Lehrstuhl für Steuerungs- und Regelungstechnik
Technische Universität München
Univ.-Prof. Dr.-Ing./Univ. Tokio Martin Buss

Aspects of Multi-Focal Vision

Kolja Kühnlenz

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. sc. techn. (ETH) Andreas Herkersdorf

Prüfer der Dissertation:

1. Univ.-Prof. Dr.-Ing./Univ. Tokio Martin Buss
2. Univ.-Prof. Dr.-Ing. Georg Färber

Die Dissertation wurde am 18.12.2006 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 17.04.2007 angenommen.

Preface

This thesis has emerged from three years of work at the Institute of Automatic Control Engineering, Technische Universität München where I stayed from 2003 until now.

I would like to gratefully thank my doctoral advisor Prof. Martin Buss, for all the inspiration, encouragement, and invaluable freedom of research, for pushing me to my intellectual, mental, and physical limits, and for all his extensive support to all of us his students — and for making me conscious of even the smallest step being a step.

My deepest thanks to my colleague and friend Dr. Sandra Hirche (Fujita Lab, Tokyo Institute of Technology). I cannot count all the ups and downs we shared. Grateful thanks to Dr. Nicholas Gans (Nonlinear Controls and Robotics, University of Florida) and Prof. Seth Hutchinson (University of Illinois, Urbana-Champaign) for their visual servo code for performance comparisons and for their hospitality. To my former colleague Dr. Javier Fernandez-Seara (The Boston Consulting Group) and to Klaus Strobl-Diestro (Institute of Robotics and Mechatronics, Deutsches Zentrum für Luft- und Raumfahrt (DLR)) for their simulation code. To all my students, in particular, Tingting Xu, Stefan Sosnowski, Florian Laquai, Jens Hölldampf, Wei Xiong Cheng, and Kiran Sivakumar, for all your extraordinary efforts. To all my colleagues, for your support and discussions, for taking over duties, making experiments work, and for making me always feel welcome in spite of my continuous absence from TU Mensa. To all my friends, thank you for your continuous empathy and patience.

In deep gratitude I want to thank my parents, for all your patience and encouragement, and for your love.

Munich, 2006. Kolja Kühnlenz

For my parents

Abstract

Vision has become a powerful tool in a variety of technological and scientific areas such as measurement, automation, and robotics. The visual channel provides detailed non-contact measurements of geometry, dynamics, and texture of object and environmental structures. Due to limitations of sensory and computational resources vision faces a trade-off between field of view and accuracy. Moreover, a vision system providing both large field of view and high accuracy would not only require considerable resources, but produce a vast amount of unnecessary data resulting in a low density of usable information.

Multi-focal vision overcomes these drawbacks providing several sensor devices which differ in accuracy and field of view. Well-known embodiments are foveated systems inspired by the human eye. These consist of a low-accuracy vision device with large field of view and a coaxially mounted high-accuracy vision device with small field of view. To date, only few works are known exploiting the particular nature of multi-focal vision. Comparative evaluations quantifying the benefits of multi-focal vision are not known.

This thesis focuses on the investigation of multi-focal vision for measurement and robotics applications. The approach considers three different abstraction levels covering static, dynamic, and planning issues. Light is shed on the configuration dependent measurement performance of multi-focal vision systems. So far unique are multi-focal vision-based control strategies of robot manipulators and active vision approaches for mobile robot navigation. Advantages are significant improvements of control performance and localization accuracy as well as an extension of the workspace compared to conventional approaches. Flexibility of multi-focal resource allocation facilitates various scenarios which are not realizable with conventional vision. Examples are whole scene observation while assuring a certain control or localization performance over the entire workspace.

In this work the benefits of multi-focal vision are quantified in comparative evaluation studies for the first time. The contributions provide fundamental insight in the multifaceted concept of multi-focal vision and serve as a signpost for future research.

Zusammenfassung

Maschinelles Sehen gewinnt zunehmend an Bedeutung in vielen technischen und wissenschaftlichen Bereichen, wie z.B. in Messtechnik, Automation und Robotik. Der visuelle Kanal ermöglicht kontaktlose Messungen von Geometrie, Dynamik und Textur von Objekten und Umgebung. Sichtfeld und Genauigkeit bilden aufgrund der Leistungsgrenzen von Sensorik und Rechenkapazität gegensätzliche Anforderungen, welche stets einen Kompromiss bedingen. Sichtsysteme, welche beide Anforderungen gleichermassen befriedigen, würden darüberhinaus ein hohes Maß unnötiger Daten und geringe Nutzinformationsdichten erzeugen.

Multifokales Sehen bietet mehrere optische Sensoren, welche sich in Sichtfeld und Genauigkeit unterscheiden. Hierdurch werden die Nachteile konventioneller Sichtsysteme überwunden. Typische Ausführungen orientieren sich an dem fovealen und peripheren Sehen des menschlichen Auges. Sie bestehen aus einer coaxialen Anordnung gering auflösender Weitwinkelsensoren und hochauflösender Schmalwinkelsensoren. Die besonderen Eigenschaften multifokalen Sehens werden bis heute nur in wenigen bekannten Arbeiten gezielt genutzt. Vergleichende quantitative Untersuchungen der Performanz multifokalen Sehens sind nicht bekannt.

Diese Dissertation befasst sich mit der Erforschung multifokalen Sehens für Messtechnik und Robotik. Statische, dynamische sowie planerische Aspekte bilden drei Abstraktionsebenen des Ansatzes. Die Performanz multifokaler Sichtsysteme wird abhängig von den Sensorkonfigurationen untersucht. Die in dieser Arbeit vorgestellten bildbasierten Strategien zur Regelung von Robotikmanipulatoren und Aufmerksamkeits-Steuerungen zur Navigation mobiler Roboter sind bislang einzigartig in der Literatur. Wesentliche Vorteile sind signifikante Steigerungen von Regelungsperformanz und Lokalisierungsgenauigkeit sowie erweiterte Arbeitsräume im Vergleich zu konventionellen Methoden aus dem Schrifttum. Erst die flexible Zuweisung multifokaler Sensorressourcen ermöglicht einige spezielle Anwendungen. Beispiele sind eine Überwachung der gesamten Umgebung bei gleichzeitiger Gewährleistung einer definierten Regelungs- oder Lokalisierungsperformanz.

In dieser Arbeit werden erstmalig die Vorzüge multifokalen Sehens in vergleichenden Untersuchungen quantifiziert. Die Beiträge liefern grundlegende Einblicke in das facettenreiche Konzept und bilden einen Wegweiser für die zukünftige Forschung.

Contents

Notations

Abbreviations

DOF	degrees of freedom
EKF	extended Kalman-filter
FOV	field of view
SC	switching condition

Conventions

Scalars, Vectors, and Matrices

x	vector unless declared otherwise
x_i	sub-vector or scalar
X	matrix or scalar
$f(\cdot)$	vector function
$\dot{x}$, $\ddot{x}$	equivalent to $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}$ and $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{x}$
$\overline{x}$	mean of x
$\hat{x}$	estimated or predicted values
$\|\cdot\|$	Euclidian norm

Subscripts and Superscripts

x^d	desired value of x, set value for the control loop
$x_{\max}$	maximum value of x
$x_{\min}$	minimum value of x
$(\cdot)^{-1}$	inverse
$(\cdot)^{+}$	pseudo-inverse
$(\cdot)^{T}$	transposed
$(\cdot)^{*}$	optimal or expected value
$(\cdot)_{pos}$	position
$(\cdot)_{pose}$	pose
$(\cdot)_{tran}$	translation
$(\cdot)_{rot}$	rotation
$(\cdot)_{mono}$	mono-focal
$(\cdot)_{multi}$	multi-focal

Geometric Aspects of Multi-Focal Vision

i, j	feature point and sensor i, j
λ	focal-length
ξ	vector of all feature points in image space
ξ_i	vector of feature point i
ξ_u, ξ_v	feature point components in u- and v-direction
S_M	memory frame
S_I	image frame
S_c	camera frame
S_r	vision system frame
${}_rx$, ${}_cx$	Cartesian point
x	vision system pose
X, Y, Z	Cartesian point coordinates
s	scalar multiplier
P	perception matrix
p	element or vector of perception matrix
${}^rR_{c,i}$	rotation matrix
r	components of rotation matrix
${}^rT_{c,i}$	homogeneous transformation matrix
t_c	translation vector
κ	distortion coefficient
E	sensor noise covariance matrix in Cartesian space
E_{uv}	sensor noise covariance matrix in image space
F	auxiliary matrix
b	auxiliary vector, stereo baseline
q	auxiliary vector
e	error vector
$\Delta(\cdot)$	error
g	function
J_{uv}	Jacobian of g
h_ξ	sensor
$\mathcal{H}_n$	manifold of sensors of rank n
J_v	visual Jacobian
R	rotation matrix
${}_0x$	vision system pose
m	number of Cartesian degrees of freedom
n	number of cameras or feature points
α, β, γ	yaw-, pitch-, roll-angles
d	stereo disparity
A	matrix
U	matrix containing the eigenvectors of JJ^T
V	matrix containing the eigenvectors of J^TJ
Σ	matrix containing square roots of eigenvalues of J^TJ
σ	singular value
u	element or vector of U
v	element or vector of V
c	condition number

w	perceptibility
Δ	difference, range
$(\cdot)_s$	sensitivity
$(\cdot)_p$	perceptibility
$(\cdot)_c$	condition
Ψ	objective function
r	vector
c_r	scalar, ratio
Q	objective function
h, l	high-, low-sensitivity
${}_r s$	sensitivity in r-direction
s_0	smallest sensitivity in a range
c_0	smallest condition number in a range

Multi-Focal Control of Robot Manipulators

t	continuous time
i, j	feature point and sensor i, j
m	number of degrees of freedom in image space
n	number of Cartesian degrees of freedom
λ	focal-length
ξ	vector of all feature points in image space
ξ_i	vector of feature point i
ξ_u, ξ_v	feature point components in u- and v-direction
q	joint angle positions
$\tilde{\xi}$	control error in image space
ξ^d	desired feature point positions in image space
x, x_{pose}	end-effector pose
M	manipulator inertia matrix
$C\dot{q}$	centripetal and Coriolis torques
g	gravitational torques
τ	joint torques
J_v	visual Jacobian
h_ξ	sensor
R	rotation matrix
K_p, K_v	control gains
f_s	nonlinear system dynamics
e_{pos}	average remaining position error
e_{rot}	average remaining rotation error
e_{pose}	pose error
σ	standard deviation
$\sigma_{e,z}$	standard deviation in z-direction
σ^2	variance, noise power
f^η	hybrid switching controller
J_v^η	selected visual Jacobian in hybrid controller
h_ξ^η	selected sensor
$\mathcal{J}_v$	set of visual Jacobians
$\mathcal{J}^m$	manifold of visual Jacobians of rank m
$\mathcal{H}_\xi$	set of sensors
$\mathcal{H}^m$	manifold of sensors with rank m
η	discrete control input
V	Lyapunov function
$[\cdot, \cdot]$	Lie bracket
Σ_x	performance region in Cartesian space
Σ_x^0	desired performance
Σ_x^*	expected performance
$< \cdot, \cdot >$	tuple
Ψ	side-condition, e.g. field of view, resolution
σ_x^0	desired performance band in Cartesian space
σ^*	expected performance
w	perceptibility
$\sigma_{trans,z}$	translational pose error variance in z-direction

v	threshold
α	aperture angle
$(\cdot)^{multi}$	multi-sensor, multiple types of sensors
$(\cdot)^{single}$	single-sensor, one type of sensor
s	singular value
v	eigenvector of row-space of $A^T A$

Multi-Focal View Direction Planning for Mobile Robots

x_k	discrete time variable at step k
$(\cdot)_\xi$	feature point in image space
S_F	robot foot frame
S_0	world frame
S_r	vision system reference frame
$(\cdot)_s$, $(\cdot)^s$	footstep
${}_0x$	robot foot pose
${}_F x_s \ {}_F y_s]^T$	commanded footstep position
${}_F\theta_s$	commanded footstep orientation
u	control vector
γ	binary variable
v_ξ	sensor noise
w	dead-reckoning error
$\Delta(\cdot)$	error
ξ	vector of all feature points in image space
ξ_i	vector of feature point i
h_ξ	sensor
${}_0l$	vector of landmarks
$x_{l,i}$, $y_{l,i}$, $z_{l,i}$	landmark i's position components
${}_F x_l$	vector of landmarks
${}_F P$	perception matrix
$(\cdot)_{ci}$	camera i
${}^F T_{ci}$	homogeneous transformation matrix
E	sensor noise covariance matrix in Cartesian space
F	auxiliary matrix
z	measurements
$(\cdot)_m$	measurement
$(\cdot)_e$	environment
Rot	rotation matrix
Q	process noise covariance matrices
R	measurement noise covariance matrix
$(\cdot)^{lin}$	linearized
A, B, W, H, V	partial derivative matrices
C	estimated covariance matrix in Kalman-filter
σ^2	variance
K	Kalman-gain
S	innovation covariance
v_0^s	robot position incertitude
e	eigenvalue
δ_0	binary variable
ε_0	binary variable
$(\cdot)_{pan}$, $(\cdot)_{tilt}$	pan, tilt
$\alpha_{pan,tilt}$	vector containing pan- and tilt-angles
T_α	point set, field of view
C_p	point set, confidence ellipsoid

X, Y, Z	components of Cartesian point in vision system frame
p	confidence level
Ω	vector containing view directions of all vision devices
Ω_j	vector containing the pan-/tilt-angles of device j
r	interest operator
A_{90}	area of 90%-confidence ellipse
N_{vis}	average number of visible landmarks

List of Figures

List of Tables

1 Introduction

Vision provides a substantial source of information and has become a key aspect in a variety of research areas such as measurement, surveillance, automation, and robotics. One of the main reasons for the increased interest of the different communities during the last three decades are advances in sensor and semiconductor technology significantly improving measurement accuracy and computational power of commercially available off-the-shelf vision products.

Vision and vision systems with three and more vision devices are state-of-the-art and systems are commercially available in a variety of embodiments. Simultaneous utilization of a multitude of vision sensors is an important topic in computer vision and robotics. The fusion of visual information of a multi-camera system significantly increases observation range, measurement accuracy, and robustness. Multi-camera vision has been an intensive research field since the last decade and systems are widely-used.

A special area in multi-camera vision is vision with several vision devices which differ in field of view and accuracy. This is called *multi-focal* vision in the remainder of this thesis. The relative poses of the individual vision devices may or may not be independently controllable and the intrinsic parameters, e.g. focal-length, may be variable. Multi-focal vision is originally inspired by the organ of sight of the biological paradigm 'human'. The retina of the human eye, e.g., is covered by different types and distributions of photoreceptors which result in a small region of high accuracy near the optical axis (fovea centralis) and a large region of lower accuracy towards the borders of the retina (peripheral vision). *Foveated vision* is widely accepted as a synonym, however, originated from biology it refers strictly speaking to a variation of measurement accuracy over the visible field with one foveal and one peripheral region and does not cover changes of the relative poses of the vision device. Main advantages of multi-focal vision are a partial examination with high accuracy keeping a large part of the environment in view. By this flexible allocation of perceptual resources the density of information of interest in the visual data stream is increased and computational cost reduced as resources are only allocated to an extend really needed in the current situation. Most multi-focal systems and methods in known literature cover configurations with fixed relative poses of the individual sensors. Only few approaches consider the combination of sensors with variable relative poses. Most prominent examples are vision systems for autonomous vehicles, surveillance installations, and humanoid robot heads.

