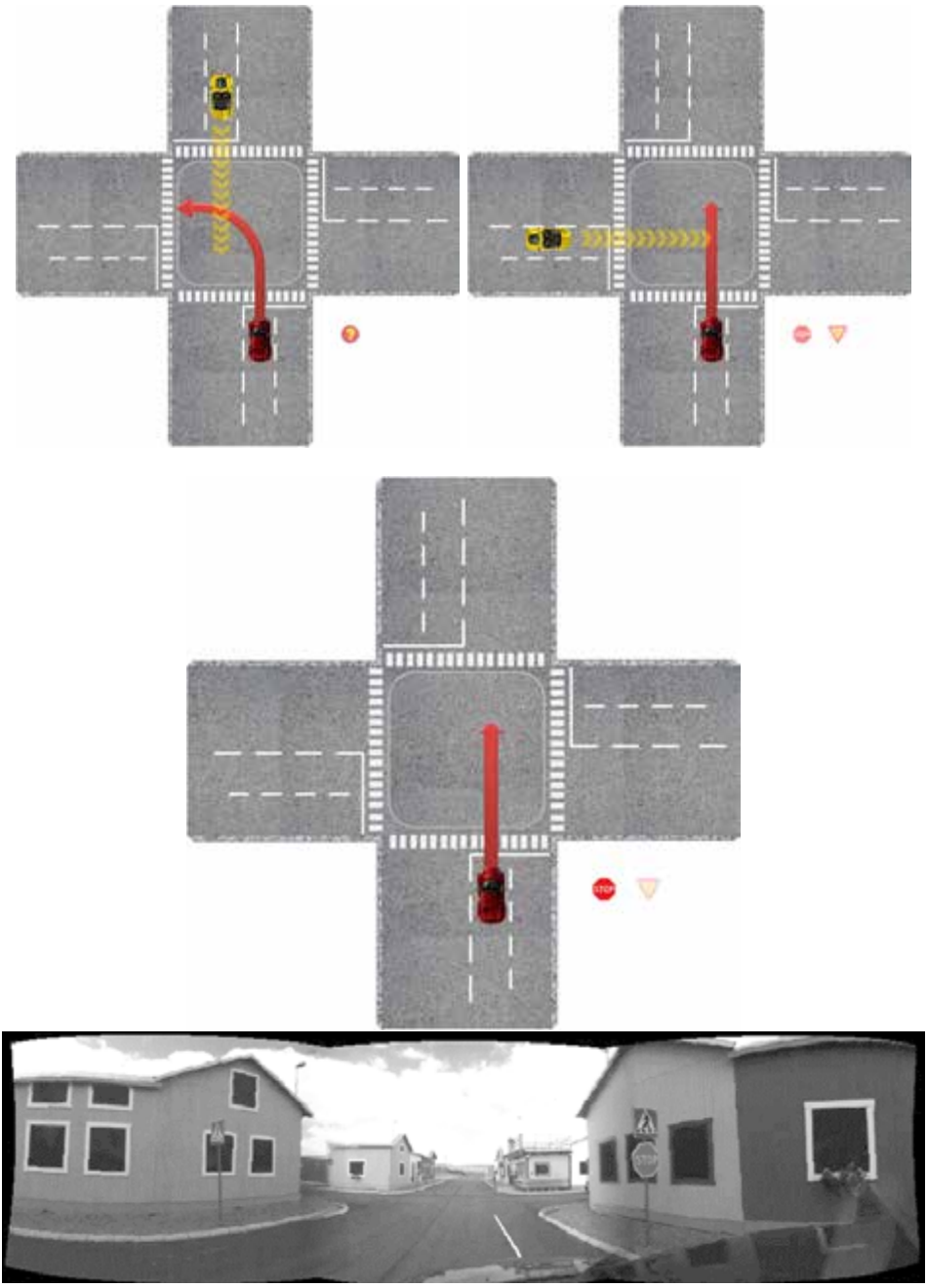


Figure 1.4: Upper left: Unknown sign. Upper right: Based on driver behavior the likelihoods of *give way sign* and *stop sign* are equal. Middle: Based on behavior, the system is confident that the sign is a *stop sign*. Bottom: View while approaching the junction.

1.3.3 ETT: Extended Target Tracking

The project ETT, *Extended Target Tracking*, running 2011-2014 aims at multiple and extended target tracking. Traditionally targets have been represented by their kinematic state (position, velocity, etc.). The project investigates new ways of extending the state vector and moving away from just a point target description. Early results, described in the included paper B, have been in the area of bicycle tracking where the bicycle is treated as a weakly articulated object and the observations consist of the projected ellipses originating from the bicycle wheels, see Fig. 1.5.

Figure 1.5: Image of a bike with estimated ellipses belonging to the bike wheels. The estimated ellipses are halfway between the colored lines.

1.4 Publications

This is a complete list of publications by the author.

Journal Papers

F. Larsson, M. Felsberg, and P.-E. Forssén. Correlating Fourier descriptors of local patches for road sign recognition. *IET Computer Vision*, 5(4):244–254, 2011

F. Larsson, E. Jonsson, and M. Felsberg. Simultaneously learning to recognize and control a low-cost robotic arm. *Image and Vision Computing*, 27(11):1729–1739, 2009

Peer-Reviewed Conference Papers

T. Ardeshiri, F. Larsson, F. Gustafsson, T. Schön, and M. Felsberg. Bicycle tracking using ellipse extraction. In *Proceedings of the 14th International Conference on Information Fusion*, 2011. **Honorable mention, nominated for the best student paper award**

F. Larsson and M. Felsberg. Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition. In *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, volume 6688 of *Lecture Notes in Computer Science*, pages 238–249, 2011

M. Felsberg and F. Larsson. Learning object tracking in image sequences. In *Proceedings of the International Conference on Cognitive Systems*, 2010

M. Felsberg, F. Larsson, W. Han, A. Ynnerman, and T. Schön. Torchlight navigation. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, 2010

M. Felsberg and F. Larsson. Learning higher-order Markov models for object tracking in image sequences. In *Proceedings of the International Symposium on Visual Computing (ISVC)*, volume 5876 of *Lecture Notes in Computer Science*, pages 184–195. Springer-Verlag, 2009

F. Larsson, M. Felsberg, and P-E. Forssén. Patch contour matching by correlating Fourier descriptors. In *Digital Image Computing: Techniques and Applications (DICTA)*, Melbourne, Australia, December 2009. IEEE Computer Society

M. Felsberg and F. Larsson. Learning Bayesian tracking for motion estimation. In *Proceedings of the European Conference on Computer Vision (ECCV), International Workshop on Machine Learning for Vision-based Motion Analysis*, 2008

F. Larsson, E. Jonsson, and M. Felsberg. Visual servoing for floppy robots using LWPR. In *Workshop on Robotics and Mathematics (ROBOMAT)*, pages 225–230, 2007

Other Conference Papers

F. Larsson and M. Felsberg. Traffic sign recognition using Fourier descriptors and spatial models. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2011

M. Felsberg, F. Larsson, W. Han, A. Ynnerman, and T. Schön. Torch guided navigation. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2010. **Awarded a paper award at the conference.**

F. Larsson, P-E. Forssén, and M. Felsberg. Using Fourier descriptors for local region matching. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2009

F. Larsson, E. Jonsson, and M. Felsberg. Learning floppy robot control. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2008

F. Larsson, E. Jonsson, and M. Felsberg. Visual servoing based on learned inverse kinematics. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2007

Theses

F. Larsson. *Methods for Visually Guided Robotic Systems: Matching, Tracking and Servoing.* Linköping Studies in Science and Technology. Thesis No. 1416, Linköping University, 2009

F. Larsson. *Visual Servoing Based on Learned Inverse Kinematics.* M.Sc. Thesis LITH-ISY-EX–07/3929, Linköping University, 2007

Reports

F. Larsson. Automatic 3D Model Construction for Turn-Table Sequences - A Simplification. LiTH-ISY-R, 3022, Linköping University, Department of Electrical Engineering, 2011

Part I

Background Theory

Chapter 2

Shape Matching

Shape matching is an ever popular area of research within the computer vision community that, as the name implies, concerns representing and recognizing arbitrary shapes. This chapter contains a brief introduction to the field of 2D shape matching and is intended as preparation for papers A-D which, to varying degree, deal with shape matching. Included is also a section on conics, containing an extended derivation of the relationship linking relative pose and the reflection of a light beam from a torchlight, used in paper A.

2.1 Overview

A common classification of shape matching methods is into *region based* and *contour based* methods. Contour based methods aim to capture the information contained on the boundary/contour only, while region based methods also include information about the internal region. Both classes can further be divided into *local* or *global* methods. Global methods treat the whole shape at once while local methods divide the shape into parts that are described individually in order to increase robustness to e.g. occlusion. See [69, 86, 91] for three excellent survey papers on shape matching.

When dealing with shape matching, an important aspect to take into consideration is which invariances are appropriate. Depending on the task at hand a particular invariance might either be beneficial or harmful. Take optical character recognition, OCR, as one example. For this particular application, full rotation invariance would be harmful since a *9* and a *6* would be confused. This is similar to the situation we face in the included papers C and D that deal with traffic sign recognition; We do not want to confuse the numbers on speed signs nor the diamond shape of Swedish main road signs with the shapes of square windows.

Depending on the desired invariance properties, different methods aim for different invariances, for example invariance under projective transformations [79], affine transformations [4], or non-rigid deformations [14, 31], just to mention a few. For an overview of invariance properties of different shape descriptors see the extensive listing and description of over 40 different methods in [86].

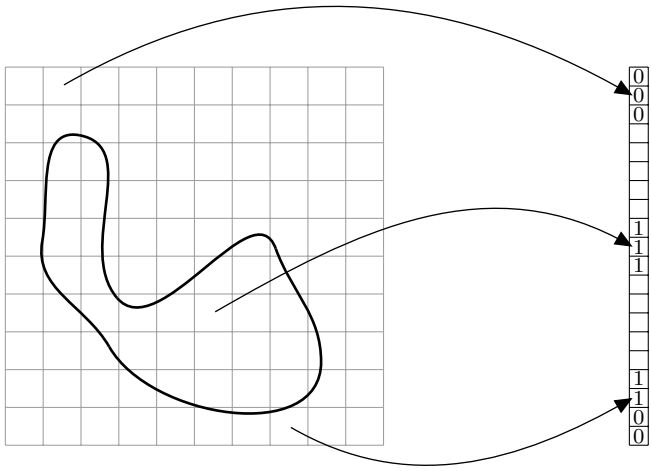


Figure 2.1: Illustration of a grid based method for describing shape. The grid is transformed into a vector and each tile is marked with hit=1, if it touches or is within the boundary, or miss=0, otherwise.

2.1.1 Region Based Matching

Region based methods aim to capture information not only from the boundary but also from the internal region of the shape. A simple and intuitive example of a region based method is the grid based method [70] illustrated in Fig. 2.1. This approach places a grid over the *canonical version*, i.e. normalized with respect to rotation, scale etc, of the shape. The grid is then transformed into a binary feature vector with the same length as the number of tiles in the grid. Ones indicate that the corresponding grid tiles touch the shape and zeros that the tiles are completely outside the shape. Note that this simple method does not capture any texture information. Popular region based approaches are moment based methods [43, 52, 82], generic Fourier descriptors [89] and methods based on the medial axis/skeleton such as [78].

2.1.2 Contour Based Matching

Contour based methods only account for the information given by contour itself. A simple example of a contour based method is shape signatures [19]. Shape signatures are basically representations based on a one-dimensional parameterization of the contour. This can be achieved using scalar valued functions, e.g. the distance to the center of gravity as a function of distance traveled along the contour as in Fig. 2.2, or functions with multivariate output, e.g using the full vector to the center of gravity not just the distance.

Shape signatures provide a periodic one-dimensional parameterization of the shape. It is thus a natural step to apply the Fourier transform to this periodic

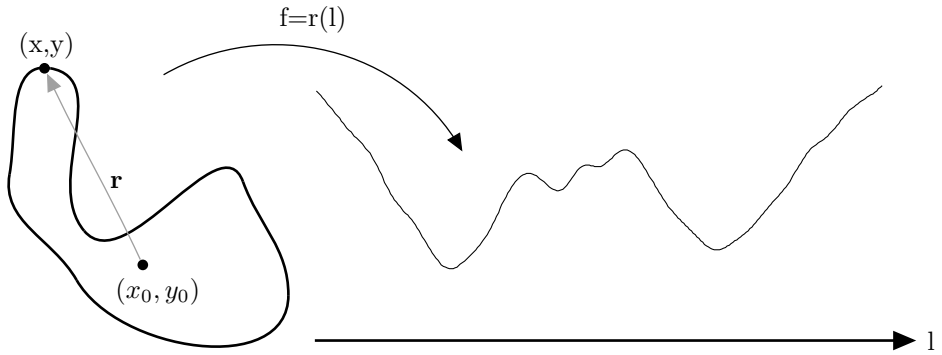


Figure 2.2: Shape signature based on distance to the center of gravity.

signal and this is exactly what is done in order to obtain Fourier Descriptors (FDs) [37, 88]. FDs use the Fourier coefficients of the 1D Fourier transform of the shape signature. Different shape signatures have been used with the Fourier descriptor framework, e.g. distance to centroid, curvature and complex valued representation. For more details on FDs see the included papers [58, 59] where we show that it is possible to retain the phase information and perform sum-of-squared differences matching without explicitly de-rotating the FDs.

Another popular contour based method is the *curvature scale space*, CSS, [73] that is incorporated in the MPEG-7 visual shape descriptors standard [12]. The CSS descriptor is based on inflection points of successively smoothed versions of the contour. The authors of [90] present an extensive comparison between FDs and CSS. In their study they show that FDs outperform CCS on the MPEG-7 contour shape database.

Shape context is another popular global contour based descriptor [10]. The descriptor is computed as log-polar histograms of edge energy around points sampled from the contour. Matching individual histograms is commonly done using χ^2 test statistics and the matching cost between shapes is given from the pairing of points that minimize the total sum of individual costs. Shape context allows for small non-rigid deformations.

One limitation with contour based methods is that they tend to be sensitive to noise and errors in the contour segmentation process. Small changes in the contour may result in big changes in the shape descriptor making matching impossible. Region based methods are less sensitive to noise since small changes of the contour leave the interior relatively unchanged. For an in depth discussion of pros and cons of the different approaches see [91].

2.1.3 Partial Contour Matching and Non-Rigid Matching

The difficulties involved in achieving reliable segmentation of shapes in natural images have led to work on shape matching based on local contour segments and different voting techniques [9, 33, 34, 68, 77]. Another rapidly evolving area is that of non-rigid shape matching based on chord angles [18], shape contexts [10], triangulated graphs [31] and shape-trees [32]. The interested reader is referred to [33], regarding partial contour matching, and to [14], regarding non-rigid matching, for the numerous references therein.

Many of the successful methods for recognition of deformable shapes tend to be very slow. The mentioned papers [18] and [32] take about 1h and 136h respectively for the MPEG-7 dataset (for which they currently rank 11th and 3rd in bulls eye score). The current best methods on the MPEG-7 dataset [8, 87] do not report any running times. As a comparison, our FDs based matching method in paper B takes less than 30 seconds on the same dataset, although at worse bulls eye score due to not dealing with non-rigid deformations.

In our work we have focused on recognition of closed contours and this fits well with our main application, traffic sign recognition. Traffic signs are designed to have easily distinguishable regions and are placed in such a way that they are rarely occluded. Traffic signs are also rigid objects meaning that invariance to non-rigid deformations could be harmful in this application domain.

2.2 Conics

Conics have a prominent role in two of the included papers and thus deserve a thorough introduction. A conic, or rather conic section, is the result of the intersection between a cone and a plane.

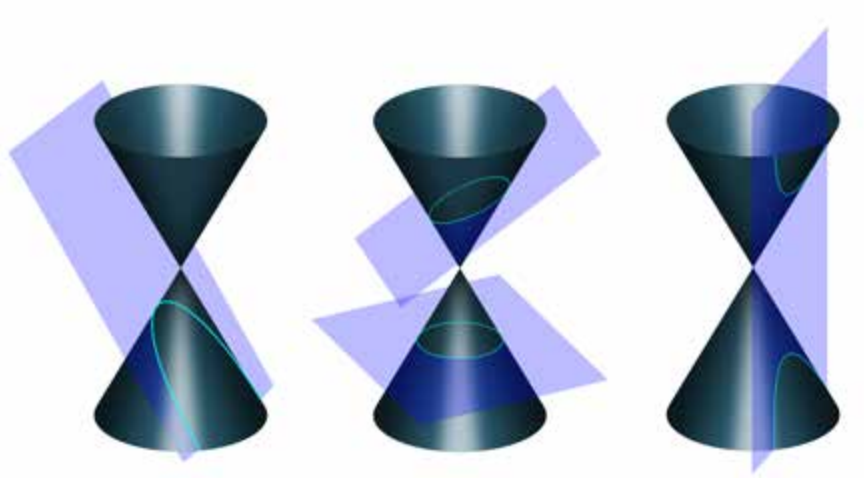


Figure 2.3: Illustration of the three types of conics. Left: Parabolas. Center: Ellipses. Right: Hyperbolas. Image adapted from Wikimedia Commons [17].