An embodiment of a multi-focal vision system which has been developed at the Institute of Automatic Control Engineering (LSR), Technische Universität München, is shown in Figure 1.1 [71]. This system comprises several vision sensors with different accuracies and fields of view integrating wide-angle and telephoto lenses which is typical for multi-focal systems. Their fields of view may or may not overlay and in contrast to most common systems their relative poses are variable. Given such a flexible vision system a manifold of potential applications are imaginable, such as acquisition of environmental data, vision-

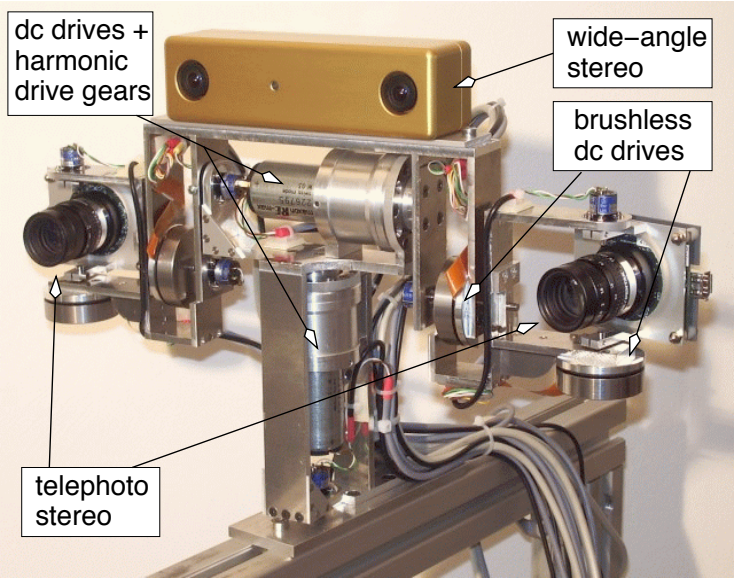

Figure 1.1: Multi-focal vision system [71].

based control of robot manipulators, and vision guided mobile robots just to mention a few.

Fundamental questions on overall performance, sensor selection, and optimal configuration arise which have not yet been considered in known literature. A variety of tools for multi-sensor data fusion of general vision devices which may differ in performance characteristics are known. However, the measurement performance of general multi-focal vision systems has not been investigated yet. Given a particular task or situation, another issue is control of the individual vision sensors in an efficient and intelligent way. Some vision-based control approaches are generic in order to allow the use of multi-focal systems, however, neither the impact of multi-focal vision on system dynamics nor the potentially dynamical (re-)configuration of the vision system has been considered yet. There are approaches to active vision using multi-focal systems, however, the only known application to vision-guided mobile systems is the guidance of automobiles. All other works either do not consider the system locomotion or do not explicitely exploit multi-focal vision.

The common main deficiencies of mono-focal approaches to vision systems, vision-based control, and active vision are the contradictory parameters field of view and measurement accuracy due to limitations of resolution and size of the vision sensor. In consequence a trade-off between workspace and task performance exists. In vision-based control the result is a limitation of the operating range of the manipulator due to field of view, i.e. visibility of the manipulator or the observed reference object, and control performance, i.e. increasing pose error and pose error variance with distance between sensor and target. In vision-guided mobile robotics this trade-off results in either less visible reference objects for localization or lower measurement accuracy.

The work presented in this thesis focuses on the investigation of multi-focal approaches for vision-based manipulator control and active vision of mobile vision-guided robots in order to overcome the mentioned drawbacks.

The main challenges faced by the design and control of multi-focal vision systems in visual perception, robot control, and active vision are summarized in the following.

1.1 Challenges

The design and control of vision systems face multiple challenges in the intersection of the fields computer vision, robot control, and mobile robots covering system configuration aspects, dynamical issues, and task and situation related coordination of the individual vision sensors. Some of the key issues targeted in this thesis are summarized as follows:

Design Issues

A major drawback of common vision systems are the contradictory requirements of measurement accuracy and field of view due to limitations of resolution and size of the vision sensors, and of computational resources. In consequence either higher computational cost has to be accepted in order to process the larger amount data of vision devices providing both, large field of view *and* high accuracy, or a smaller field of view, respectively, a lower accuracy has to be accepted in order to satisfy computational resources limitations. Another aspect is the potentially low density of task relevant information in the data stream processing the data of a large field of view and high accuracy system. Zooming cameras fail in those cases where large field of view and high accuracy have to be provided simultaneously or fast configuration changes are necessary, and require more complex modeling and calibration. The mentioned drawbacks can be overcome utilizing multi-focal vision providing several vision devices with different performance characteristics, e.g. with various fields of view and accuracies, which can selectively be allocated by changing the intrinsic and/or extrinsic configuration of the vision system.

A main challenge addressed in this thesis is the investigation of the measurement performance of general multi-focal vision systems with variations of the relative poses and of the focal-lengths of the individual vision devices. Another aspect addressed are deliberate configuration changes in order to achieve a particular measurement performance of the vision system.

Control Issues

Vision-based control of robot manipulators provides accurate free-space motion in weakly structured environments. Only local knowledge of the environmental structure is needed to control the trajectory of the robot end-effector. However, control performance in terms of pose error and pose error variance, and stability are strongly dependent on the measurement accuracy of the vision system and on the distance to the observed reference object. Conventional vision-based control methods cannot guarantee constant control quality over a wide spatial operating range and even stability. Zooming cameras are a considerable means, however, cannot provide high accuracy and wide field of view at the same time, modeling and calibration complexity is significantly increased, and focus adjustments limit the dynamic range.

Main challenges focused on are the investigation of multi-focal vision-based control strategies and the dynamical adaptation of vision system configurations in order to satisfy given control performance requirements.

Planning Issues

Mobile vision guided robots localize themselves and plan their paths assessing visual information. This information is fused with internal sensor data in order to exploit the benefits of the absolute nature of visual measurements and the relative nature of odometry. Vision sensors with high accuracy provide better measurements, but on the one hand the local

surroundings are not perceived due to the strongly limited field of view and on the other hand reference objects for localization along the robot's path cannot be detected if their positions are not known a priori. Zooming cameras can switch between high accuracy and wide field of view, however, cannot provide both at the same time. Thus, it is not possible to detect potentially interesting or even dangerous objects or events outside of the zoomed region.

The main challenge addressed are multi-focal view direction planning strategies for mobile robots with active vision in combination with multi-focal robot localization in order to improve localization accuracy and reactivity due to enhanced perceptual capabilities. An aspect is to exploit the flexibility of sensor resource allocation of multi-focal vision with independent pose control of the individual vision devices.

1.2 Main Contributions and Outline of the Thesis

The presented work focuses on the design of multi-focal vision systems and multi-focal methods for robot manipulator control and for vision-guided mobile robots with active vision. For efficient and goal-oriented action and reaction of an entity sufficient information on the surrounding environment has to be perceived. Multi-focal vision is a flexible and powerful means to improve the performance of visual perception as it provides both, high measurement accuracy and keeping a large part of the scene in view. Large workspaces are covered, geometrical structures are acquired and localized with high certainty, and due to a multitude of sensors high robustness and consistency of environmental knowledge are given. The individual selection of the performance of perception for different parts of the environment improves efficiency by increasing the density of usable information and reducing the amount of unnecessary data.

In order to exploit the beneficial characteristics of multi-focal vision it is essential to understand the fundamental issues determining the performance of such a system. How to combine the different types of sensors in a selective and geometrical way if particular task or situation specific and performance requirements have to be met. *Geometric aspects* constitute the fundamentals of multi-focal perception of the environment and provide an essential tool for multi-focal manipulator and active vision control. Multi-focal *manipulator control* overcomes the common drawbacks of vision-based control. Dependency of control performance on operating distance and stability margins are significantly improved due to selectively increasing accuracy and field of view. For vision-guided mobile robots a task and situation related *planning* of the *view direction* of a multi-focal active vision system is provided in order to filter out usable information on the environment and ensure good quality of perception.

These main aspects of multi-focal vision are addressed in this thesis. Representing orthogonal research directions by themselves these aspects constitute a hierarchical approach as the level of information abstraction increases from the basic perception *of* the environment to intelligent action *within* the environment accounting for the particular benefits of multi-focal vision and give the structure to this thesis as highlighted in Figure 1.2. The main contributions of this work are presented in the following.

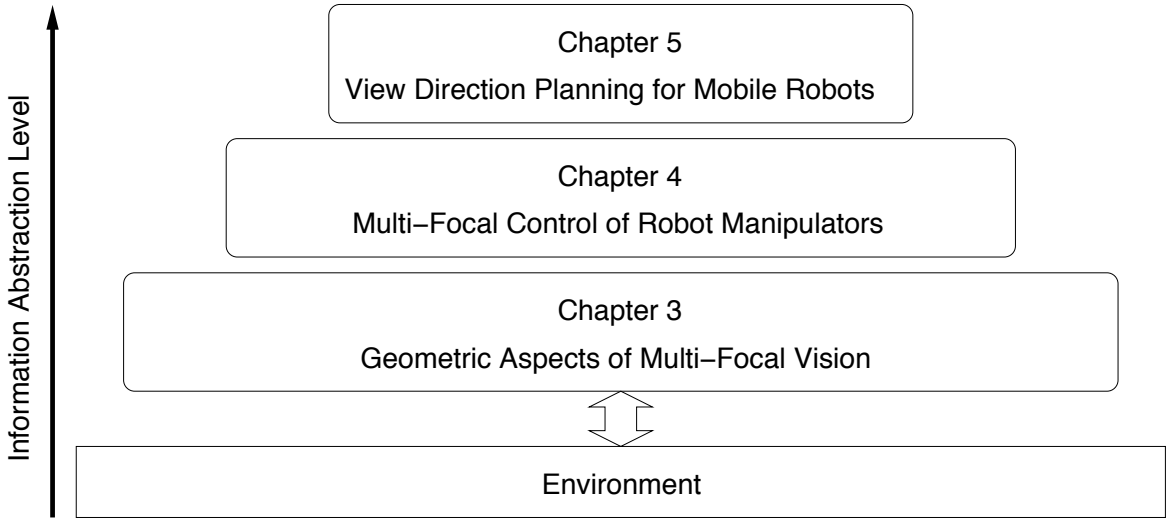

Figure 1.2: Outline of the thesis.

Geometric Aspects - System Design and Performance of Perception

Multi-focal vision - the combination of sensors with high measurement accuracies but small fields of view and sensors with large fields of view but low accuracy - provides a solution to the problem statement in Section 1.1. High measurement accuracy, wide field of view, and high density of usable information in the visual data stream are provided. Yet, the issue *which* sensors to choose, *how* to combine them, and what is the expected performance of the complete system has not been addressed in known literature. The performance of the complete system is composed of the performance properties of the individual sensors and the combined use of the individual sensors results in potentially different performance properties if system configurations are altered. Inaccurate knowledge of the system properties potentially results in poor measurement accuracy, control performance and may even render vision-based control unstable. The assessment of the performance of the multi-focal vision system taking geometrical and sensor specific constraints into account is an essential issue in order to fuse the visual data optimally, to evaluate the accuracy of the complete system, and to facilitate optimal design and controlled configuration changes according to task and situational requirements. In Chapter 3 these aspects are addressed in a systematical investigation providing fundamental insight for design and application of multi-focal vision systems.

Manipulator Control - Multi-Focal Vision-Based Robot Control

While only static issues and configurations are considered in Chapter 3 this research direction focuses on dynamic aspects of multi-focal vision and its impact on control of robot manipulators. Sensor characteristics have a direct impact on control performance and stability. Due to the dependency of measurement accuracy on distance to the observed environmental structures control performance degrades with distance significantly limiting workspace and control is potentially even rendered unstable. The limited image plane further reduces workspace. In this thesis manipulator control based on multi-focal vision is investigated for the first time. In Chapter 4 a switching approach to vision-based control is proposed comprising instantaneous configuration changes of the multi-focal system in order to satisfy performance and task specific constraints. A selective improvement

of control quality and field of view is provided facilitating large workspaces and largely distance independent performance. Another innovation is vision-based control with selective observation of individual object features with high-accuracy sensors in addition to a large field of view sensor for scene overview. Thereby, a significant improvement of control performance is achieved.

View Direction Planning - Control of Active Vision Systems for Mobile Robots
Chapter 5 addresses multi-focal vision from a higher-level information specific perspective. Aspects of task and situation dependent coordination of multi-focal active vision devices for visually guided mobile robots are investigated. Focused are methods to control the view directions of the individual vision devices with different characteristics in order to increase localization accuracy, to cope with large environments and sparsely distributed reference objects, and to flexibly respond to events of potential interest trading those versus the primary mission – following a planned path. Foveated state-of-the-art vision systems and multi-focal systems with independent active vision devices are considered in this thesis for the first time in the context of active vision for mobile robots with missions requiring locomotion. Major benefits are significantly improved localization accuracies and flexibility of reactive behavior.

The aspects addressed in this thesis contribute to a fundamental understanding of multi-focal vision as an integrated concept. Although, a variety of multi-focal vision systems exists only few methodical approaches are known exploiting their particular nature. It is the aim of this work to bring together and integrate very different facets of the concept of multi-focal vision in order to act as a guidepost and source of inspiration for future research in this field. A variety of applications and examples are selected to underline the integrated and multifaceted character of multi-focal vision.

2 State-of-the-Art Vision Systems, Control, and Planning

Multi-focal vision is a relatively young field of research at the intersection of a multitude of different areas. Originated as an analogy to the vision system of vertebrates earliest related works are concerned with the human vision system and its foveated control mechanisms from anatomical, psychological, and neuroscience perspectives. For more than one century mechanisms of human oculomotor control and visual attention have been an intensively investigated topic in psychology and neuroscience research [2, 3, 10, 16, 40, 44, 68, 91, 98, 104, 110, 128, 134, 137]. An important aspect in this context are intentionally and stimuli triggered fast ballistic gaze shifts (saccades) based on the evaluation of peripheral vision and mechanisms of *preattentive selection*, e.g. cf. [3, 10, 68, 98, 107, 110, 128, 134]. To date, the importance of these research areas has even increased as in the eighties the first computational and implementable attention models to direct the fovea have been developed. Transfer research towards technically implementable models in foveated vision particularly covers computational neuroscience models of attentional mechanisms, e.g. [35, 38, 62, 68, 126, 136], and control models, e.g. [19, 107, 119].

In engineering and information sciences some of the closest related research fields are multiple view geometry, multi- and spatially varying resolution vision, optimal sensor placement, control aspects (visual servoing) taking spatially varying resolution and limitations of the visible field into account, gaze control, i.e. higher-level camera coordination mechanisms, vision guided robotics, multi-sensor data fusion, and many more. Apparently, this spectrum is far too manifold to be discussed here completely. First works concerned with multi-focal robot vision as defined in this thesis are noted in the nineties [14, 27, 48, 111]. To date no survey articles are known giving an overview on challenges and known approaches.

Selected fields closest related to the topics investigated in this thesis are multi-resolution vision systems, multi-focal vision-based control and selection methods, vision-based control with particular emphasis on multi-camera, multi-resolution, planning, sensor selection and placement approaches, and performance issues, and view direction selection and control in vision guided robotics. These are surveyed in the following.