Conics are represented by the following second order polynomial

$$ax^2 + 2bxy + cy^2 + 2dx + 2ey + f = 0 \tag{2.1}$$

where x, y denote coordinates in the plane and a, b, c, d, e, f denote the coefficients defining the conic. Using homogeneous coordinates and matrix notation, (2.1) can be written as

$$\mathbf{p}^{\mathrm{T}}\mathbf{C}\mathbf{p} = 0 \tag{2.2}$$

where

$$\mathbf{p} = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{2.3}$$

and

$$\mathbf{C} = \begin{pmatrix} a & b & d \\ b & c & e \\ d & e & f \end{pmatrix} . \tag{2.4}$$

Note that any multiple of $\mathbf{C}$ defines the same conic, thus a conic has only five degrees of freedom.

A conic with $\det(\mathbf{C}) \neq 0$ is called a non-degenerate conic. Three types of non-degenerate conics exist in the Euclidean case: parabolas, hyperbolas and ellipses with circles being a special case of ellipses, see Fig. 2.3. It is possible to classify a non-degenerate/degenerate conic based on the determinant of the upper left 2×2 submatrix

$$\mathbf{C}_{22} = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \tag{2.5}$$

according to Table 2.1 [50, 66, 67].

	$\det(\mathbf{C}) \neq 0$	$\det(\mathbf{C}) = 0$	
$\det(\mathbf{C}_{22}) > 0$	$a + c < 0$: real ellipse $a + c > 0$: imaginary ellipse	point ellipse	
$\det(\mathbf{C}_{22}) = 0$	parabola	$\mathrm{rank}(\mathbf{C}) = 2$:	two unique parallel lines
		$\mathrm{rank}(\mathbf{C}) = 1$:	two coincident parallel lines
$\det(\mathbf{C}_{22}) < 0$	hyperbola	two intersecting lines	

Table 2.1: Classification of the different types of conics.

For the case of an ellipse, the center of the conic is given as

$$\begin{pmatrix} x_c \\ y_c \end{pmatrix} = \mathbf{C}_{22}^{-1} \begin{pmatrix} -d \\ -e \end{pmatrix} , \tag{2.6}$$

and the directions of the major and minor axes are given by the eigenvectors of $\mathbf{C}_{22}$.

The relations and properties mentioned above hold for the Euclidean case. For more information on the properties of conics in different spaces see [11, 40, 50].

2.3 The Conic From a Torchlight

This is an extended version of the derivation of the resulting conic from a reflected light beam used in paper A. This conic relates the reprojection of the light beam emitted by a torchlight to the relative pose of the illuminated object. Related work dealing with pose estimation from (multiple) conics can be found in [48, 51, 81].

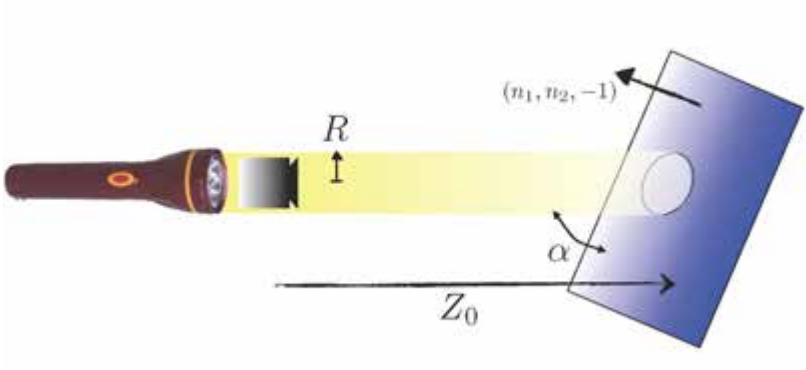


Figure 2.4: The torchlight setup used in paper A.

For the rest of this section; capital scalars X,Y,Z denote world coordinates while lower case scalars x,y denote image coordinates. The subscripts o, p are used if there is need to distinguish between *orthographic camera*, i.e. parallel projection, or *pinhole camera*. The same definitions as in [40] are used regarding orthographic and pinhole camera.

Assume that the world coordinate system is placed at the optical center of a pinhole camera and that the optical axis is aligned to the world Z-axis. The emitted light is assumed to form a perfect cylinder with radius R and that propagates in the direction of the optical axis, see Fig. 2.4. The light beam is intersected by a plane $\mathcal{P}$ and this will, under the mild assumption that the plane normal is not orthogonal to the optical axis, result in an ellipse [42]. If the plane normal is orthogonal to the optical axis the result is a line, or rather two coincident parallel lines according to the previous section. The camera views the illuminated plane, which results in a bright ellipse in the image plane, described by $\mathbf{C}_p$, that is directly related to the relative pose. This is the same assumptions as made in the included paper A.

We are looking for the resulting conic $\mathbf{C}_p$ in the image of the pinhole camera. One way is to first find the expression, in world coordinates, of the resulting quadric describing the intersection of the light beam and the plane $\mathcal{P}$, and then project this quadric into the pinhole camera. However, an easier way is to first assume that we use an orthographic camera placed at the same position as the pinhole camera and find $\mathbf{C}_o$, the conic in the orthographic image. Finding $\mathbf{C}_o$ is trivial since the optical axis is assumed to be the same as the direction of light propagation. This conic can then be transformed into the pinhole camera using a homography resulting in the desired $\mathbf{C}_p$.

The rest of this section is structured as follows. First the derivation of the homography relating the two cameras is described, secondly the resulting conic in the orthographic camera is discussed, and thirdly the homography and the orthographic conic are used in order to find the desired conic in the pinhole camera.

Finding the Homography

Under the assumption that the two cameras are viewing the same plane $\mathcal{P}$, a homography relates the coordinates in the orthographic camera to coordinates in the pinhole camera. This homography $\mathbf{H}$ is, up to a scalar, given by the relation

$$\mathbf{H}\mathbf{p}_o = \mathbf{p}_p \tag{2.7}$$

where $\mathbf{p}_o, \mathbf{p}_p$ denote homogeneous coordinates in the two cameras. This can further be written as

$$\mathbf{H}\mathbf{P}_o \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} = \mathbf{P}_p \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \ , \tag{2.8}$$

where $\mathbf{P}_o, \mathbf{P}_p$ denote corresponding projection matrices and $(X, Y, Z) \in \mathcal{P}$. The orthographic projection matrix is given as

$$\mathbf{P}_o = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.9}$$

while the actually used camera is modelled as a pinhole camera with focal length f and projection matrix

$$\mathbf{P}_p = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \ . \tag{2.10}$$

Further assume that the the plane $\mathcal{P}$ lies at distance Z_0 with normal $(n_1, n_2, -1)^{\mathrm{T}}$ and is parametrized over (X, Y) as

$$Z(X, Y) = n_1 X + n_2 Y + Z_0 \ . \tag{2.11}$$

Combining equations (2.8)-(2.11) result in

$$\mathbf{H}\mathbf{P}_o \begin{pmatrix} X \\ Y \\ n_1 X + n_2 Y + Z_0 \\ 1 \end{pmatrix} = \mathbf{P}_p \begin{pmatrix} X \\ Y \\ n_1 X + n_2 Y + Z_0 \\ 1 \end{pmatrix} \tag{2.12}$$

$$\mathbf{H} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} = \begin{pmatrix} fX \\ fY \\ n_1 X + n_2 Y + Z_0 \end{pmatrix} \tag{2.13}$$

and the final homography is identified as

$$\mathbf{H} = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ n_1 & n_2 & Z_0 \end{pmatrix} . \tag{2.14}$$

Finding the Conic in the Orthographic Camera

The light beam/cylinder is given as

$$L(X,Y,Z) = \begin{cases} 1 & X^2+Y^2 \leq R^2 \\ 0 & X^2+Y^2 > R^2 \end{cases} , \tag{2.15}$$

where X, Y, Z denote world coordinates and R is the radius of the beam. The conic describing the image of the outer contour in the orthographic camera $\mathbf{P}_o$, see (2.9), is readily identified as

$$x_o^2 + y_o^2 = R^2 \tag{2.16}$$

where (x_o, y_o) denote the coordinates in the image plane. This can further be written as

$$\mathbf{p}_o^{\mathrm{T}} \mathbf{C}_o \mathbf{p}_o = 0 \tag{2.17}$$

using homogeneous coordinates $\mathbf{p}_o = [x_o, y_o, 1]^{\mathrm{T}}$ and the matrix representation of the conic, where

$$\mathbf{C}_o = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -R^2 \end{pmatrix} . \tag{2.18}$$

Transforming the Conic into the Pinhole Camera

Equation (2.14) describes the mapping from coordinates in the orthographic image into coordinates in the pinhole image, see (2.7). According to [40], the corresponding transformation of $\mathbf{C}_o$ into $\mathbf{C}_p$ is

$$\mathbf{C}_p = \mathbf{H}^{-\mathrm{T}} \mathbf{C}_o \mathbf{H}^{-1} . \tag{2.19}$$

This can be verified by manipulating (2.17) according to

$$\begin{aligned} 0 &= \mathbf{p}_o^{\mathrm{T}} \mathbf{C}_o \mathbf{p}_o && (2.20) \\ 0 &= \mathbf{p}_o^{\mathrm{T}} (\mathbf{H}^{\mathrm{T}} \mathbf{H}^{-\mathrm{T}}) \mathbf{C}_o (\mathbf{H}^{-1} \mathbf{H}) \mathbf{p}_o && (2.21) \\ 0 &= (\mathbf{H}\mathbf{p}_o)^{\mathrm{T}} \mathbf{H}^{-\mathrm{T}} \mathbf{C}_o \mathbf{H}^{-1} (\mathbf{H}\mathbf{p}_o) && (2.22) \end{aligned}$$

and identifying $\mathbf{p}_p = (\mathbf{H}\mathbf{p}_o)$ which gives

$$\begin{aligned} 0 &= \mathbf{p}_p^{\mathrm{T}} \mathbf{H}^{-\mathrm{T}} \mathbf{C}_o \mathbf{H}^{-1} \mathbf{p}_p && (2.23) \\ 0 &= \mathbf{p}_p^{\mathrm{T}} \mathbf{C}_p \mathbf{p}_p . && (2.24) \end{aligned}$$

Combining (2.14), (2.18) and (2.19) gives

$$\mathbf{C}_p = \begin{pmatrix} \frac{1}{f^2} - \frac{R^2 n_1^2}{Z_0^2 f^2} & -\frac{R^2 n_1 n_2}{Z_0^2 f^2} & \frac{R^2 n_1}{Z_0^2 f} \\ -\frac{R^2 n_1 n_2}{Z_0^2 f^2} & \frac{1}{f^2} - \frac{R^2 n_2^2}{Z_0^2 f^2} & \frac{R^2 n_2}{Z_0^2 f} \\ \frac{R^2 n_1}{Z_0^2 f} & \frac{R^2 n_2}{Z_0^2 f} & -\frac{R^2}{Z_0^2} \end{pmatrix} . \tag{2.25}$$

$\mathbf{C}_p$ being a projective element, allows simplification of (2.25) by multiplication with $\frac{Z_0^2 f^2}{R^2}$ giving

$$\mathbf{C}_p = \begin{pmatrix} \frac{Z_0^2}{R^2} - n_1^2 & -n_1 n_2 & f n_1 \\ -n_1 n_2 & \frac{Z_0^2}{R^2} - n_2^2 & f n_2 \\ f n_1 & f n_2 & -f^2 \end{pmatrix} , \tag{2.26}$$

which is also the form used in paper A.

Chapter 3

Tracking

This chapter is an extended version of the brief introductions to Bayesian tracking contained in the included papers B and E. Included is also a section on the *channel representation* used in paper E. The channel representation is a sparse localized representation that, among other things, can be used for estimation and representation of probability density functions.

3.1 Bayesian Tracking

Throughout this thesis, the term *tracking* refers to *Bayesian tracking* unless otherwise stated. This should not be confused with *visual tracking* techniques, such as the KLT-tracker [71], which minimizes a cost function directly in the image domain.

Bayesian tracking (or Bayesian filtering) techniques address the problem of estimating an object's state vector, which may consist of arbitrary abstract properties, based on measurements, which are usually not direct measurements of the tracked state dimensions. Applications can be estimating the 3D position of an object based on the (x, y)-position in the image plane or estimating the pose vector of a bicycle based on observations of the wheels, as in paper B. Bayesian tracking techniques are often applied to visual data, see e.g. [13, 45, 74, 83].

Assume a system that changes over time and a way to acquire measurements from the same system. The task is then to estimate the probability of each possible state of the system given all measurements up to the current time step. To put it more formally: In Bayesian tracking, the current system state is represented as a probability density function (pdf) over the system's state space. The state density for a given time is estimated in a two separate steps. First, the pdf from the previous time step is propagated through the system model which gives a prior estimate for the current state. Secondly, new measurements are used to update the prior distribution which results in the state estimate for the current time step, i.e. the posterior distribution. The process is commonly illustrated as a closed loop with two phases, see Fig. 3.1.

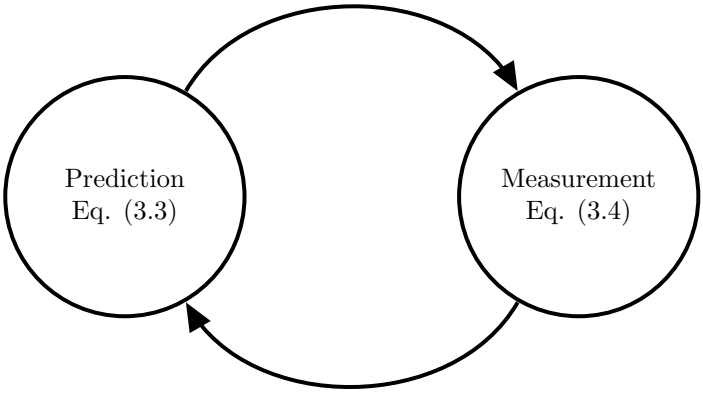


Figure 3.1: Illustration of the Bayesian tracking loop. The loop alternates between making predictions and incorporating new measurements.

Using the same notation as in [7, 26], the system model $\mathbf{f}$ is given as:

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}, \mathbf{v}_{k-1}) \ , \tag{3.1}$$

where $\mathbf{x}_k$ denotes the state space of the system and $\mathbf{v}_k$ denotes the noise term, both at time k. The system model describes how the system state changes over time k. The measurement model $\mathbf{h}$ is defined as:

$$\mathbf{z}_k = \mathbf{h}(\mathbf{x}_k, \mathbf{n}_k) \ , \tag{3.2}$$

where $\mathbf{n}_k$ denotes the noise term at time k. The task is thus to estimate the pdf $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, were $\mathbf{z}_{1:k}$ denotes all measurements from time 1 to k. This is achieved by combining the old state estimate with new measurements. The old state estimate is propagated through the system model resulting in a prediction/prior distribution for the new time step. Given the previous measurements and the system model, the prior distribution is

$$p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1} \ . \tag{3.3}$$

Which is the result of (3.1) representing a first order Markov model. When new measurements become available, the prior distribution is updated accordingly and the estimate of the posterior distribution is obtained as

$$\begin{aligned} p(\mathbf{x}_k|\mathbf{z}_{1:k}) \quad &= \quad p(\mathbf{x}_k|\mathbf{z}_{1:k-1}, \mathbf{z}_k) = \frac{p(\mathbf{z}_k|\mathbf{x}_k, \mathbf{z}_{1:k-1}) p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} = \\ &\overbrace{=}^{(3.2)} \quad \frac{p(\mathbf{z}_k|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \ . \end{aligned} \tag{3.4}$$

The denominator in (3.4),

$$p(\mathbf{z}_k|\mathbf{z}_{1:k-1}) = \int p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})d\mathbf{x}_k \ , \tag{3.5}$$

acts as a normalizing constant ensuring that the posterior estimate is a proper pdf.

It is possible to estimate $\mathbf{x}_k$ by recurrent use of (3.3) and (3.4) given an estimate of the initial state $p(\mathbf{x}_0)$, and assuming $p(\mathbf{x}_0|\mathbf{z}_0) = p(\mathbf{x}_0)$.

Equation (3.4) can be solved exactly or only approximately depending on the assumptions made about the system. Under the assumption of a linear system model and a linear measurement models combined with Gaussian white noise [49] the Kalman filter is the optimal recursive solution in the maximum likelihood sense. Various numerical methods exist for handling the general case with non-linear models and non-Gaussian noise, e.g. particle filters [36] and grid-based methods [7]. For a good introduction and overview of Bayesian estimation techniques see [7, 15].