2.1 Multi-Resolution Vision Systems

Vision systems with spatially varying resolutions are known in manifold embodiments. The underlying technological principles reach from multiple-lens optics, e.g. [80], over multi-camera, e.g. [37], and mirror systems, e.g. [63] to custom sensor designs [129] and image processing methods, e.g. [132]. Vision devices are commonly mounted on kinematic platforms comprising 2 to 7 degrees of freedom (DOF) corresponding to basic pan/tilt-platforms and platforms with additional 2 DOF per "eye", respectively. *Multi-focal* vision systems comprise more than one vision sensor, usually a camera device, with different

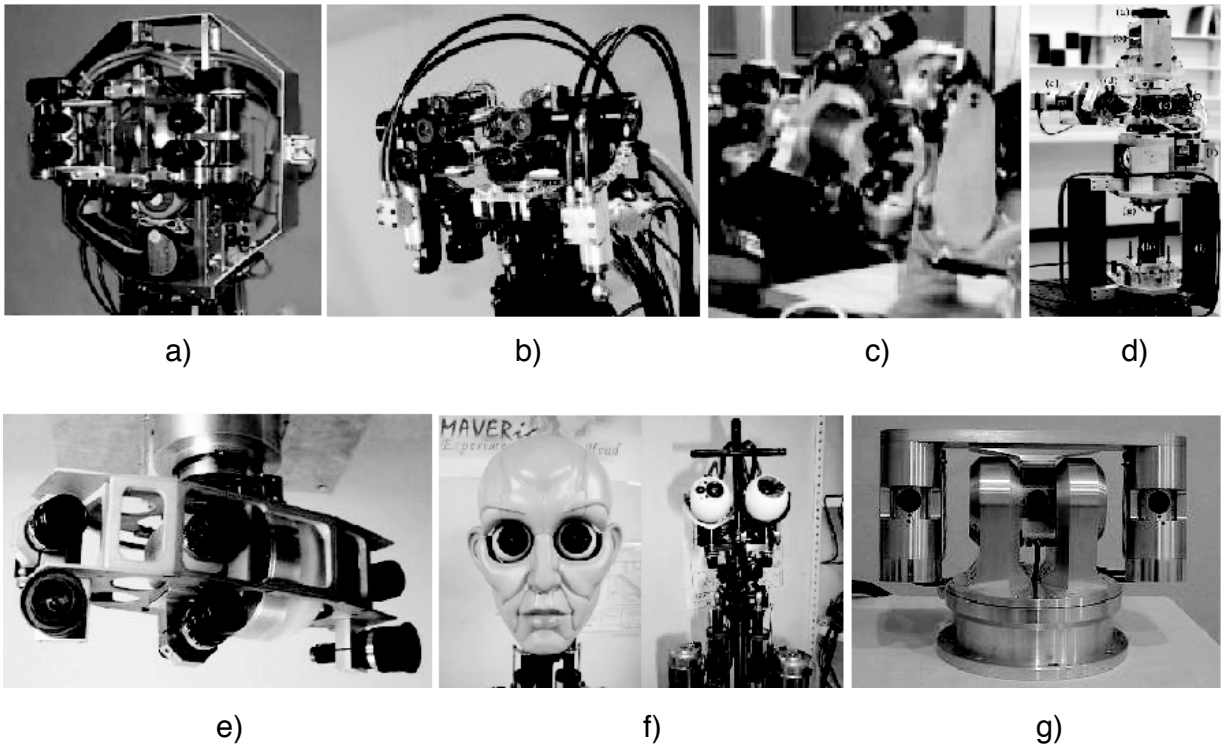

Figure 2.1: Selected multi-focal vision systems; a) Cog [14]; b) DB head [127]; c) Macaco [7]; d) surveillance system [63]; e) MarVEye4 [114]; f) Maveric [131]; g) Triclops [48].

accuracies and fields of view covering mainly robot heads, ground vehicles, and surveillance systems.

A variety of multi-focal robot vision systems are known, e.g. [5, 7, 12–14, 31, 48, 94, 111, 127]. Most prominent examples are the systems of MIT, USA, e.g. Cog [14] (see Figure 2.1a), Macaco [7] (see Figure 2.1c) (Artificial Intelligence Lab), and Kismet [13] (Media Lab) with coupled tilt of both cameras and 6 DOF, and the vision system of the humanoid DB [127] (see Figure 2.1b) built by SARCOS, USA, in cooperation with RIKEN Brain Science Institute, Japan, and the Computational Learning and Motor Control Lab, University of Southern California, with 7 DOF. Each of these comprises two cameras per eye with wide-angle and telephoto lenses, respectively. The relative poses of these two cameras of Cog and Macaco are fixed, while Kismet's wide-angle stereo camera is mounted on the central pan/tilt-platform. The wide-angle stereo pair of Leonardo [12] (Media Lab, MIT, USA) is fixed and mounted behind the robot, while the telephoto pair is mounted on the ceiling facing downwards. The Triclops system [48] (see Figure 2.1g) (Intelligent Systems, NIST, USA) consists of one central wide-angle camera and a telephoto stereo camera pair mounted on a 2 DOF platform. Additionally, the pan-angles of the telephoto cameras are controllable. Recently, also the 2 DOF vision system of the humanoid HRP-2 (AIST, Japan) has been equipped with a wide-angle camera in addition to its trinocular rig [31]. These systems mimic the human eye with its foveated and peripheral vision.

Another main field of applications are autonomous ground vehicles and driver assistance systems, e.g. [5, 37, 93, 95, 106, 114]. Those vision systems mostly consist of one telephoto camera and one or more cameras with lower resolution and wide fields of view. The most

Table 2.1: Performance characteristics of selected multi-focal vision systems.

	MarVEye4	DB head	Maveric	Triclops	Cog
DOF	2	7	7	4	6
fields of view [°]	100, 23, 7	100, 24		80, 24	116, 25
velocity, accel.	∼40°/300ms		1875°/s	1830°/s, 63000°/s^2	75000°/s^2

prominent example is the MarVEye [37, 93, 106, 114] (see Figure 2.1e) series of Institut für Systemdynamik und Flugmechanik, Universität der Bundeswehr München, Germany.

Within the field of surveillance, vision systems are integrated or distributed. Embodiments reach from basic combinations of two or more cameras with telephoto and wide-angle lenses, respectively, e.g. [11, 28, 59], over lens-mirror combinations, e.g. [42], to complex multi-focal catadioptric systems, e.g. [63] (see Figure 2.1d).

Selected multi-focal vision systems and performance characteristics are shown in Figure 2.1 and Table 2.1.

2.2 Multi-Focal Methods

Exploiting the particular characteristics of multi-focal vision systems only few strategies are known which mainly make use of peripheral vision to detect regions of interest in order to direct foveated attention. Other methods are concerned with matching of corresponding objects in multi-resolution and wide-baseline stereo views. Selected approaches are discussed in the following.

The related works on gaze control of MarVEye, e.g. cf. [37], in the context of autonomous visual guidance of ground vehicles are probably the most advanced known. In the implemented system a gaze control unit determines the viewing direction of the camera head not only *ad hoc* for the present moment, but also plans and optimizes the viewing behavior for a certain amount of time. Therefore, periods of smooth pursuit are interrupted by quick changes of viewing direction, so-called saccades. Saccades are triggered by visual stimuli, in particular, taking peripheral vision into account or by intention. The underlying principle of the strategy is to predict a varying sequence of actions and then to optimize this sequence according to a heuristical balance of the so-called *physical situation*, *perceptive situation*, and *subjective situation*. The new optimal view direction is obtained by evaluating the predicted information gained modeled with the so-called *knowledge decay* and *knowledge gain* functions. Only some parts of this gaze control have been implemented to present. The definition of information modeling functions has not been published yet.

In [127] a control scheme is proposed to maintain the view of an object in the foveated image using information from the wide-angle view. Geometric relations between these two views with fixed relative pose are studied in order to avoid disparity computation from the stereo cameras.

In [120] the peripheral optical flow field is evaluated. Motion stimuli are modeled by a spatial neural network and a saliency map is generated from the activation dynamics. The strongest stimulus is foveated. After the saccade the control schemes are switched to smooth pursuit in order to keep the object within the field of view. The work of this group in cooperation with CNS, ATR, Japan, concentrates on the biologically inspired integration

of the mechanisms of the human eye covering the vestibuloocular and optokinetical reflexes, smooth pursuit, and bottom-up triggered saccades.

The method proposed in [95] detects candidates for traffic signs in the wide-angle image using color, intensity, and shape information. For each candidate, the telephoto camera is directed to its predicted position to capture the candidate in a larger size in the image.

In [108] the benefit of implementing foveated vision is formulated as an optimization problem, since a trade-off appears between having a small window which would yield small computational delays but tighter control objectives or relaxing the control objectives but obtaining more challenging dynamics. Following this approach, the size of the fovea is chosen as the one giving best tracking capabilities, as measured by the size of the signals which the system is guaranteed to track. A switching controller is proposed to change between saccades and smooth pursuit.

An eye finding algorithm for a foveated active vision system is proposed in [112]. The system uses a motion-based prefilter to identify potential face locations in the peripheral view. These locations are analyzed for faces with a template-based algorithm. Detected faces are tracked in real time, and the active vision system saccades to the face using a learned sensorimotor mapping. Once gaze has been centered on the face, a high-resolution image of the eye can be captured from the foveal camera using a self-calibrated peripheral-to-foveal mapping.

In [59] the geometric and kinematic coupling between a static wide-angle camera and a rotating telephoto camera is analyzed. A solution for this coupling for a general kinematic mechanism and for a simpler pan/tilt model is derived. A unique solution allowing to rotate the camera such that it gazes towards an object within the scene is given. This solution is parameterized by a depth parameter (the distance from the static camera to the object).

An interesting approach directing a high-resolution camera evaluating peripheral vision from a fixed camera in order to actively recognize human behaviors is presented in [27]. Possible actions, i.e. camera fixations, human states, and observables are modeled by a Partially Observable Markov Decision Process (POMDP), which is solved by deriving policies to direct the telephoto view.

A distributed cooperative system of mobile robots is considered in [99] proposing an approach towards cooperative stereo, i.e. wide-baseline stereo. Particularly the key problems, significantly different views and different scales are addressed which are also a fundamental issue in multi-focal vision. An algorithmic approach based on geometric feature descriptor and extensive filterig pipeline is presented. Other approaches in this context are based on sufficiently invariant features, e.g. invariant to rotation and scale [115] and affine transformations [9, 84].

A very recent approach fuses depth images from several vision sensors with different resolutions [96] to generate a 3D surface model of an object. Correspondences between the images are obtained based on similarities of points and their topological information. Each point is assigned a certainty of belonging to the object surface. Points in overlapping areas are integrated based on these two constraints.

Another work proposing a strategy of activity recognition in the peripheral view and deriving control commands for the foveated view is [11].

Except for the works on MarVEye known multi-focal control methods are only exemplarily applied to very limited experimental scenarios and mainly cover very specific problems.

2.3 Vision-based Control

Visual control of robot manipulators, commonly referred to as *visual servoing*, has been a research field of continued and increasing interest for more than three decades. The use of visual data within a feedback loop to position a robot has several benefits. The configurations of the robot effector and the environment are directly related via the visual perception providing accurate free-space motion control with even coarse knowledge of manipulator parameters. Only knowledge of local environmental geometry is needed allowing for application in weakly or unstructured environments.

Yet, only few survey or taxonomy articles exist, e.g. [21, 61, 69]. A comprehensive bibliography can also be found in [17]. Taxonomies are mainly formulated from a systems or application design perspective. A widely accepted classification of approaches is based on the following criteria: Domain of the task function, configuration of the vision system (type / placement), control of the components of the camera velocity screw, architecture hierarchy. A good approach for classification can be found in [69]. However, most classifications simply capture kinematic configurations. From a control point of view, main differences consist in the kinematic and dynamic effects and performance issues, e.g. sensitivity, condition, stability domain, robustness, and control performance. In this sense the distinction of [20, 22] between visual kinematic and visual dynamic control applies.

In the following sections a brief introduction in selected basic concepts of visual servoing and a survey of state-of-the-art approaches relevant to this thesis are given.

Image-, Position-based, and Partitioned Approaches

Image-based visual servo control is the earliest and most basic kinematic architecture. The basic principle is a resolved rate controller computing a manipulated value in Cartesian space from the control error defined in image space based on a differential relationship between Cartesian and image space - the visual or image Jacobian, also referred to as interaction matrix in combination with another transformation. The visual Jacobian first introduced in [133] is mostly based on the pinhole camera model computing the projections of Cartesian point motions onto the image plane for a moving camera.

Advantages of this method are robustness against calibration errors and no need of a 3D environment model. The major drawbacks of this method are potentially large, complex, and unefficient movements in Cartesian or joint space, singularities, the weak condition of the controller, and potential unstability in case feature points are occluded or leaving the image plane. Another shortcoming is the need of depth estimation. However, it can be noted that in many visual servoing schemes depth acts simply as a gain. An estimate of the depth at the desired position provides acceptable performance and decoupling near the desired pose. Works based on this concept mainly prove stability based on kinematic considerations completely neglecting manipulator dynamics.

Due to the relatively low framerates of the visual sensor most applications introduce an additional joint level controller. This controller is formulated, e.g. in [66] taking manipulator dynamics into account and proving stability of this non-linear system using Lyapunov's direct method [65, 66].

Position-based visual servoing relies on relative pose estimation between camera and reference object evaluating visual features based on a geometric model of the object or environment. The control error is, thus, defined in Cartesian space. As reconstruction of 3D-structures is based on camera parameters position-based methods can be susceptible

to calibration errors. Most approaches are based on the epipolar geometry and the homography estimation, which is susceptible to noise. An advantage is the formulation of the task in Cartesian space as common in robotics.

In order to overcome potential drawbacks of image-based visual servoing several partitioned approaches are known [23, 32, 89], which control particular Cartesian degrees of freedom using different methods, e.g. based on homography estimation. The method of Corke and Hutchinson [23], e.g., is concerned with avoiding large translations along the optical axis of the camera at particular tasks. Therefore, the translational and rotational Cartesian components along and around the optical axis, respectively, are decoupled and controlled using different methods. Image features are kept within the field of view using a repellant potential function. 2-1/2D visual servoing [89] avoids the need of a 3D model and is robust to calibration errors. However, this method is more susceptible to noise than classical image-based visual servoing for it is based on homography estimation used to control particular degrees of freedom. A complementary approach is the method of Deguchi [32], which has similar advantages and disadvantages.

Geometric, Invariance, and Subspace Methods

Some recent approaches to visual servoing exploit particular aspects of differential geometry and invariance regions, e.g. [24, 26, 41, 53, 87, 116], and subspaces, e.g. [33, 100, 109]. In [41] methods are proposed, which exploit properties of the Lie algebra of affine transformations. Observed affine and projective deformations to target planar contours of an object are directly related to Cartesian robot motion. One of the key advantages is robustness to a large range of perturbations of the Jacobian due to the fact that vectors through the origin of the space of deformations are geodesics in the manifold of the Lie group. Another advantage is the avoidance of suboptimal trajectories in Cartesian space. In [26] a diffeomorphism from a visible set, i.e. a subset of rigid body transformations relative to the camera that keep all features visible, to an image space is defined. Using the resulting Jacobian control is done in image space. The impact is a global method keeping features visible and avoiding self-occlusions at the cost of a specifically designed visual target.

Methods for visual control in invariant spaces are, e.g. proposed in [53, 87, 116]. In [53, 87] a projective space invariant to camera intrinsic parameters is used. The control error is defined in the invariant space. Some advantages are that a picture taken with a different camera can be used to derive the desired pose and invariance to the knowledge of the 3D model of the object. However, in order to estimate the Jacobian intrinsics are needed. Thus, robust methods are essential to ensure stability. The use of scale-invariant feature transform (SIFT) is, e.g., proposed by [116].

In [33, 100] a subspace of the visual workspace, a space capturing all possible appearance variations within a given task, is defined, which is the eigenspace. The visual workspace is, thus, represented as a continuous appearance manifold within the eigenspace. Motion parameters are obtained from the projection of a sample image to eigenspace and then to the manifold. In [109] the visual servoing task is formulated in projective space taking visibility and mobility constraints into consideration. Thus, trajectories are defined, which are visually and globally feasible.

Many of these methods might be considered less susceptive to noise and quantization effects. However, evidence and comparative investigations are pending. Independency of sensor scaling factors to a large extend is particularly beneficial in multi-resolution vision-

based control. These approaches, thus, may provide valuable tools for multi-focal vision in future research.

Multi-Camera Visual Servoing

In order to improve the accuracy and robustness and facilitate depth estimation, two or more vision sensors are used. This field of visual servoing approaches can be divided into two areas: approaches taking multiple view constraints, e.g. epipolar geometry or data fusion methods, into account and approaches where the individual sensors control either different Cartesian degrees of freedom or share the same degrees of freedom in a sequential manner.

Most common visual servoing methods considering multiple view constraints are based on stereo vision. The vision sensor is either fixed within the environment (eye-to-hand) [1, 4, 53, 58, 101] or mounted on the end-effector of the manipulator (eye-in-hand) [6, 25, 92, 105].

Several works on two-camera visual servoing are concerned with eye-in-hand/eye-to-hand cooperation strategies [43, 49, 85, 97]. In [49] the two cameras control rotational, respectively, translational Cartesian degrees of freedom of the robot manipulator. In [43] an end-effector camera measures object poses while a workspace camera acquires the end-effector pose. In [97] an eye-in-hand visual servoing setup serves as an eye-to-hand camera for a second robot manipulator. These approaches use the information of the different sensors for different objectives. Data fusion of both sensors towards a common objective using an extended Kalman-filter is proposed by [85].

Visual servo systems using more than two cameras are rare mainly due to time consuming matching across the different views, e.g. [117]. Further works are [52] using three stationary cameras estimating the pose of a workpiece and afterwards controlling the robot manipulator picking it up. In [113] a target is tracked in six views and a point-to-point positioning task is accomplished. Trinocular vision is used by [70] for grasping. In [135] methods for dealing with redundant sensors are presented as well as the effect of the image processing and visual servoing on robustness. In [90] several cameras with fixed relative poses are mounted on an end-effector and a resolved rate controller with a visual Jacobian composed of the individual sensor Jacobians is defined. Experiments are conducted using two cameras. Tuning of the individual control gains is proposed according to sensor accuracy or reliability, but no results are given.

Multi-Resolution Visual Servoing

Considering vision sensor data with different accuracies leads to a general sensor data fusion problem if all sensors contribute to a common control goal. However, only few vision-based control approaches make use of data fusion methods in order to integrate information of different vision sensors, e.g. [85]. Works on multi-resolution data fusion in visual servoing are not known.

If only one sensor at a time is used, e.g. using a zooming camera, the performance characteristics change over time. Works on zooming cameras are, e.g. [18, 57, 60, 88].

Other approaches consider invariant spaces, e.g. [53, 87, 116] (cf. Section 2.3), thereby, overcoming the problem of tracking objects observed with different sensor scaling factors.

In [90] the multi-resolution problem is addressed by tunig the gains corresponding to the individual cameras with respect to their accuracy (cf. Section 2.3). However, no results on this matter are presented.

Performance Issues

Comparability of results is a major problem in visual servoing research. In order to evaluate the quality of a positioning or tracking task performed by a vision-based control strategy performance measures are necessary. A common measure is the remaining control error which is considered in most works. However, the control error of image-based strategies is defined in image-space. Only very few works consider the Cartesian position or tracking error of these methods. Although, the control error is dependent on system parameters, e.g. gains, in most works these parameters are not mentioned. Another problematic aspect is negligence of manipulator dynamics in simulations and design. Thus, evaluation of dynamic properties is not possible. Sensor noise significantly impairs control performance. The stochastics of the visual servoing task are, thus, an important measure to assess control quality. Yet, there are only few works taking noise into account mainly modeling sensor or feature tracker noises, e.g. [51, 81, 103]. Recent approaches make a move to consider the propagation of sensor noise through controller and plant, however, not accounting for manipulator dynamics [81]. Quantification of the propagation of sensor noise is nontrivial as some parameters are not known or hard to be determined precisely, e.g. time delay.

The properties sensitivity and condition number of the controller are of particular interest in order to evaluate the performance of vision-based control. Sensitivity, also referred to as resolvability, is a measure for the ability of the controller to resolve motion in Cartesian or sensor space, respectively. This is an important aspect as it determines the impact of sensor noise on Cartesian motion. Resolvability investigations are conducted in [102] considering single camera, stereo, and orthogonal camera setups with fixed relative poses. Worst case sensitivity, i.e. the minimum singular value of the controller, is considered in [55] in order to investigate various combinations of feature points. The relationship between sensitivity and condition number is shown in [47] and both measures are considered in order to select optimal image features to be tracked. Perceptibility, the product of all singular values of the controller, as a non-directional global performance measure is considered mainly in sensor placement and trajectory planning, e.g. [34, 118].

Switching Approaches

Recent works in visual servo control propose switching approaches. Thereby, the controller, i.e. the vision-based strategy, switches between several different strategies in order to overcome some drawbacks of the individual methods [34, 50].

In [50] the visual controller switches from position-based visual servoing to image-based visual servoing whenever the visibility problem of position-based visual servoing is imminent. If the camera retreat problem occurs the controller switches back to the position-based method. This approach is shown to be locally asymptotically stable. In [34] the controller switches between the same strategies, thereby, avoiding singular configurations and local minima of the local visual controller.

A potential switching scheme that enlarges the stable region of image-based visual servoing is proposed in [56]. Relay images which interpolate initial and reference image features are generated by using affine transformations. Artificial potentials defined by the relay images are patched around the reference point of the original potential to enlarge the stable region.

2.4 Active Vision for Mobile Robots

The control of the camera view directions of vision guided robots is motivated by limitations of computational resources. Only a selected part of the environment is focused in order to reduce visual data to be processed. The camera is controlled in order to acquire as much information as possible with respect to certain constraints. The selection of the view direction of known approaches is influenced by the current situation and task (top-down) and/or potentially unforeseen stimuli (bottom-up). A variety of selection mechanisms exists from basic finite state machines [8] over multi-agent systems [45] and Markov Decision Processes (MDP) [123] to biologically inspired approaches, e.g. computational models of visual cortex processes [62].

Optimal gaze control goes hand in hand with goal-oriented scene understanding, i.e. understanding and interpreting of complex visual environments in a manner that depends on the robot's higher intentions and goals. Two main directions are entropy-, respectively, information-based approaches and top-down modulation of early sensory processing. Entropy-based approaches, i.e. methods based on information maximization, are mainly used in simultaneous localization and mapping. The basic principle is a task- or situation-dependent formulation of the information content of the scene. Based on this information model the gain of information for possible view directions is predicted and the direction providing a maximum information increase is chosen. Examples in this field are the works of the Active Vision Lab, University of Oxford [29, 30, 67], Imperial College London [130], the works around the humanoid JOHNNIE at the Institute of Automatic Control Engineering (LSR), Technische Universität München [45, 46], and the control of the MarVEye platform, e.g. [36, 37], discussed in Section 2.2.

There are a number of works on camera coordination, particularly, accounting for the combination of cameras with different characteristics in field of view and accuracy. A selection has been reviewed in the Section 2.2. However, only few have found their way towards an application in mobile robotics as task formulations requiring locomotion are not considered.

2.5 Summary

To date, only few works exist exploiting the particular nature of multi-focal vision. Most works are application oriented. In consequence generalizability of the approaches in order to work for a larger class of systems and settings is mostly weak. Most advanced approaches can be noted in the autonomous vehicles field. Many system embodiments exist, yet, only few of those have been used for investigating methodical multi-focal concepts. In the mobile and humanoid robot fields where most foveated systems have been developed almost no works have been done which go beyond basic oculomotor functions. Higher-level functions and connections to locomotion tasks are not covered. Comparative evaluations quantifying the benefits of multi-focal vision are not known.

For the first time, in this thesis hierachical investigations and comparative studies of multi-focal vision are conducted. Unique concepts of multi-focal vision-based control and active vision for mobile robots are presented which are largely independent of particular system embodiments and scenarios. The contributions of the presented work advance the state-of-the-art in machine vision, vision-based control, and vision-guided robotics.

3 Geometric Aspects of Multi-Focal Vision

Vision devices in measurement and robotics are commonly used in order to estimate geometrical or dynamical properties of the environment or the robot. Stereo-camera systems and systems comprising three and more vision devices have been introduced in order to reduce the estimation error, avoid singular configurations, overcome ambiguities, and extend the workspace, therefore, fusing the data of the individual devices. Multi-camera data fusion is commonly referred to as, e.g., multiple view or n-camera problem and a variety of approaches to its resolution have been proposed. Known approaches are formulated in a general way such that individual characteristics of the vision devices can be considered. Yet, methodical approaches examining the impact of multi-focal system configurations, i.e. how the combination of vision devices differing in their measurement performances influences overall system performance, are not known.

The measurement quality of a vision system depends on a number of factors: the intrinsic parameters of the device which determine the optical projection and optronical conversion, the geometrical configurations of the observing vision devices with respect to the observed structure, and the configuration of the observed structure itself. Works covering these aspects considering vision systems comprising only one device type at a time are known.

The scientific questions being answered in this chapter are how measurement quality changes if the individual intrinsic parameters and geometrical configurations of the vision devices vary independently and how this can be exploited in order to deliberately improve measurement quality. The contribution of the presented work is an improvement of the performance of visual perception using multi-focal multi-camera vision. Key challenges are the investigation and design of multi-focal vision systems in terms of condition and sensitivities. It is shown how multi-focal vision influences performance of perception, methods are proposed which environment points to observe with which sensor depending on the individual sensor characteristics and tools to assess and predict the expected performance change if the configuration of the multi-focal system is altered.

The remainder of this chapter is organized as follows: Assumed multi-focal perception models are defined in Section 3.1. In Section 3.2 selected performance properties are discussed and the capabilities of conventional single and multi-camera configurations are investigated. The performance of the multi-focal approach is assessed in Section 3.3. Measures and design tools to assess the performance of the proposed systems based on configuration and parameter changes are given in Section 3.4.

3.1 Assumptions and Perception Models

A general multi-focal vision system is shown in Figure 3.1 consisting of at least two cameras with individual characteristics. These characteristics are considered fully defined by their extrinsic and intrinsic parameters. A variety of formalisms exist to describe such a system. A well known description of the two-camera problem is given by the fundamental matrix relating the projections of a point in Cartesian space onto the image planes of both cameras

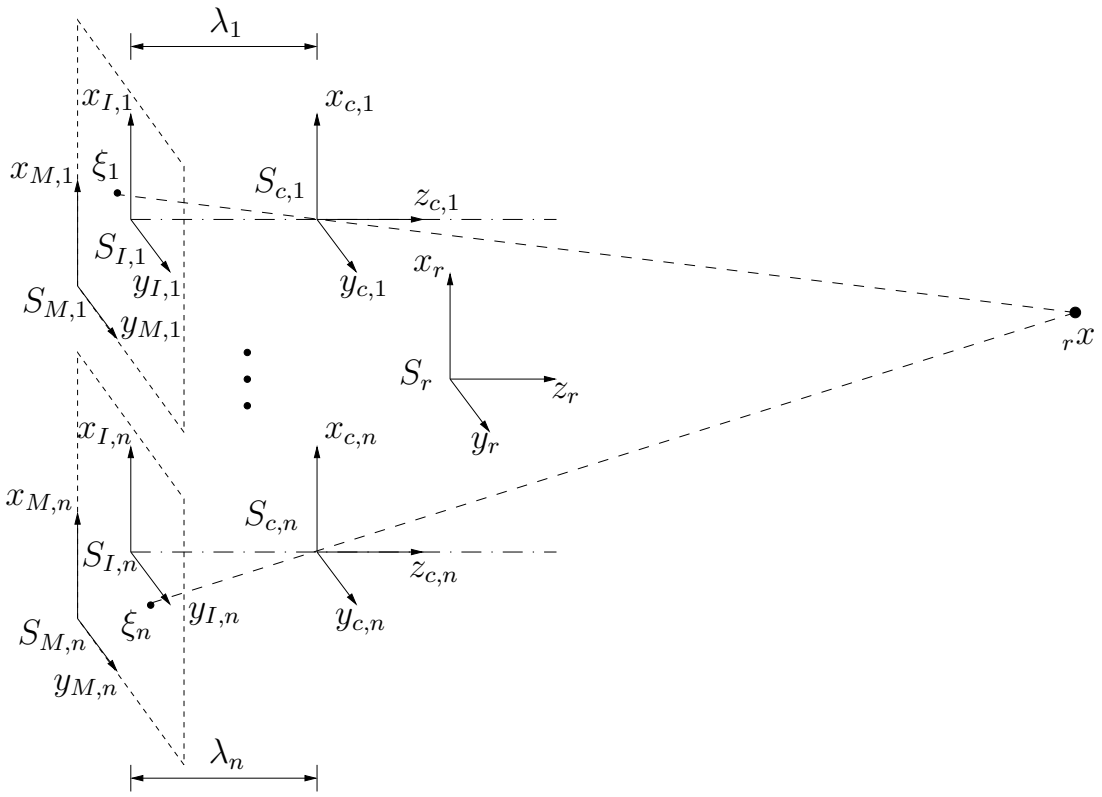

Figure 3.1: Multi-focal vision system with n cameras and individual focal-lengths λ_i, camera-head reference frame S_r, camera reference frames $S_{c,i}$, image frames $S_{I,i}$, and memory frames $S_{M,i}$; point ${}_rx = [X\,Y\,Z]^T$ in Cartesian space is projected to $\xi_i = [\xi_{u,i}\,\xi_{v,i}]^T$ in image plane i based on the pinhole camera model.

to each other [54]. The fundamental matrix is uniquely determined by a pair of camera matrices containing the camera parameters. A system of more than two cameras can be described by the multi-focal tensor, which is the extension of the fundamental matrix to the multi-camera case [54]. Yet, the tensor method does not extend to more than four views. However, a multiple view reconstruction problem may be decomposed into several three- or four-camera problems. Well known methods to solve the general multiple view problem are, e.g., projective factorization and bundle adjustment [54].

In this thesis, a perception model based on the method of Tsai [125] is considered. The vision sensor is represented by a pinhole camera model. A point in three-dimensional Cartesian space is projected onto the image planes of n cameras. The reconstruction of the Cartesian point from these projections leads to a static optimization problem, which is solved by using an iteratively reweighted least squares technique. This model is an extension to the two-camera model in [45].