3.2 Data Association

The problem of *data association* arises whenever measurements might come from multiple sources, such as in multi-target tracking, or in the presence of false and/or missing measurements. The problem is to correctly associate the acquired measurements to the tracked targets. This is one of the greatest and most fundamental challenges when dealing with Bayesian tracking in computer vision [6].

There are numerous reasons why this is a hard and still largely an unsolved problem. At each time step, the prediction from the previous one is to be matched to the new measurements. If there are no new measurements matching the prediction, this might be due to occlusion, incorrect prediction, or that the tracked object have ceased to exist. If there are multiple measurements matching the prediction, a decision has to be made regarding which one, if any, to use. If there are multiple targets matching a single measurement, this situation must also be dealt with.

The most straightforward way of dealing with the problem is the greedy nearest neighbor principle. Target-measurement associations are simply made such that each prediction is paired with the nearest still unused measurement. This approach requires making hard associations at each time step. Consequently, if an incorrect association is made, recovery is unlikely.

Other approaches postpone the association decision by looking at the development over a window in time, e.g. Multiple Hypotheses Tracking (MHT). Another strategy is to update each prediction based on all available measurements, but to weight the importance of each measurement according to their agreement with the prediction, e.g. Probabilistic Data Association Filter (PDAF) [76], Joint PDAF and Probabilistic Multiple Hypotheses Tracking (PMHT) [80]. Much research is undertaken within this field, see e.g. approaches based on *random finite sets* such as the Probability Hypothesis Density (PHD) filter [72].

3.3 Channel Representation

This section contains an extended version of the brief introduction to the channel representation found in paper E. The channel representation is a sparse localized representation [38], which is used in the included paper to represent probability density functions.

Channel encoding is a way to transform a compact representation, such as numbers, into a sparse localized representation. For an overview and definitions of the aspects of compact/sparse/local representations see [35]. This introduction to the channel representation is limited to the encoding of scalars, but the representation readily generalizes to multiple dimensions.

Using the same notation as in [47]; a *channel vector* $\mathbf{c}$ is constructed from a scalar x by the nonlinear transformation

$$\mathbf{c} = [B(x - \tilde{x}_1), B(x - \tilde{x}_2), \ \ldots \ , B(x - \tilde{x}_N)]^{\mathrm{T}} \ . \tag{3.6}$$

Where $B(\cdot)$ denotes the basis/kernel function used. B is often chosen to be symmetric, non-negative and with compact support. The kernel centers $\tilde{x}_i$ can be placed arbitrarily in the input space, but are often uniformly distributed. The process of creating a channel vector from a scalar or another compact representation is referred to as *channel encoding* and the opposite process is referred to as *decoding*. Gaussians, B-splines, and windowed $\cos^2$ functions are examples of suitable kernel functions [35].

Using the windowed $\cos^2$ function

$$B(x) = \begin{cases} \cos^2(ax) & \text{if } |x| \leq \frac{\pi}{2a} \\ 0 & \text{otherwise} \end{cases} \ , \tag{3.7}$$

and placing 10 kernels centered on integer values $\tilde{x}_i \in [1, 10]$, gives the basis functions seen in Fig. 3.2. For this example the kernel width is set to $a = \frac{\pi}{3}$, which means that there are always three simultaneously non-zero kernels for the domain $[1.5, 9.5]$. How to properly choose a depending on required spatial and feature resolution is addressed in [22]. Encoding the scalar $x = 3.3$ using these

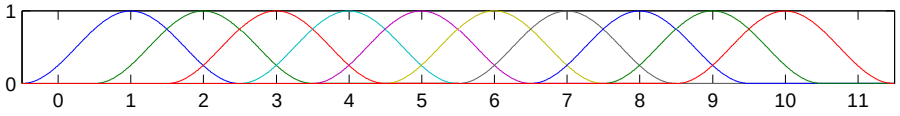


Figure 3.2: Ten $\cos^2$ kernels with respective kernel centered on integer values.

kernels results in the channel vector

$$\begin{aligned} \mathbf{c} &= [B(2.3), B(1.3), B(0.3), \ \ldots \ , B(-6.7)]^{\mathrm{T}} \\ &= [\ 0 \quad 0.04 \quad 0.90 \quad 0.55 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \ ]^{\mathrm{T}} \ . \end{aligned} \tag{3.8}$$

Note that only a few of the channels have a non-zero value, and that only channels close to each other are activated. This illustrates how channel encoding results in

a sparse localized representation. The basic idea when decoding a channel vector is to consider only a few neighboring channels at a time in order to ensure that the locality is preserved in the decoding process as well. The decoding algorithm for the $\cos^2(\cdot)$ kernels in (3.7) is adapted from [35] and is repeated here for completeness

$$\hat{x}^l = l + \frac{1}{2a} \arg\left(\sum_{k=l}^{l+M-1} c^k e^{i2a(k-l)} \right) \quad . \tag{3.9}$$

Here, c^k denotes the kth element in the channel vector, l indicates the element position in the resulting vector and $M = \frac{\pi}{a}$ indicates how many channels that are considered at a time, i.e. $M = 3$ in our case. An estimate $\hat{x}^l$ that is outside its valid range $[l + 1.5, l + 2.5]$ is rejected. Additionally each decoded value is accompanied by a certainty measure r

$$r^l = l + \frac{1}{M} \sum_{k=l}^{l+M-1} c^k \quad . \tag{3.10}$$

Applying (3.9) and (3.10) to (3.8) results in

$$\hat{\mathbf{x}} = [\ -0.02 \quad 3.30 \quad 3.31 \quad 4.00 \quad 5.00 \quad 6.00 \quad 7.00 \quad 8.00\ ]^{\mathrm{T}} \tag{3.11}$$

$$\mathbf{r} = [\ 0.95 \quad 1.50 \quad 1.46 \quad 0.55 \quad 0.00 \quad 0.00 \quad 0.00 \quad 0.00\ ]^{\mathrm{T}} \quad . \tag{3.12}$$

Note that only the second element in $\hat{\mathbf{x}}$ is within its valid range, leaving only the correct estimate of 3.3 that also has the highest confidence.

Adding a number of channel vectors results in a soft histogram, i.e. a histogram with overlapping bins. Using the same kernels as above, and encoding $x_1 = 3.3$ and $x_2 = 6.8$ results in

$$\begin{array}{rcl} \mathbf{c}_1 & = & [\ 0 \quad 0.04 \quad 0.90 \quad 0.55 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\ ]^{\mathrm{T}} \\ \mathbf{c}_2 & = & [\ 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0.48 \quad 0.96 \quad 0.96 \quad 0\ ]^{\mathrm{T}} \end{array} \tag{3.13}$$

and the corresponding soft histogram

$$\mathbf{c} = \mathbf{c}_1 + \mathbf{c}_2 = [\ 0 \ \ 0.04 \ \ 0.90 \ \ 0.55 \ \ 0 \ \ 0 \ \ 0.48 \ \ 0.96 \ \ 0.96 \ \ 0\ ]^{\mathrm{T}} \quad . \tag{3.14}$$

Due to the locality of the representation, the two different scalars do not interfere with each other. Retrieving the original scalars is straightforward as long as they are sufficiently separated with respect to the kernels used. In the case of interference, retrieving the cluster centers is a simple procedure. For more details on decoding schemes see [35, 47]. The ability to simultaneously being able to represent multiple values can be used for e.g. estimating the local orientation in an image or representing multiple hypotheses for the state of a tracked target.

A certainty measure is also obtained while decoding making it possible to recover multiple modes with decreasing certainty. A certainty measure can also be included in the encoding process by simply multiplying the channel vector by the certainty. Examples of how this has been used can be found in paper E where this property is used for encoding noisy measurements.

As mentioned above, a soft histogram is obtained by adding channel vectors. This can be used for estimating and representing probability density functions (pdfs). It is simple to find the peaks of the pdf by decoding the channel vector, quite similar to locating the bin with most entries in ordinary histograms. However, the accuracy of an ordinary histogram is limited to the bin size. In the channel case, sub-bin accuracy is possible due to the fact that the channels are overlapping and that the distance to the channel-center determines the influence of each sample. It has been shown [24] that the use of the channel representation reduces the quantization effect by a factor up to 20 compared to ordinary histograms. Using channels instead of histograms allows for reducing the computational complexity, by using fewer bins, or to obtain a higher accuracy while using the same number of bins. It is also possible to obtain a continuous reconstruction of the underlying pdf, instead of just locating the peaks [47].

As previously stated, this is a very brief introduction to the channel representation. The interested reader is referred to [23, 35, 38, 46, 47] for in depth presentations.