3.1.1 Multi-Focal Perception Model

Perception Model. Considered is a 3D feature point ${}_rx$ in Cartesian space with respect to reference frame S_r. Following the method of Tsai [125] the corresponding projection

$\xi_i = \begin{bmatrix}\xi_{u,i} & \xi_{v,i}\end{bmatrix}^T$ into the image space of camera i is given by

$$s \begin{bmatrix}\xi_i^T & 1\end{bmatrix}^T = P_{ir} x, \tag{3.1}$$

where s is an arbitrary factor and the perspective projection matrix P_i of camera i is given by

$$P_i = \begin{bmatrix} -\lambda_i r_{11} & -\lambda_i r_{12} & -\lambda_i r_{13} & -\lambda_i t_x \\ -\lambda_i r_{21} & -\lambda_i r_{22} & -\lambda_i r_{23} & -\lambda_i t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} = \begin{bmatrix} p_{i,1}^T & p_{i,14} \\ p_{i,2}^T & p_{i,24} \\ p_{i,3}^T & p_{i,34} \end{bmatrix}, \tag{3.2}$$

with focal-length λ_i and r_{kl} and t_m the elements of the rotation matrix ${}^rR_{c,i}$ and translation vector $t_{c,i}$ of the homogenous transformation ${}^rT_{ci}$ between camera frame $S_{c,i}$ of camera i and S_r, respectively. For better readability the indices c, i have been left out.

Lens distortions can be represented by a first order model

$$\xi_{u,i} = \xi_{u,i,d}\left(1 + \kappa_i\left(\xi_{u,i,d}^2 + \xi_{v,i,d}^2\right)\right), \quad \xi_{v,i} = \xi_{v,i,d}\left(1 + \kappa_i\left(\xi_{u,i,d}^2 + \xi_{v,i,d}^2\right)\right),$$

where $(.)_d$ denotes the corresponding distorted position in image space and κ is a scalar distortion coefficient. Lens distortions are not considered in the remainder of this work as they do not restrict the general investigations on multi-focal vision presented and can easily be introduced in the framework as shown above.

Reconstruction of 3D Points. In order to reconstruct ${}_rx$ from its projections, the ξ_i are assumed to be known, the stereo correspondence problem is assumed solved, and quantization effects are neglected. The objective is to find the corresponding Cartesian coordinates ${}_rx$ fulfilling (3.1) for all cameras. Thus, a system of $2n$ linear equations is obtained, which can be written and solved in various ways [45]. Here the system is solved by writing

$$F_r x^* = b,$$

with ${}_rx^*$ the homogeneous coordinates of ${}_rx$ and

$$F = \begin{bmatrix} \xi_{u,1} p_{1,3}^T & -p_{1,1}^T \\ \xi_{v,1} p_{1,3}^T & -p_{1,2}^T \\ \vdots & \vdots \\ \xi_{u,n} p_{n,3}^T & -p_{n,1}^T \\ \xi_{v,n} p_{n,3}^T & -p_{n,2}^T \end{bmatrix}, b = \begin{bmatrix} p_{1,14} - \xi_{u,1} p_{1,34} \\ p_{1,14} - \xi_{v,1} p_{1,34} \\ \vdots \\ p_{n,14} - \xi_{u,n} p_{n,34} \\ p_{n,14} - \xi_{v,n} p_{n,34} \end{bmatrix}. \tag{3.3}$$

Matrix F and vector b are uncertain due to errors in the measurements of $\xi_{u,i}$ and $\xi_{v,i}$. For a first estimation of ${}_r\hat{x}^*$ the perturbations are neglected. The solution can then be found as the least squares solution of $\hat{F}$ given by

$${}_r\hat{x}^* = \left(\hat{F}^T \hat{F}\right)^{-1} \hat{F}^T \hat{b},$$

with estimated values $\hat{F}$ and $\hat{b}$. With this first solution of ${}_r\hat{x}^*$ an error vector e can be defined as

$$e = F_c \hat{x}^* - b \simeq \Delta F_c \hat{x}^* - \Delta b = g\left(\xi_1, \xi_2, ..., \xi_n\right),$$

with perturbations ΔF and Δb of F and b, respectively. The errors in image space $\Delta\xi_{u,i}$, $\Delta\xi_{v,i}$ are assumed uncorrelated with known variances $\sigma^2_{u,i}$, $\sigma^2_{v,i}$. With the Jacobian of g the error e can be approximated

$$e \simeq J_{uv} \begin{bmatrix} \Delta\xi_{u,1} & \Delta\xi_{v,1} & \dots & \Delta\xi_{u,n} & \Delta\xi_{v,n} \end{bmatrix}^T,$$

with $J_{uv} = diag\,(J_{uv,1}, ..., J_{uv,n})$.

A weighted least squares solution for the estimation of ${}_r\hat{x}^*$ is then given by

$${}_r\hat{x}^* = \left(\hat{F}^T E^{-1} \hat{F}\right)^{-1} \hat{F}^T E^{-1} \hat{b}, \tag{3.4}$$

with

$$E = J_{uv} E_{uv} J_{uv}^T, \tag{3.5}$$

and

$$E_{uv} = \text{diag}\left(\sigma^2_{u,1}, \sigma^2_{v,1}, \dots, \sigma^2_{u,n}, \sigma^2_{v,n}\right),$$

the error covariance matrix in image space.

An initial solution can be found setting $E = 1$. The iteratively reweighted least squares method is terminated, when the difference of two consecutive estimations of ${}_r\hat{x}^*$ is below an arbitrary threshold. This method has successfully been applied to a two-camera system [86].

Non-linear Representation. In the remainder of this thesis this vision sensor is represented by a non-linear model $h_\xi(x, P)$ and projections ξ_i are considered projections of particular features ${}_rx_i$ in Cartesian space. Each feature in Cartesian space is transformed to sensor space writing

$$\xi = h_\xi({}_rx, P, \lambda) = \begin{bmatrix} h_{\xi,1}({}_rx_1, P_1) \\ h_{\xi,2}({}_rx_2, P_2) \\ \vdots \\ h_{\xi,n}({}_rx_n, P_n) \end{bmatrix} \in \mathcal{H}_n, \tag{3.6}$$

with sensor space feature vector $\xi = [\xi_1^T \dots \xi_n^T]^T$, camera projection matrices P_i with focal-lengths λ_i, where the individual transformations $h_{\xi,i}({}_rx_i, P_i)$ may belong to one or individual vision devices. A particular parametrization of the sensor model is an element of the manifold of all possible sensor configurations of dimension n denoted by $\mathcal{H}_n$. The purpose of this work is to investigate the impact of simultaneous utilization of several vision sensors which differ in accuracy and field of view. Therefore, investigations are based on simple observed geometrical structures. Point features $\xi_i = [\xi_{u,i}\ \xi_{v,i}]^T$ and ${}_rx_i = [X_i\ Y_i\ Z_i]^T$ are considered. This representation will be referred to in Chapter 5 forming the perception model of a mobile robot. This model will then be integrated in the measurement equation of a Kalman-filter in order to estimate the robot pose from a acquired map of landmarks.

3.1.2 Perception of Motion

The underlying principle for the perception of camera motion considered in this thesis is the differential relation between Cartesian and sensor space motion - the visual Jacobian, which is a linearized sensor model of h_ξ. This Jacobian can be formulated for arbitrary features reaching from simple points to complex geometrical structures as well as geometrical transformations or deformations.

For an arbitrary multi-camera configuration the differential relationship can be formulated as

$$\dot{\xi} = J_v(\lambda, \xi, Z) R_0 \dot{x}, \tag{3.7}$$

with

$$\dot{\xi} = \begin{bmatrix} \dot{\xi}_1^T & \dot{\xi}_2^T & \dots & \dot{\xi}_n^T \end{bmatrix}^T, \; J_v = \mathrm{diag}\Big(J_{v,1}(\lambda_1, \xi_1, Z_1), J_{v,2}(\lambda_2, \xi_2, Z_2), \dots, J_{v,n}(\lambda_n, \xi_n, Z_n) \Big),$$

$$R = \mathrm{diag}\Big(R_1, R_1, R_2, R_2, \dots, R_n, R_n \Big), \quad R_i \in SO(3),$$

$${}_0x = [{}_0x_1^T \; {}_0x_2^T \; \dots \; {}_0x_n^T]^T, \quad {}_0x_i = [X_i \; Y_i \; Z_i \; \alpha_i \; \beta_i \; \gamma_i]^T,$$

with feature vectors in sensor space ξ_i, corresponding visual Jacobian $J_{v,i}$, which is element of manifold $\mathcal{J}_n$, whereas several Jacobians can belong to one or individual sensors, and R_i the transformation between camera i and a reference frame within the vision system. Vector ${}_0\dot{x}$ is the Cartesian velocity screw of the vision system. Considering only the focal-length λ as intrinsic parameter, the Jacobian $J_{v,i}$ can, e.g., be written

$$J_{v,i} = \begin{bmatrix} \frac{\lambda_i}{Z} & 0 & -\frac{\xi_{u,i}}{Z_i} & -\frac{\xi_{u,i}\xi_{v,i}}{\lambda_i} & \frac{\lambda_i^2+\xi_{u,i}^2}{\lambda_i} & -\xi_{v,i} \\ 0 & \frac{\lambda_i}{Z_i} & -\frac{\xi_{v,i}}{Z_i} & -\frac{\lambda_i^2+\xi_{v,i}^2}{\lambda_i} & \frac{\xi_{u,i}\xi_{v,i}}{\lambda_i} & \xi_{u,i} \end{bmatrix}. \tag{3.8}$$

Taking the epipolar constraint into consideration individual feature points may be perceived by two or more sensors simultaneously, which are, thus, forming *effective stereo-pairs*. Consequently, N cameras observing the same point form $(N^2 - N)/2$ effective stereo-pairs. The optical axes of the individual stereo-pairs are considered parallel. Assuming a planar configuration of cameras, i.e. disparities only in ξ_u-direction, and the stereo correspondence problem solved, then the individual Jacobians for the cameras of one effective stereo-pair can be written in case the cameras are parallel

$$J_{v,i} = J_{v,ij}^e = \begin{bmatrix} \frac{\lambda_i d_n}{b_{ij}} & 0 & \frac{\xi_{u,i} d_n}{b_{ij}} & \frac{\xi_{u,i}\xi_{v,i}}{\lambda_i} & \frac{\lambda_i^2+\xi_{u,i}^2}{\lambda_i} & \xi_{v,i} \\ 0 & \frac{\lambda_i d_n}{b_{ij}} & \frac{\xi_{v,i} d_n}{b_{ij}} & \frac{\lambda_i^2+\xi_{v,i}^2}{\lambda_i} & \frac{\xi_{u,i}\xi_{v,i}}{\lambda_i} & \xi_{u,i} \end{bmatrix},$$

with a normed multi-focal stereo disparity

$$d_n = \frac{\xi_{u,i}}{\lambda_i} - \frac{\xi_{u,j}}{\lambda_j},$$

with $(.)_i$ and $(.)_j$ referring to the left and the right camera of an effective stereo-pair, respectively, and the stereo baseline b_{ij}.

In the remainder of this chapter investigations are based on the visual Jacobian in order to obtain comparable results with earlier works on mono-focal vision, e.g. cf. [102]. In Chapter 4 the visual Jacobian will be used for the formulation of a vision-based controller in order to control a robot manipulator.

3.2 Performance of Single- and Multi-Camera Perception

In order to assess the performance of a multi-focal system, it is essential to know the performance of mono-focal systems as a reference. Known literature considers various measures for performance assessment. In [102] sensitivity is used to investigate various camera and feature point configurations, however, translational and rotational components are completely decoupled. The condition number is considered, e.g., in [47]. Perceptibility [118] as a global non-directional measure is mainly focused by sensor placement and planning problems.

Due to their different character, in this thesis the performance of visual perception is evaluated considering all these measures, i.e. sensitivity, perceptibility, and condition of the respective sensor Jacobian, simultaneously. In the following these measures are explained in brief. The performance of selected mono-focal single- and multi-camera configurations is evaluated based on these measures.

3.2.1 State-of-the-Art Tools for Performance Assessment

Sensitivity, also referred to as resolvability [102], is a directionally dependent measure for the dependency of different spaces by a transforming system A. Sensitivity is represented by the singular values and eigenvectors of matrix A^TA obtained by singular value decomposition (SVD).

The SVD of a matrix A is given by

$$A = U\Sigma V^T,$$

with the diagonal matrix Σ containing the singular values of A, matrix U containing the eigenvectors of AA^T, and V containing the eigenvectors of A^TA.

Here, the matrix A is the visual Jacobian J_v. The eigenvectors of $J_v^TJ_v$ giving a set of basis vectors v_i of the row space of J_v multiplied with their corresponding singular values σ_i, thus, represent a measure for the ability of J_v to perceive motions in Cartesian space. The minimum singular value can be considered a measure of worst case sensitivity.

Perceptibility, similar to manipulability, is defined as the volume of the ellipsoid

$$\sum_{i=1}^{m} \frac{1}{\sigma_i^2} U^T \dot{\xi}_i \leq 1,$$

in m-dimensional space [118]. The volume is given by

$$w_v = \sqrt{\det J_v^T J_v} = \sigma_1\sigma_2 \ldots \sigma_m, \quad 2n > m,$$

for the redundant case, with feature space dimension $2n$ and Cartesian space dimension m, and

$$w_v = \sqrt{\det J_v J_v^T} = \sigma_1\sigma_2 \ldots \sigma_2 n, \quad 2n < m,$$

for the under observed case, up to a factor, which is only dependent on m or $2n$, respectively, and, therefore, neglected. It is, thus, a non-directional global measure to evaluate the ability of J_v to perceive geometrical structures and motion.