Chapter 4

Visual Servoing

This chapter is intended as an extended introduction to paper F and contains an introduction to visual servoing adapting the nomenclature from [44, 53]. The use of visual information for robot control can be divided into two classes depending on approach; *open-loop systems* and *closed-loop systems*. The term visual servoing refers to the latter approach.

4.1 Open-Loop Systems

An open-loop system can be seen as a system working in two distinct phases where extraction of visual information is separated from the task of operating the robot. Information, e.g. the position of the object to be grasped, is extracted from the image(s) during the first phase. This information is then fed to a robot control system that moves the robot arm blindly during the second phase. This requires an accurate inverse kinematic model for the robot arm as well as an accurately calibrated camera system. Also, the environment needs to remain static between the assessment phase and the movement phase.

4.2 Visual Servoing

The second main approach is based on a closed-loop system architecture, often denoted visual servoing. The extraction of visual information and computation of control signals is more tightly coupled than for open-loop systems. Visual information is continuously used as feedback to update the control signals. This results in a system that is less dependent on static environment, calibrated camera(s) etc.

Depending on the method of transforming information into robot action, visual servoing systems are further divided into two subclasses, *dynamic look-and-move systems* and *direct visual servoing systems*. Dynamic look-and-move systems use visually extracted information as input to a robot controller that computes the desired joint configurations and then uses joint feedback to internally stabilize the robot. This means that once the desired lengths and angles of the joints have been

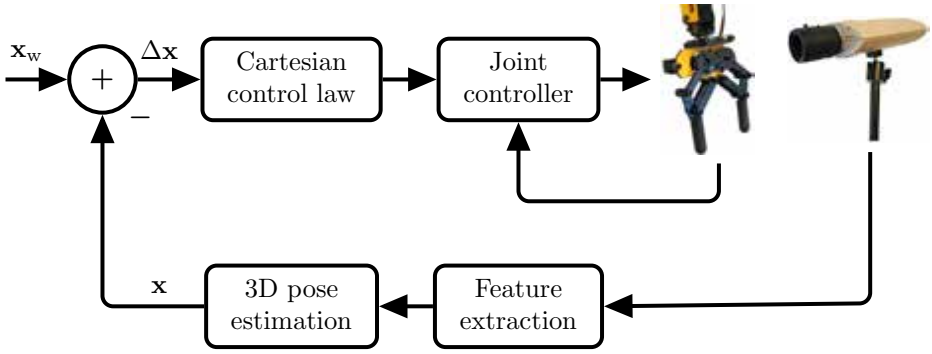


Figure 4.1: Flowchart for a position based dynamic look-and-move system. $\Delta\mathbf{x}$ denotes the deviation between target ($\mathbf{x}_w$) and reached ($\mathbf{x}$) configuration of the end-effector. All configurations are given in 3D positions for this position based setup.

computed, this configuration is reached. Direct visual servoing systems use the extracted information to directly compute the input to the robot, meaning that this approach can be used when no joint feedback is available.

Both the dynamic look-and-move and the direct visual servoing approach may be used in a *position based* or *image based* way, or in a combination of both. In a position based approach the images are processed such that relevant 3D information is retrieved in world/robot/camera coordinates. The process of positioning the robotic arm is then defined in the appropriate 3D coordinate system. In an image based approach, 2D information is directly used to decide how to position the robot, i.e. the robotic arm is to be moved to a position defined by image coordinates. See figure 4.1 and 4.2 for flowcharts describing the different system architectures.

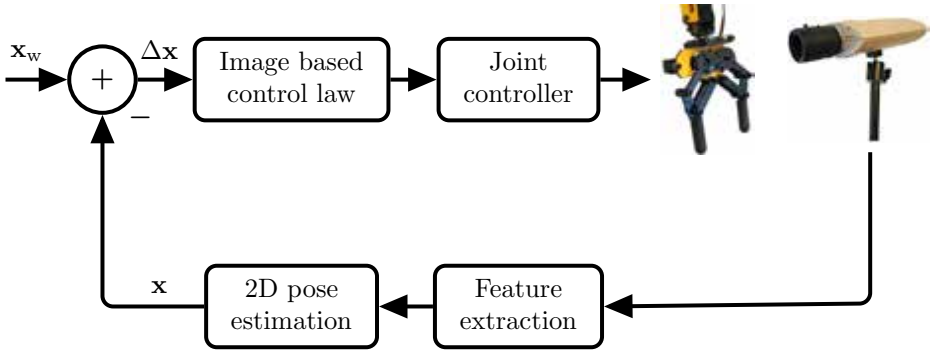


Figure 4.2: Flowchart for an image based direct visual servo system. $\Delta\mathbf{x}$ denotes the deviation between target ($\mathbf{x}_w$) and reached ($\mathbf{x}$) configuration of the end-effector. All configurations are given in 2D coordinates for this setup.

According to the introduced nomenclature the approach used in paper F is classified as image based direct visual servoing. The desired configuration is specified in terms of image coordinates for automatically acquired features which are directly mapped into control signals for the robotic arm.

4.3 The Visual Servoing Task

The task in visual servoing is to minimize the norm of the deviation vector $\Delta\mathbf{x} = \mathbf{x}_w - \mathbf{x}$, where $\mathbf{x}$ denotes the reached configuration and $\mathbf{x}_w$ denotes the target configuration. For example, the configuration $\mathbf{x}$ may denote position, velocity and/or jerk of the joints.

The configuration $\mathbf{x}$ is said to lie in the *task space* and the control signal $\mathbf{y}$ that generated this configuration is located in the *joint space*. The image Jacobian $\mathbf{J}_{\text{img}}$ is the linear mapping that maps changes in joint space $\Delta\mathbf{y}$ to changes in task space $\Delta\mathbf{x}$ such that:

$$\Delta\mathbf{x} = \mathbf{J}_{\text{img}}\Delta\mathbf{y}. \tag{4.1}$$

The term *image Jacobian* is used since the task space is often the acquired image(s). The configuration vector is then the position of features in these images. The term *interaction matrix* may sometimes be encountered instead of image Jacobian.

Furthermore, let $\mathbf{J}$ denote the inverse image Jacobian, i.e. a mapping from changes in task space to changes in joint space such that:

$$\Delta\mathbf{y} = \mathbf{J}\Delta\mathbf{x} \tag{4.2}$$

$$\mathbf{J} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}. \tag{4.3}$$

The term inverse image Jacobian does not necessarily mean that $\mathbf{J}$ is the mathematical inverse to $\mathbf{J}_{\text{img}}$. In fact, the mapping $\mathbf{J}_{\text{img}}$ does not need to be injective and hence not invertible. The word inverse simply implies that the inverse image Jacobian describes changes in joint spaces given wanted changes in task space while the image Jacobian describes changes in task space given changes in joint space.

If the inverse image Jacobian, or an estimate thereof, has been acquired, the task of correcting for an erroneous control signal is rather simple in theory. If the current position with deviation $\Delta\mathbf{x}$ originates from the control signal $\mathbf{y}$, the new control signal is then given as

$$\mathbf{y}_{\text{new}} = \mathbf{y} - \mathbf{J}\Delta\mathbf{x}. \tag{4.4}$$

However, in a non-ideal situation, the new control signal will most likely not result in the target configuration either. The process of estimating the Jacobian and updating the control signal needs to be repeated until a stopping criterion is met, e.g. the deviation is sufficiently small or the maximum number of iterations is reached.

Chapter 5

Concluding Remarks

Part I of this thesis covers some basic materials completing the publications included in Part II. This concluding section summarizes the main results and briefly discusses possible areas of future research.

5.1 Results

Much of the work within this thesis has been carried out in projects aiming for (cognitive) driver assistance systems and hopefully represents a step towards improving traffic safety. The main contributions are within the area of Computer Vision, and more specifically, within the areas of shape matching, Bayesian tracking, and visual servoing with the main focus being on shape matching and applications thereof. The different methods have been demonstrated in traffic safety applications, such as bicycle tracking, car tracking, and traffic sign recognition, as well as for pose estimation and robot control.

One of the core contributions is a new method for recognizing closed contours. This matching method in combination with spatial models has led to a methodology for traffic sign detection and recognition. Another contribution has been the extension of a framework for learning based Bayesian tracking called channel based tracking. The framework has been evaluated in car tracking scenarios and is shown to give competitive tracking performance, compared to standard approaches. The last field of contribution has been in cognitive robot control. A method is presented for learning how to control a robotic arm without knowing beforehand what it looks like or how it is controlled. Below follows a brief summary of the individual contributions in each the included papers.