The *condition number* of J_v is given by

$$c(J_v) = \|J_v\|\|J_v^{-1}\| = \frac{\sigma_{max}}{\sigma_{min}},$$

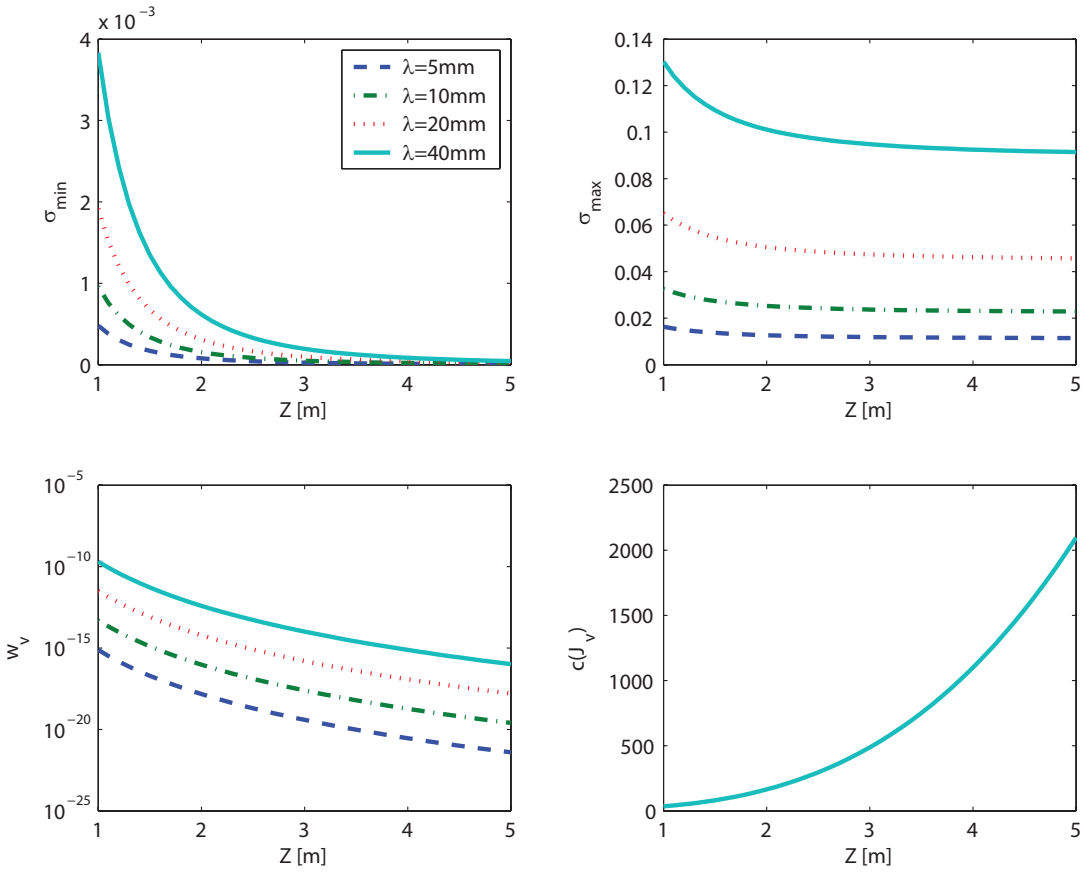

Figure 3.2: Performance of a single-camera vision system over distance Z to an observed object of five feature points forming a square in Cartesian space (edge lengths 0.5m, 0.05m displacement from optical axis in x-direction); minimum singular value σ_{min}, maximum singular value σ_{max}, perceptibility w_v, and condition number c of the visual Jacobian J_v.

where the norm may be $\|.\|_1$, $\|.\|_2$, or $\|.\|_\infty$, with maximum and minimum singular value σ_{max} and σ_{min}, respectively. Small values near $c(J_v) = 1$ imply that J_v is well-conditioned, i.e. equally sensitive in all directions. The condition number is also a non-directional measure. Contrary to perceptibility no information on sensitivity is provided.

3.2.2 Performance of Single- and Multi-Camera Vision Systems

In this section the performance of mono-focal vision systems is assessed in terms of sensitivity, perceptibility, and condition number of the visual Jacobian. Regarding sensitivity the worst and best case sensitivities are considered corresponding to minimum and maximum singular values. Varied parameters are the distance to an observed object, the focal-lengths, and the stereo baseline.

Considered are selected vision systems: a single-camera setup tracking a reference object (five feature points forming a square in Cartesian space with additional central point) and the same setup with additional stereo-camera pair tracking the central point of the object.

Figure 3.2 shows the propagation of the smallest and largest singular values, the perceptibility, and the condition number with increasing distance and varied focal-length. The dependency on distance Z is obvious as the singular values decrease, hence perceptibility

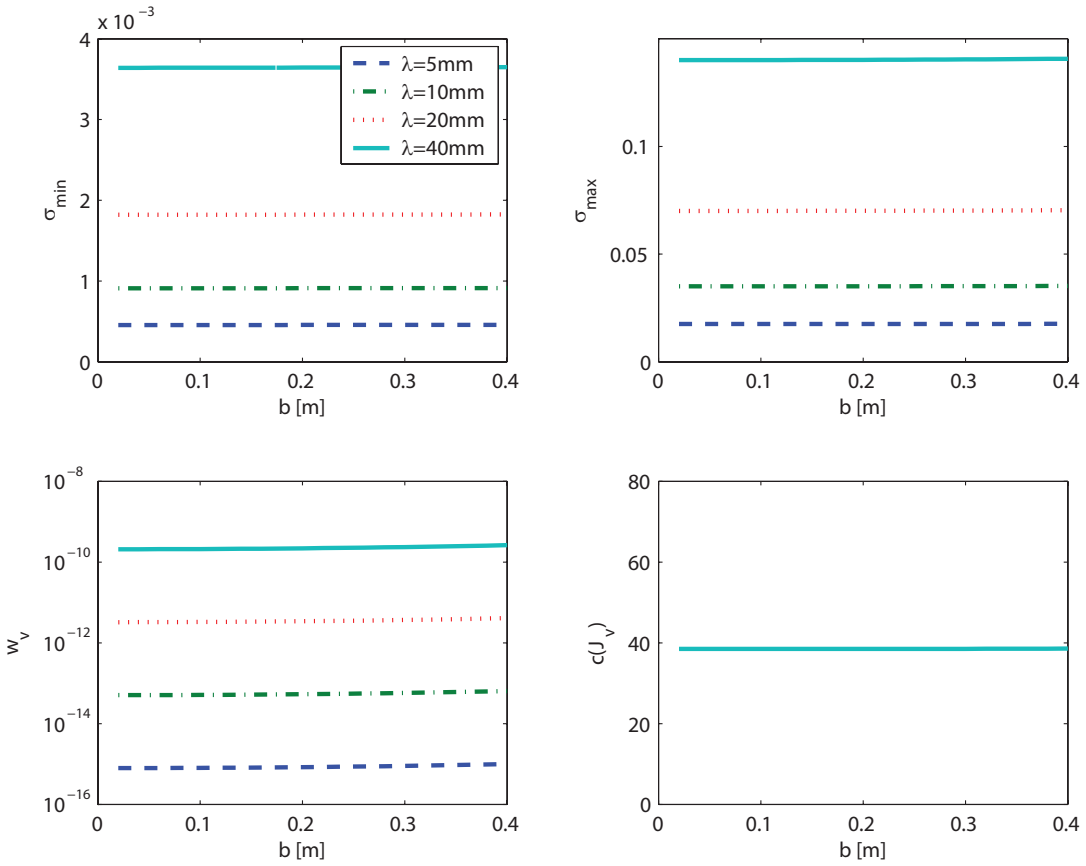

Figure 3.3: Performance of a stereo-camera vision system over stereo baseline b observing an object of five feature points forming a square in Cartesian space at distance $Z = 3$m (edge lengths 0.5m, 0.05m displacement from optical axis in x-direction); minimum singular value σ_{min}, maximum singular value σ_{max}, perceptibility w_v, and condition number c of the visual Jacobian J_v.

decreases, and the condition number rises due to the stronger decrease of the smallest singular values. The dependency on focal-length is linear, which can be seen by the propagation of perceptibility. However, due to limitations of the vision sensor chip size focal-length and field of view are approximately invers proportional. Thus, an increase of sensitivity is only possible with a decrease of the field of view. The dependency of sensitivity on distance Z is stronger than linear.

The dependency on Z for the stereo case is similar to the single-camera case up to a scaling factor corresponding to the baseline length. Thus, the distance Z is kept constant at $Z = 3$m and the baseline is varied. The results are displayed in Figure 3.3. Compared to the single-camera case the performance is improved and a very slight increase of performance with increasing baseline is notable. The singular values and perceptibility increase slightly, and the condition number falls in a non-linear manner.

Summarized, the performance in terms of sensitivity, perceptibility, and condition number decreases with distance, and increases with stereo baseline. An improvement of the performance can only be achieved by a larger focal-length increasing sensitivity at the cost of a reduced field-of view. The results will serve as a reference for the following investigations of multi-focal vision.

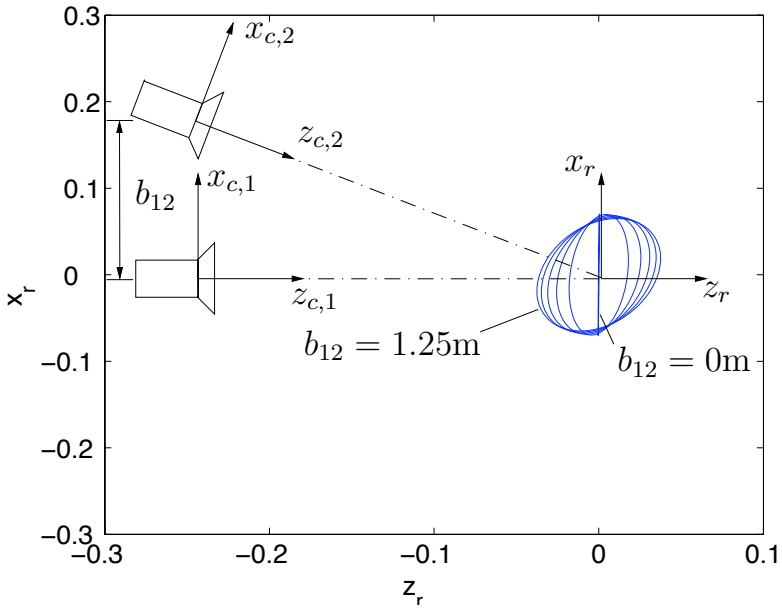

Figure 3.4: Sensitivity ellipsoids of a two-camera system with focal-lengths $\lambda_i = 5\text{mm}$ and varying stereo-base b_{12} observing an environment point at ${}_r x = [0\ 0.01\ 0]^T\text{m}$.

3.3 Multi-Focal Perception Performance

Visual perception is limited by sensor characteristics, which has been demonstrated in the previous section. Sensitivity and field of view are contradictory requirements, which cannot be overcome by the use of one or more vision sensors of the same type.

In this section the performance of a novel approach to visual perception based on multi-focal vision is assessed. The key idea is the simultaneous utilization of wide-angle vision sensors and additional vision sensors with comparably higher sensitivities and narrower aperture angles. The high sensitivity sensors observe a small region of the reference object with high accuracy. Thereby, a wide field of view is provided and the overall performance of perception in terms of sensitivity is significantly improved. The outcome of this approach is a vision system with an improved performance and the benefit of a wide field of view, which outperforms a setup with equal number of cameras and intermediate characteristics.

In the following several embodiments of the multi-focal perception models defined in Section 3.1 are discussed including the cases considered in Section 3.2.2. The performance is evaluated in terms of sensitivity, perceptibility, and condition number.

3.3.1 Sensitivity Ellipsoids

The singular values Σ of the visual Jacobian J_v define an ellipsoid. This ellipsoid is projected into Cartesian space by ${}_r\Sigma = V^T \Sigma V$ with V containing the eigenvectors of $J_v^T J_v$. The projected ellipsoid represents the ability of J_v to resolve motion in Cartesian space. A larger main axis implies better sensitivity in the Cartesian directions corresponding to the particular eigenvector.

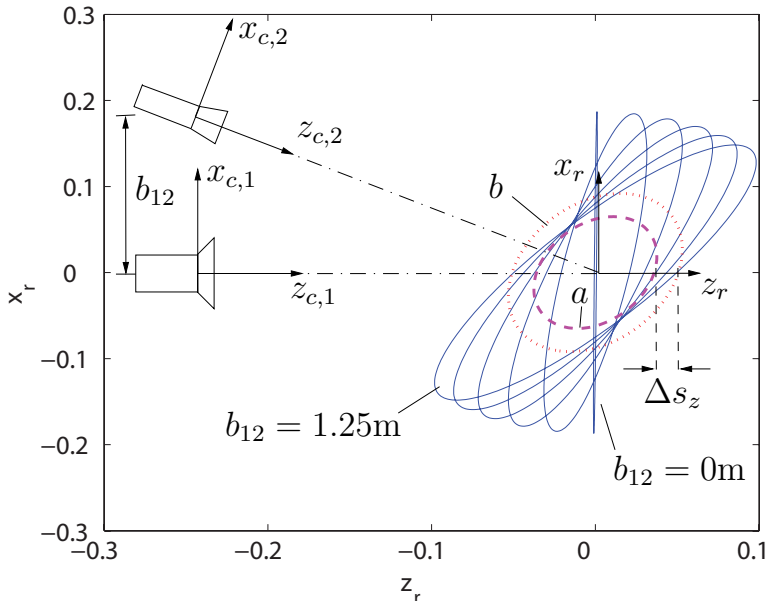

Figure 3.5: Sensitivity ellipsoids of a multi-focal two-camera system with focal-lengths $\lambda_1 = 5\text{mm}$, $\lambda_2 = 50\text{mm}$ and varying stereo-base b_{12} observing an environment point at ${}_r x = [0\ 0.01\ 0]^T\text{m}$.

The different translational sensitivity ellipsoids of a mono- and multi-focal vision system are shown in Figure 3.4 to Figure 3.6 for a two-camera vision system with equal and different focal-lengths, respectively. The optical axes are directed towards an observed feature point in Cartesian space. In Figure 3.4 and Figure 3.5 the stereo baseline is varied, while in Figure 3.6 the baseline is fixed and the focal-length of camera 2 is varied. The sensitivity ellipsoids of the individual cameras superpose resulting in a rotation of the ellipsoid of the two-camera system as the baseline or focal-length of camera 2 increase and the angle between the optical axes grows. The weak sensitivity in direction of the optical axis is obvious. As the angle between both optical axes approaches 90° the ellipsoids generally approach a sphere if focal-lengths and distances between cameras and feature point are equal.

Also the impact of an increase of the focal-length of camera 2 can be noted in Figure 3.5. The volume of the ellipsoid increases, i.e. perceptibility is improved. The dependency of the main axes lengths and rotation on focal-lengths is also shown in Figure 3.6: the larger the focal-length of camera 2 the stronger the rotation of the ellipsoid. A stronger increase of the larger main axis can be noted resulting in a weaker condition of the multi-focal setup. Due to the rotation of the larger main axis of the ellipsoid towards the optical axis of camera 1 particularly sensitivity along the optical axis of camera 1 increases significantly. Thus, the drawback of a weaker condition is turned into a particular advantage of a multi-focal system.

Concluding from these results, an improvement of sensitivity at the cost of weaker condition can be expected from a multi-focal vision system with the benefit of retaining a wide field of view. The weaker condition can be exploited to selectively increase the sensitivity in a particular direction by rotation of the sensitivity ellipsoid. The rotation is achieved

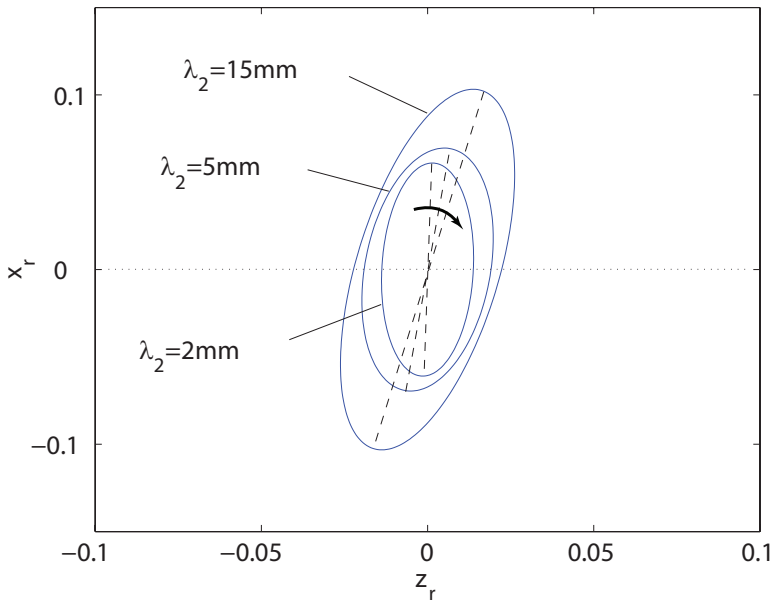

Figure 3.6: Sensitivity ellipsoids of a two-camera system; camera 1 with fixed focal-length $\lambda_1 = 5$mm and camera 2 with varied focal-length λ_2; stereo-base $b_{12} = 0.3$m; observed environment point ${}_r x = [0\ 0.01\ 0]^T$m.

by increasing the focal-length of a particular sensor and may be further increased by a rotation of the higher-sensitivity camera.

Example: Consider again the mono-focal and multi-focal setups in Figure 3.4 and Figure 3.5, respectively. As stated above due to the increased volume of the ellipsoid and the rotation of the larger main axis towards the optical axis of camera 1 the sensitivity in $z_{c,1}$-direction increases significantly.

Consider the ellipsoids resulting from a baseline length $b_{12} = 1.25$m for the multi-focal case in Figure 3.5. For comparison also the dashed ellipsoid a for the mono-focal setup is shown. The difference Δs_z of the sensitivities of the mono- and multi-focal setup in $z_{c,1}$-direction amounts approximately 0.05-0.036=0.014. Thus, an improvement of 39% is achieved by using multi-focal vision; additionally a wide field of view is provided.

Consider now the dotted ellipse b corresponding to a two-camera setup with focal-lengths of 10mm. This vision system has approximately the same sensitivity in $z_{c,1}$-direction as the multi-focal setup with focal-lengths of 5mm and 50mm, however, without the benefit of the wider field of view.

Summarizing, increasing the focal-length of one camera of a multi-camera system improves overall sensitivity retaining a wide field of view. A multi-focal setup reaches a sensitivity which lies between the sensitivities achieved with sensors of its smallest and its largest focal-length.

3.3.2 Sensitivity, Perceptibility, and Condition

The multi-focal approach consists in the combination of low-sensitivity vision sensors with large aperture angles and high-sensitivity sensors with small aperture angles to improve the

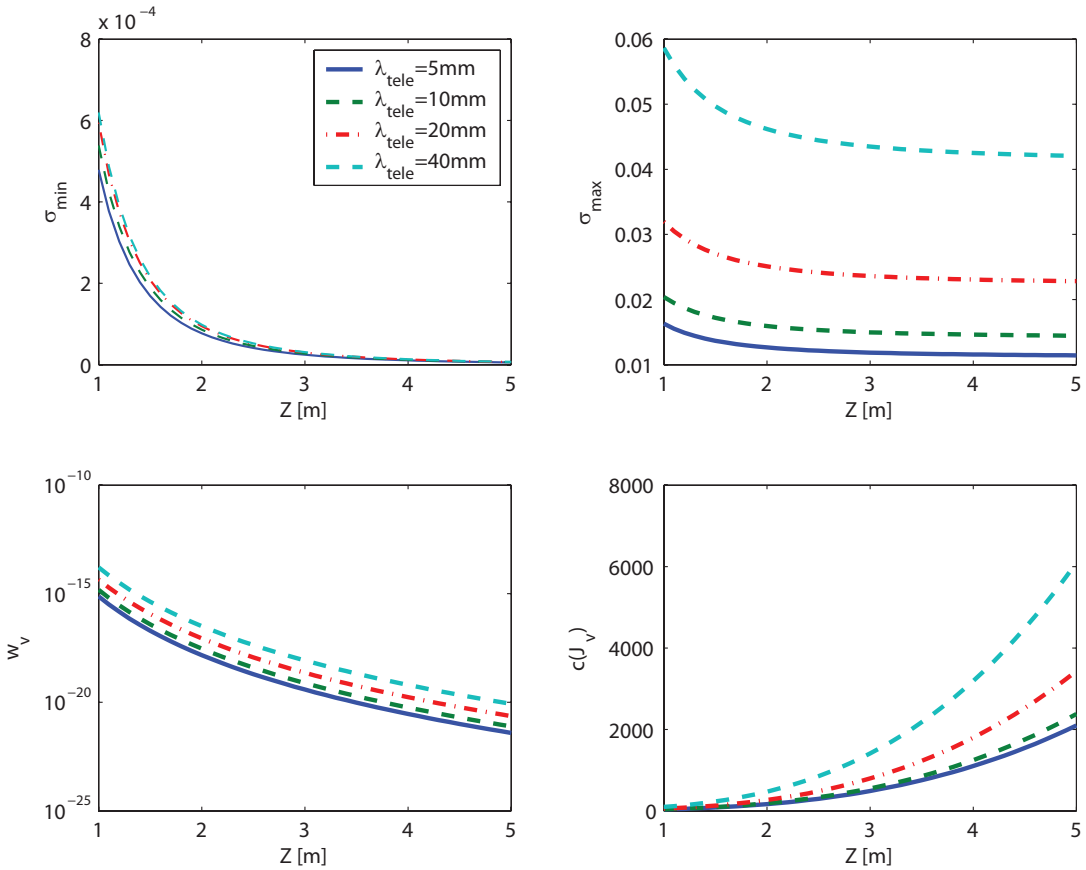

Figure 3.7: Performance of a multi-focal vision system with varied focal-length of the telephoto camera over distance Z to an observed object of five feature points forming a square in Cartesian space (edge lengths 0.5m, 0.05m displacement from optical axis in x-direction); minimum singular value σ_{min}, maximum singular value σ_{max}, perceptibility w_v, and condition number c of the visual Jacobian J_v; focal-length of the wide-angle camera $\lambda_{wide} = 5\text{mm}$.

perception of Cartesian motion while retaining the benefit of a wide field of view. In this section the performance of selected multi-focal vision system configurations is evaluated quantitatively. The aspects addressed are the progression of sensitivity, perceptibility, and condition with changing focal-length and stereo baseline.

Considered are the cases studied in Section 3.2.2, whereas one feature point (central feature point of the square object) in Cartesian space is observed with a high-sensitivity vision sensor or stereo-pair while the other points are observed with a low-sensitivity sensor. The focal-lengths of the sensors tracking the central feature point are varied, while the focal-length of the wide-angle sensor tracking the remaining features is kept constant. The results are shown in Figure 3.7 and Figure 3.8.

An increase of performance in terms of sensitivity and perceptibility can be noted as the singular values are increased as the focal-length of the telephoto sensor grows compared to the mono-focal case. The increase of the largest singular value is stronger leading to larger condition numbers as expected in Section 3.3.1. The change of performance is exemplarily shown for a distance of one meter in Table 3.1 and Table 3.2 for the mono and stereo case, respectively. As qualitatively shown in Section 3.3.1 it is noted that the performance of the

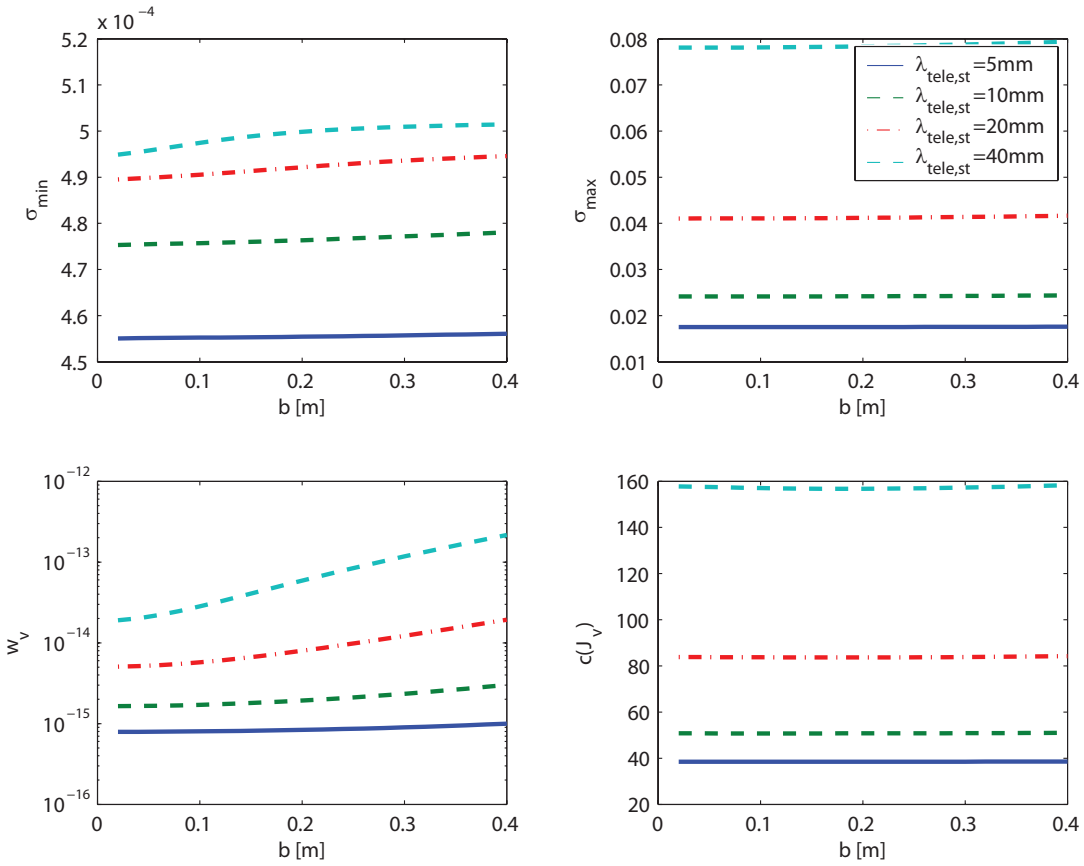

Figure 3.8: Performance of a multi-focal stereo-vision system with varied focal-length of the telephoto stereo-camera over stereo baseline b observing an object of five feature points forming a square in Cartesian space at distance $Z = 3$m (edge lengths 0.5m, 0.05m displacement from optical axis in x-direction); minimum singular value σ_{min}, maximum singular value σ_{max}, perceptibility w_v, and condition number c of the visual Jacobian J_v; focal-length of the wide-angle camera $\lambda_{wide} = 5$mm.

assessed multi-focal system with focal-lengths of 5mm and 40mm shows approximately the same performance as a system with equal numbers of cameras, but focal-lengths of 10mm. Thus, observing only one feature point with a high-sensitivity sensor improves sensitivity significantly and retains a wide field of view.

Summarized, the proposed multi-focal approach to perception significantly improves sensitivity and perceptibility at the cost of an increased condition number. Observing selected feature points with high sensitivity an overall sensitivity is achieved, which is significantly higher than that of the lowest sensitivity sensor. The multi-focal system achieves the same performance as a particular multi-camera system of sensors of a single type with intermediate sensitivity. The wide field of view is a significant advantage of the multi-focal system, which cannot be achieved with a mono-focal setup of the same sensitivity.

Table 3.1: Motion perception performance with additional telephoto camera with focal-length $\lambda_{tele} = 40$mm at $Z = 1$m.

	σ_{min}	σ_{max}	w_v	$c(J_v)$
mono-focal	$4.8 \cdot 10^{-4}$	$1.6 \cdot 10^{-2}$	$7.5 \cdot 10^{-16}$	34
multi-focal	$6.2 \cdot 10^{-4}$	$5.9 \cdot 10^{-2}$	$1.6 \cdot 10^{-14}$	95

Table 3.2: Motion perception performance with additional telephoto stereo-camera with focal-length $\lambda_{tele} = 40$mm and baseline $b = 0.2$m at $Z = 1$m.

	σ_{min}	σ_{max}	w_v	$c(J_v)$
mono-focal	$4.6 \cdot 10^{-4}$	$1.8 \cdot 10^{-2}$	$7.9 \cdot 10^{-16}$	38
multi-focal	$5.0 \cdot 10^{-4}$	$7.8 \cdot 10^{-2}$	$1.9 \cdot 10^{-14}$	157

3.4 Tools for Design, Configuration, and Performance Assessment

In the previous section a multi-focal approach to perception has been proposed and assessed. The sensitivity of perception has been improved and the field of view increased compared to the mono-focal case. This section is concerned with optimal vision system configuration and the investigation of measures to predict the expected change of the performance due to changes in the vision sensor configuration. The results can be used as design tools for construction of multi-focal vision systems and tools for dynamic configuration of vision systems in use to adapt to the current situation and performance requirements.

3.4.1 Optimal Focus of Attention

Assuming a multi-focal vision system with a wide-angle device and one or more high-sensitivity devices with strongly limited field of view the question remains, which is the best feature to be observed with high sensitivity in order to achieve an optimal performance. A possible approach is to evaluate the directional sensitivities dependent on sensor pose configurations which will be explained in Section 3.4.2. Therefore, 3D knowledge on object geometry is necessary. This section proposes an approach which only requires knowledge on the camera parameters and measurements in sensor space under the restriction that the origins of all camera frames are nearly coincident, i.e. their distances are sufficiently small. Considering the performance in terms of directional sensitivities, perceptibility, and condition number several conditions for determination of an optimal camera orientation are reasonable: maximum directional sensitivity, maximum perceptibility, minimum condition number, or a trade-off between these.

Consider again the singular value decomposition of the visual Jacobian J_v

$$J_v = U \Sigma V^T,$$

with U containing the eigenvectors of JJ^T, i.e. a set of basis vectors for the column space of J. The products of these basis vector elements and their corresponding singular values $u_{ij}\sigma_j$ are a measure for the perception of Cartesian space structures or motion in sensor space. Thus, better perception in particular Cartesian directions observing a feature point in sensor space implies high values of the row vector elements of U corresponding to this feature point and these Cartesian directions.

This fact is utilized for the determination of an optimal focus of attention for the high-sensitivity sensors of a multi-focal vision system. The key idea is to evaluate the sensor space eigenvector elements with respect to the conditions defined above. This is done for the visual Jacobian of the wide-angle sensor observing all considered features. The high-resolution sensors are then directed towards those feature points for which the conditions are met best.

Maximum Directional Sensitivity. The determination of the best focus of attention to achieve maximum sensitivity in a particular Cartesian direction can be achieved as follows. The optimal feature ξ^* to be observed is the feature ξ_k in sensor space for which

$$\xi^* = \left\{ \xi_k \middle| \Psi_s = \max_{k=1,\ldots,n} |u_{2k-1,j}| + |u_{2k,j}| \right\}, \quad \xi = \begin{bmatrix} \xi_u \\ \xi_v \end{bmatrix}, \quad u_{ij} \in U,$$

holds, with feature point ξ_k in sensor space, the number of features n, and u_{ij} the element of U corresponding to Cartesian direction j, to the feature point k and the direction in sensor space ξ_u or ξ_v, respectively. In other words, that feature with the largest eigenvectors of $J_v J_v^T$ in sensor space corresponding to the desired Cartesian direction is focused.

Maximum Perceptibility. Perceptibility can be defined as the product of all singular values. In order to achieve maximum perceptibility, that particular feature ξ_k has to be selected, for which the absolute values of all corresponding elements of U are largest, i.e. in terms of the Euclidean norm. The optimal feature point ξ^* can, e.g., be determined

$$\xi^* = \left\{ \xi_k \middle| \Psi_p = \max_{k=1,\ldots,n} \|u_{2k-1}\| + \|u_{2k}\| \right\}, \quad \xi = \begin{bmatrix} \xi_u \\ \xi_v \end{bmatrix}, \tag{3.9}$$

with the row-vectors u_{2k-1} and u_{2k} of U corresponding to feature point ξ_k.