Paper A contains work on relative pose estimation using a torch light. The reprojection of the emitted light beam creates, under certain conditions, an ellipse in the image plane. It is shown that it is possible to use this ellipse in order to estimate the relative pose between the torchlight and illuminated object.

Paper B builds on the ideas presented in paper A and contains initial work on bicycle tracking. The relative pose estimates are based on ellipses originating from the projection of the bicycle wheels into the image. This is combined with a

particle filter framework and a weakly articulated object model in order to track the bicycle in 3D. This approach is demonstrated in simulations and on real world data with encouraging results.

In paper C, a novel method for matching Fourier descriptors is presented and evaluated. One of the main conclusions is that it is important to keep the phase information when matching Fourier descriptors. Neglecting the phase corresponds to matching while minimizing the rotation difference between each individual pair of Fourier coefficients, instead of minimizing the rotation difference between the shapes. This can result in perfect matches between intrinsically different shapes. Another benefit of keeping the phase is that rotation covariant or invariant matching is achieved in the same way by using complex valued correlation. The only difference is to either consider the magnitude, for rotation invariant matching, or just the real value, for rotation covariant matching, of the complex valued correlation.

In paper D, the matching method presented in paper C is used in combination with an implicit star-shaped object models for traffic sign recognition. The presented method works fully automatically on query images with no need for regions-of-interests. It is shown that the presented method performs well for traffic signs that contain multiple distinct contours, while some improvement still is needed for signs defined by a single contour. The presented methodology is general enough to be used for arbitrary objects, as long as they can be defined by a number of regions. Another major contribution is the release of the first publicly available large database not only containing small patches around traffic signs, allowing for comparison of different approaches.

Paper E contains work on learning based object tracking and extends a framework for Bayesian tracking called channel based tracking. Compared to earlier work, the multi-dimensional case has been reformulated in a sound probabilistic way and the learning algorithm itself has been extended. The framework is evaluated in car tracking scenarios and is shown to give competitive tracking performance, compared to standard approaches.

Paper F describes a method that allows simultaneous learning of appearance and control of a robotic arm. The method achieves sufficient accuracy for simple assembly tasks by combining autonomous recognition with visual servoing, based on a learned mapping between percepts and actions. The paper demonstrates that limitations of inexpensive hardware, such as web cameras and low-cost robotic arms, can be overcome using powerful algorithms.

All in all, the methods developed and presented in this thesis can all be used for different components in a system guided by visual information, and hopefully represents a step towards improving traffic safety.

5.2 Future Work

The methods and results presented in this thesis are currently being developed and used within the projects ETT and GARNICS (Gardening with a Cognitive System [3]).

ETT, as described earlier, focuses on extended target tracking, with applications to the traffic safety domain. An interesting research direction would be to investigate to what degree the tracking framework could be incorporated directly in the matching of contours. By tracking contours over time it would be possible to learn what potential transformations that the contour can undergo. Given a few observations of a contour, the system could predict how the contour should look in the next time step, and this could be exploited in the matching step. Another obvious extension of the method presented in paper D would be to include color in the traffic sign prototypes. This is a straightforward extension and will likely lead to increased matching performance, although at a slightly higher computational cost.

GARNICS is a European project within the cognitive system domain and aims at 3D sensing of plant growth and building perceptual representations for learning the links to actions of a robot gardener. The Fourier descriptor based matching combined with the spatial models can potentially be used to keep track of the growth of plants by recognizing the individual leaves and their relative position. In an embodied setting such as GARNICS the tracking and recognition framework could be utilized for guiding the actions in case of uncertainties. If the system is uncertain of the identity of an object, actions can be chosen by consulting the tracking model in order to resolve these ambiguities.

Bibliography

[1] The COSPAL project. http://www.cospal.org.

[2] The DIPLECS project. http://www.diplecs.eu.

[3] The GARNICS project. http://www.garnics.eu.

[4] K. Arbter, W. Snyder, and H. Burkhardt. Application of Affine-Invariant Fourier Descriptors to Recognition of 3-D Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):640–647, 1990.

[5] T. Ardeshiri, F. Larsson, F. Gustafsson, T. Schön, and M. Felsberg. Bicycle tracking using ellipse extraction. In *Proceedings of the 14th International Conference on Information Fusion*, 2011.

[6] H. Ardö. *Multi-target Tracking Using on-line Viterbi Optimisation and Stochastic Modelling*. PhD thesis, Centre for Mathematical Sciences LTH, Lund University, Sweden, 2009.

[7] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[8] X. Bai, X. Yang, L. Latecki, W. Liu, and Z. Tu. Learning context sensitive shape similarity by graph transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):861–874, 2010.

[9] D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

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

[11] R. Bix. *Conics and Cubics*. Springer, 2006.

[12] M. Bober, F. Preteux, and Y.-M. Kim. MPEG-7 visual shape descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):716–719, June 2001.

[13] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online Multiperson Tracking-by-Detection from a Single, Uncalibrated Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1820–1833, 2011.

[14] A. Bronstein, M. Bronstein, A. Bruckstein, and R. Kimmel. Analysis of Two-Dimensional Non-Rigid Shapes. *International Journal of Computer Vision*, 78(1):67–88, 2008.

[15] Z. Chen. Bayesian filtering: From Kalman filters to particle filters, and beyond. Technical report, Communications Research Laboratory, McMaster University, 2003.

[16] Commission for Global Road Safety. *Make Roads Safe, A Decade of Action for Road Safety, ISBN-13: 978-0-9561403-2-6.* 2010.

[17] Wikimedia Commons. File:conic_sections_with_plane.svg. http://commons.wikimedia.org/wiki/File:Conic_sections_with_plane.svg.

[18] M. Donoser, H. Riemenschneider, and H. Bischof. Efficient partial shape matching of outer contours. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2009.

[19] A. El-ghazal, O. Basir, and S. Belkasim. Farthest point distance: A new shape signature for Fourier descriptors. *Signal Processing: Image Communication*, 24(7):572 – 586, 2009.

[20] L. Ellis and R. Bowden. Learning responses to visual stimuli: A generic approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2007.

[21] L. Ellis, M. Felsberg, and R. Bowden. Affordance mining: Forming perception through action. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2010.

[22] M. Felsberg. Spatio-featural scale-space. In *Proceedings of the International Conference on Scale Space Methods and Variational Methods in Computer Vision*, volume 5567 of *Lecture Notes in Computer Science (LNCS)*, 2009.

[23] M. Felsberg. Adaptive filtering using channel representations. In L. M. J. Florack, R. Duits, G. Jongbloed, M.-C. van Lieshout, and L. Davies, editors, *Locally Adaptive Filters in Signal and Image Processing*, pages 35–54. Springer, 2011.

[24] M. Felsberg, P.-E. Forssén, and H. Scharr. Channel smoothing: Efficient robust smoothing of low-level signal features. *IEEE Transactions on Pattern Analysis and Machine*, 28(2):209–222, 2006.

[25] M. Felsberg and F. Larsson. Learning Bayesian tracking for motion estimation. In *Proceedings of the European Conference on Computer Vision (ECCV), International Workshop on Machine Learning for Vision-based Motion Analysis*, 2008.

[26] M. Felsberg and F. Larsson. Learning higher-order Markov models for object tracking in image sequences. In *Proceedings of the International Symposium on Visual Computing (ISVC)*, volume 5876 of *Lecture Notes in Computer Science*, pages 184–195. Springer-Verlag, 2009.

[27] M. Felsberg and F. Larsson. Learning object tracking in image sequences. In *Proceedings of the International Conference on Cognitive Systems*, 2010.

[28] M. Felsberg, F. Larsson, W. Han, A. Ynnerman, and T. Schön. Torch guided navigation. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2010.

[29] M. Felsberg, F. Larsson, W. Han, A. Ynnerman, and T. Schön. Torchlight navigation. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, 2010.

[30] M. Felsberg, A. Shaukat, and D. Windridge. Online learning in perception-action systems. In *Proceedings of the European Conference on Computer Vision (ECCV), Workshop on Vision for Cognitive Tasks*, 2010.

[31] P. Felzenszwalb. Representation and Detection of Deformable Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):208–220, 2005.

[32] P. Felzenszwalb and J. Schwartz. Hierarchical matching of deformable shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[33] V. Ferrari, F. Jurie, and C. Schmid. From images to shape models for object detection. *International Journal of Computer Vision*, 87(3):284–303, 2010.

[34] S. Fidler, M. Boben, and A. Leonardis. Learning hierarchical compositional representations of object structure. In S. Dickinson, A. Leonardis, B. Schiele, and M.J. Tarr, editors, *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press, 2009.

[35] P-E. Forssén. *Low and Medium Level Vision using Channel Representations*. PhD thesis, Linköping University, Sweden, SE-581 83 Linköping, Sweden, March 2004. Dissertation No. 858, ISBN 91-7373-876-X.

[36] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, 1993.

[37] G. H. Granlund. Fourier Preprocessing for Hand Print Character Recognition. *IEEE Transactions on Computers*, C–21(2):195–201, 1972.

[38] G.H. Granlund. An associative perception-action structure using a localized space variant information representation. In *Proceedings of Algebraic Frames for the Perception-Action Cycle (AFPAC)*, 2000.

[39] G.H. Granlund. Organization of architectures for cognitive vision systems. In H.I Christensen and H.H. Nagel, editors, *Cognitive Vision Systems: Sampling the spectrum of approaches*, pages 37–55. Springer-Verlag, Berlin Heidelberg, Germany, 2006.

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

[41] J. Hedborg, P.-E. Forssén, and M. Felsberg. Fast and accurate structure and motion estimation. In *International Symposium on Visual Computing*, number Volume 5875 in Lecture Notes in Computer Science, pages 211–222. Springer-Verlag, 2009.

[42] D. Hilbert and S. Cohn-Vossen. *Geometry and the imagination*. Chelsea Publishing Company, New York, 1952.

[43] M. Hu. Visual Pattern Recognition by Moment Invariants. *IRE Transactions on Information Theory*, IT-8:179–187, 1962.

[44] S. A. Hutchinson, G. D. Hager, and P. I. Corke. A tutorial on visual servo control. *IEEE Transaction on Robotics and Automation*, 12(5):651–670, 1996.

[45] M. Isard and A. Blake. CONDENSATION – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[46] B. Johansson. *Low Level Operations and Learning in Computer Vision*. PhD thesis, Linköping University, Sweden, SE-581 83 Linköping, Sweden, December 2004. Dissertation No. 912, ISBN 91-85295-93-0.

[47] E. Jonsson. *Channel-Coded Feature Maps for Computer Vision and Machine Learning*. PhD thesis, Linköping University, Sweden, SE-581 83 Linköping, Sweden, February 2008. Dissertation No. 1160, ISBN 978-91-7393-988-1.

[48] F. Kahl and A. Heyden. Using Conic Correspondences in Two Images to Estimate the Epipolar Geometry. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1998.

[49] R. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[50] K. Kanatani. *Geometric computation for machine vision*. Oxford University Press, Inc., 1993.

[51] J. Kannala, M. Salo, and J. Heikkilä. Algorithms for computing a planar homography from conics in correspondence. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2006.

[52] A. Khotanzad and Y. Hong. Invariant Image Recognition by Zernike Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, 1990.

[53] D. Kragic and H. I. Christensen. Survey on visual servoing for manipulation. Technical report, ISRN KTH/NA/P–02/01–SE, Jan. 2002., CVAP259, 2002.

[54] F. Larsson. *Visual Servoing Based on Learned Inverse Kinematics.* M.Sc. Thesis LITH-ISY-EX–07/3929, Linköping University, 2007.

[55] F. Larsson. *Methods for Visually Guided Robotic Systems: Matching, Tracking and Servoing.* Linköping Studies in Science and Technology. Thesis No. 1416, Linköping University, 2009.

[56] F. Larsson. Automatic 3D Model Construction for Turn-Table Sequences - A Simplification. LiTH-ISY-R, 3022, Linköping University, Department of Electrical Engineering, 2011.

[57] F. Larsson and M. Felsberg. Traffic sign recognition using Fourier descriptors and spatial models. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2011.

[58] F. Larsson and M. Felsberg. Using Fourier Descriptors and Spatial Models for Traffic Sign Recognition. In *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, volume 6688 of *Lecture Notes in Computer Science*, pages 238–249, 2011.

[59] F. Larsson, M. Felsberg, and P.-E. Forssén. Correlating Fourier descriptors of local patches for road sign recognition. *IET Computer Vision*, 5(4):244–254, 2011.

[60] F. Larsson, M. Felsberg, and P-E. Forssén. Patch contour matching by correlating Fourier descriptors. In *Digital Image Computing: Techniques and Applications (DICTA)*, Melbourne, Australia, December 2009. IEEE Computer Society.

[61] F. Larsson, P-E. Forssén, and M. Felsberg. Using Fourier descriptors for local region matching. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2009.

[62] F. Larsson, E. Jonsson, and M. Felsberg. Visual servoing based on learned inverse kinematics. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2007.

[63] F. Larsson, E. Jonsson, and M. Felsberg. Visual servoing for floppy robots using LWPR. In *Workshop on Robotics and Mathematics (ROBOMAT)*, pages 225–230, 2007.

[64] F. Larsson, E. Jonsson, and M. Felsberg. Learning floppy robot control. In *Proceedings of the Swedish Symposium on Image Analysis (SSBA)*, 2008.

[65] F. Larsson, E. Jonsson, and M. Felsberg. Simultaneously learning to recognize and control a low-cost robotic arm. *Image and Vision Computing*, 27(11):1729–1739, 2009.

[66] J. W. Lasley. On degenerate conics. *The American Mathematical Monthly*, 64(5):362–364, 1957.

[67] J. D. Lawrence. *A Catalog of Special Plane Curves*. Dove Publications, Inc., 1972.

[68] M. Leordeanu, M. Hebert, and R. Sukthankar. Beyond Local Appearance: Category Recognition from Pairwise Interactions of Simple Features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[69] S. Loncaric. A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001, 1998.

[70] G. Lu and A. Sajjanhar. Region-based shape representation and similarity measure suitable for content-based image retrieval. *Multimedia Systems*, 7(2):165–174, 1999.

[71] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1981.

[72] R. Mahler. Multitarget Bayes filtering via first-order multitarget moments. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4):1152–1178, 2003.

[73] F. Mokhtarian, S. Abbasi, and J. Kittler. Robust and efficient shape indexing through curvature scale space. In *Proceedings of the British Machine Vision Conference (BMVC)*, 1996.

[74] V. Pavlovic, J.M. Rehg, T.J. Cham, and K.P. Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1999.

[75] N. Pugeault and R. Bowden. Driving me around the bend: Learning to drive from visual gist. In *Proceedings of the 1st IEEE Workshop on Challenges and Opportunities in Robot Perception, in parallel to the IEEE International Conference on Computer Vision (ICCV)*, 2011.

[76] B. Shalom and E. Tse. Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11(5):451–460, 1975.

[77] J. Shotton, A. Blake, and R. Cipolla. Contour-based learning for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2005.

[78] K. Siddiqi, A. Shokoufandeh, S. Dickinson, and S. Zucker. Shock Graphs and Shape Matching. *International Journal of Computer Vision*, 35(1):13–32, 1999.

[79] P. Srestasathiern and A. Yilmaz. Planar Shape Representation and Matching Under Projective Transformation. *Computer Vision and Image Understanding*, In press, 2011.

[80] R. L. Streit and T. E. Luginbuhl. Probabilistic multi-hypothesis tracking. Technical report, 10, NUWC-NPT, 1995.

[81] A. Sugimoto. A Linear Algorithm for Computing the Homography from Conics in Correspondence. *Journal of Mathematical Imaging and Vision*, 13(2):115–130, 2000.

[82] M. R. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America (1917-1983)*, 70(8):920–930, 1980.

[83] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision*, 48(1):9–19, 2002.

[84] United Nations General Assembly. *Improving global road safety, A/RES/64/255 (2010). Resolution of the United Nations General Assembly, 64th session.* 2010.

[85] World Health Organization. *Global status report on road safety: time for action.* 2009.

[86] M. Yang, K. Kpalma, and J. Ronsin. A Survey of Shape Feature Extraction Techniques. In *Pattern Recognition Techniques, Technology and Applications*, pages 978–953. IN-TECH, 2008.

[87] X. Yang, S. Koknar-Tezel, and L. Latecki. Locally constrained diffusion process on locally densified distance spaces with applications to shape retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[88] C.T. Zahn and R.Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, C-21(3):269–281, 1972.

[89] D. Zhang and G. Lu. Generic Fourier descriptor for shape-based image retrieval. In *Proceedings on the IEEE International Conference on Multimedia and Expo*, 2002.

[90] D. Zhang and G. Lu. A Comparative Study of Curvature Scale Space and Fourier Descriptors for Shape-based Image Retrieval. *Journal of Visual Communication and Image Representation*, 14(1):39–57, 2003.

[91] D. Zhang and G. Lu. Review of shape representation and description techniques. *Pattern Recognition*, 37(1):1 – 19, 2004.