Minimum Condition Number. The condition number can be defined as the ratio of maximum to minimum singular value. Thus, the ratios of the elements u_{ij} corresponding to these singular values and the respective feature point ξ_k are to be maximized yielding

$$\xi^* = \left\{ \xi_k \middle| \Psi_c = \min_{k=1,\ldots,n} |\frac{u_{2k-1,1}}{u_{2k-1,s}}| + |\frac{u_{2k,1}}{u_{2k,s}}| \right\}, \quad \xi = \begin{bmatrix} \xi_u \\ \xi_v \end{bmatrix}, \quad u_{ij} \in U,$$

with s the number of singular values

Optimizing Perceptibility and Condition. A possible approach to achieve a good compromise between perceptibility and condition number is to evaluate the difference between the normalized definitions for perceptibility and condition number according to

$$\xi^* = \{ \xi_k \big| \Psi_{p,c} = \max_{k=1,\ldots,\frac{m}{2}} \frac{\|u_{2k-1}\| + \|u_{2k}\| - \Psi_{p,min}}{\Psi_p^*} - \frac{|\frac{u_{2k-1,1}}{u_{2k-1,n}}| + |\frac{u_{2k,1}}{u_{2k,n}}| - \Psi_c^*}{\Psi_{c,max}} \},$$

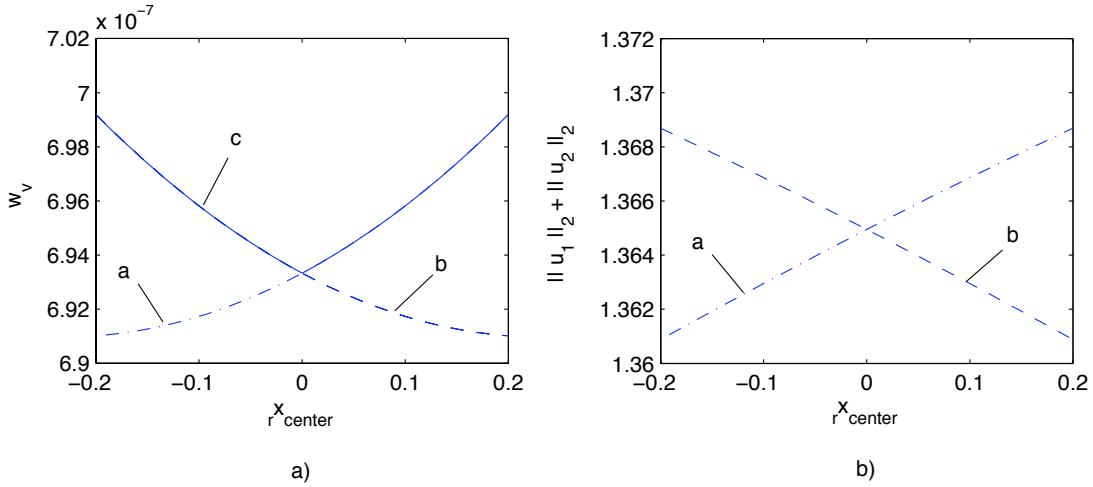

Figure 3.9: a) Perceptibility of a multi-focal vision system with two coaxial cameras observing a triangular object of three feature points at ${}_rx_1 = [-0.3\ 0.3\ 5]^T$m, ${}_rx_2 = [0.3\ 0.3\ 5]^T$m, and ${}_rx_3 = [0\ 0.5\ 5]^T$m and focal-lengths of 5mm (wide-angle camera) and 25mm (telephoto camera); the telephoto camera is observing either ${}_rx_1$ (curve a) or ${}_rx_2$ (curve b); perceptibility curve c is obtained by optimal focus of attention of the telephoto camera; b) 2-norms of the row-vectors of input matrix U containing the eigenvectors of $J_v J_v^T$.

$$\begin{aligned} \Psi_{p,min} &= \min_{k=1,\ldots,\frac{m}{2}} \|u_{2k-1}\|_2 + \|u_{2k}\|_2, \\ \Psi_{c,max} &= \max_{k=1,\ldots,\frac{m}{2}} |\frac{u_{2k-1,1}}{u_{2k-1,n}}| + |\frac{u_{2k,1}}{u_{2k,n}}|, \end{aligned}$$

with the eigenvectors evaluated regarding minimum perceptibility $\Psi_{p,min}$ and maximum condition number $\Psi_{c,max}$. The normalization is defined by the difference between actual and minimum value devided by the maximum value. The optimum is defined as the maximum difference between normalized perceptibility and normalized condition number.

Example Consider a uniform object motion along the x_r-axis orthogonal to the optical axes of an ideal two-camera system with coaxial optical axes, common image plane, and $S_{c,i}$ and S_r are coincident. The observed object consists of three feature points at ${}_rx_1 = [-0.3\,0.3\,5]^T$m, ${}_rx_2 = [0.3\,0.3\,5]^T$m, and ${}_rx_3 = [0\,0.5\,5]^T$m with respect to its center of gravity, forming a triangle in Cartesian space. Its center of gravity moves from ${}_rx_{center} = -0.2$m to ${}_rx_{center} = +0.2$m. The optical axes are, thus, orthogonal to the object surface. The vision system consists of a wide-angle camera with a focal-length of $\lambda_1 = 5$mm and a telephoto camera with $\lambda_2 = 25$mm. Due to its narrow field of view the telephoto camera can only observe a selected point, whereas the other two points are perceived by the wide-angle camera. The objective is now to dynamically decide, which is the best feature point to be observed by the telephoto camera in order to achieve maximum perceptibility.

Applying (3.9) the eigenvectors of the visual Jacobian of the wide-angle camera observing all three points are evaluated. That feature point of the object is then focused with the telephoto camera, which fulfills the criterion for maximum perceptibility, i.e. that corresponding to the largest 2-norm of the particular eigenvectors of $J_v J_v^T$ in image space.

The results are shown in Figure 3.9. Figure 3.9a shows the progression of the perceptibility with the telephoto camera focussing either ${}_rx_1$ (curve a) or ${}_rx_2$ (curve b). Feature point ${}_rx_3$ is

not considered as its contribution is neglectable. The corresponding 2-norms of the eigenvectors are shown in Figure 3.9b. Applying the proposed focus of attention mechanism that feature point is selected dynamically which provides the maximum perceptibility. The resulting progression of the perceptibility is shown in Figure 3.9a curve c, which is clearly the maximum possible perceptibility.

3.4.2 Design Considerations

In the preceding sections the impact of multi-focal vision on perception has been investigated and mechanisms for optimal foci of attention of the individual sensors of such a vision system have been proposed. In this section quantitative measures are given to assess the resulting change of performance due to configuration changes.

Change of Focal-Length. It is a well known fact that sensitivity of a mono-focal vision system depends linearly on focal-length. This is clearly seen by the change of the singular values and perceptibility shown in Figure 3.2. Thus, a straight-forward measure to assess the expected change of performance of a mono-focal vision system depending on the change of focal-length is given by

$$\frac{w_{v,i}}{w_{v,j}} = \frac{\sigma_{i,1}\sigma_{i,2}\ldots\sigma_{i,m}}{\sigma_{j,1}\sigma_{j,2}\ldots\sigma_{j,m}} = \frac{\prod_{k=1}^{m}\lambda_i}{\prod_{l=1}^{m}\lambda_j} = \left(\frac{\lambda_i}{\lambda_j}\right)^m,$$

with perceptibilities w_v, focal-lengths λ, singular values σ, $(.)_i$ and $(.)_j$ denoting the sensors, and number of singular values m, respectively, the number of Cartesian degrees of freedom if the number of feature points is at least $\frac{m}{2}$.

Assuming a current sensor configuration $J_{v,i}$ in a current situation with perceptibility $w_{v,i}$ and focal-length λ_i utilizing this measure an appropriate focal-length can be selected to achieve a particular perceptibility.

General Multi-Focal System Configuration. In the general case the focal-lengths of the individual sensors can change independently. The kinematic configuration, i.e. relative poses of the sensors, may also change arbitrarily. For general multi-focal systems expressing performance measures like sensitivities in terms of focal-lengths and homogeneous transformations is more complex.

In order to assess performance changes of a general multi-focal vision system due to changes of focal-lengths and geometrical configuration the evaluation of the directional sensitivity of its Jacobian is proposed. The sensitivities in a desired direction after the change are predicted and compared with the current values. Based on directional sensitivities also perceptibility and condition can be computed. In the following the procedure is described in detail:

(i) **Definition of Desired Directional Sensitivity.** First, the desired value of the sensitivity in a particular direction ${}_rs$ and the corresponding directional vector ${}_rr^d \in \mathbb{R}^m$ are defined with number of Cartesian degrees of freedom m.

(ii) **Computation of Singular value Ellipsoid of the Jacobian.** The Jacobian J_v of the parametrized multi-focal vision system is computed. The parameters are, e.g. focal-lengths λ_i of the individual sensors, their relative poses expressed by homogeneous transformations ${}^rT_{ci}$, and the observed feature points ${}_rx_i$. In order to obtain the sensitivity ellipsoid, a singular value decomposition of J_v is computed according to

$$J_v = U\Sigma V^T, \tag{3.10}$$

with singular matrix Σ and U, V^T the projectors into image, respectively, Cartesian space. The ellipsoid hull is given by ${}_rr^T\Sigma_r r = 1$, ${}_rr \in \mathbb{R}^m$.

(iii) **Projection into Cartesian Space.** The sensitivity ellipsoid is projected into Cartesian space by

$${}_r\Sigma = V^T\Sigma V, \quad {}_r\Sigma = f(\lambda, {}^rT_{ci}, {}_rx), \quad i = 1, ..., n, \tag{3.11}$$

with ${}_r(.)$ denoting the reference frame in which the relative sensor poses and observed feature points are defined.

(iv) **Computation of Sensitivities.** The sensitivity in a particular desired direction ${}_rr^d \in \mathbb{R}^m$ is computed satisfying the projection

$${}_rr^T{}_r\Sigma_r r = 1, \quad {}_rr \in \mathbb{R}^3, \tag{3.12}$$

with point ${}_rr$ on ellipsoid hull and satisfying for the scalar product of desired direction ${}_rr^d$ and point ${}_rr$

$${}_rr^d{}_rr = \|{}_rr^d\|\|{}_rr\|. \tag{3.13}$$

The solution is a vector ${}_rr^*$. The Euclidean norm $\|{}_rr^*\|$ of which represents the sensitivity in the desired direction.

(v) **Evaluation of Sensitivities.** The sensitivities of a current and desired configuration of a multi-focal system can be compared, e.g. computing the ratio

$$c_r = \frac{\|{}_rr_2^*\|}{\|{}_rr_1^*\|}, \tag{3.14}$$

with $\|{}_rr_1^*\|$ and $\|{}_rr_2^*\|$ the sensitivities in a desired direction ${}_rr^d$ of the Jacobians $J_{v,1}$ and $J_{v,2}$, respectively, i.e. of the desired and current configuration.

If desired sensitivities are known the reverse problem has to be solved. In this case, the corresponding focal-lengths and geometrical configurations of the multi-focal system are the wanted parameters and the equations above have to be solved for λ and ${}^rT_{ci}$. However, the existence of ambiguities is obvious, e.g. the same sensitivity can be achieved by variation of the focal-length or rotation of the sensor. Therefore, it is opportune to reduce the dimension of the solution space to an acceptable extent, e.g. keeping sensor poses constant while solving for focal-lengths.

The proposed procedure can serve as a generic tool for the design and evaluation of multi-focal vision systems.

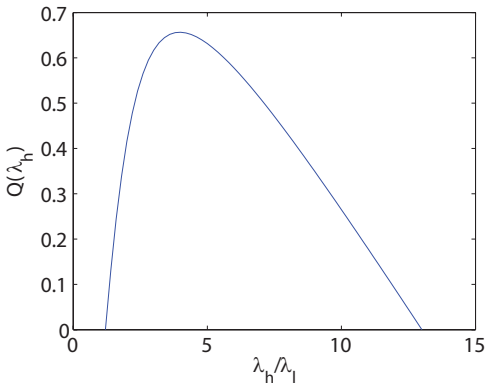

Figure 3.10: Objective function $Q(\lambda_h)$ trading sensitivity versus condition number over the ratio of focal-lengths λ_h/λ_l of a multi-focal two-camera system with coaxial optical axes observing a square object of five feature points in Cartesian space at a distance of $Z = 5$m.

Trade-off between Condition Number and Sensitivity. As discussed in Sections 3.3.1 and 3.3.2 sensitivity and condition are contradictory requirements in multi-focal vision. A selection of a larger focal-length of one of the vision sensors increases sensitivity and decreases the condition number. If both measures have to be taken into account an optimal focal-length λ_h^* of the high-sensitivity sensor can be determined

$$\lambda_h^* = \left\{ \lambda_h \in [\lambda_{h,min}, \lambda_{h,max}] \middle| Q(\lambda_h^*) = \max_{\lambda_h \in [\lambda_{h,min}, \lambda_{h,max}]} \left(\frac{{}_r s - {}_r s_0}{\Delta_r s} \right)^2 - \left(\frac{c - c_0}{\Delta c} \right)^2 \right\},$$

where $[\lambda_{h,min}, \lambda_{h,max}]$ defines the range of selectable focal-lengths, ${}_r s = f(\lambda_h)$ is the resulting sensitivity, $c = f(\lambda_h)$ is the resulting condition number, $\Delta(.)$ denotes the range of sensitivities, respectively, condition numbers corresponding to the range of focal-lengths, and $(.)_0$ denotes the magnitude corresponding to the minimum focal-length $\lambda_{h,min}$. Similarly, perceptibility and condition can be optimized.

Example: Considered is again the configuration in Section 3.3.2 of two cameras with parallel optical axes and coincident camera frames observing the square object. The focal-length λ_l of the wide-angle camera is set to 5mm and the focal-length λ_h of the telephoto camera is varied. The telephoto camera is observing the central point of the square object. The aspect addressed is the focal-length which optimizes sensitivity and condition number.

Evaluating the objective function $Q(\lambda_h)$ from (3.4.2) shown in Figure 3.10 for the minimum singular value and the condition number over the ratio of focal-lengths of both cameras a maximum at about a ratio of 4 can be noted. The optimal focal-length of the telephoto camera to be selected to provide a good trade-off between sensitivity and condition number is, thus, 20mm.

3.5 Discussion

Multi-camera vision is a powerful means for the estimation of environmental structures and motions. A system of multiple highly accurate sensors provides precise measurements, but requires extensive computational resources if their fields of view have to cover a large part of the environment. Additionally, the density of usable information might be low if

regions of interest are sparsely distributed in the visible space. Thus, accuracy is traded versus system and computational complexity and field of view.

In this chapter multi-focal vision systems, i.e. systems providing several sensors with different sensitivities and fields of view, are systematically investigated. The main focus is on the performance of perception in terms of sensitivity, perceptibility, and condition. An improvement of sensitivity by a combination of high-sensitivity sensors and low-sensitivity sensors is achieved. A multi-focal system provides a significantly wider field of view compared to a mono-focal system with equal sensitivity. Methods are proposed to determine the best environment point to be focused with the high-sensitivity sensors in order to optimize performance. A potential drawback of multi-focal vision is the weaker condition. Yet, this fact is exploited to improve sensitivity in a particular Cartesian direction by a change of the focal-length and optionally by changing the relative camera poses, thereby, rotating the sensitivity ellipsoid. Methods are given to assess the system performance allowing selective configuration changes of the vision system according to performance and situational requirements.

The contribution of this chapter facilitates the application of multi-focal vision under well-defined performance constraints providing high sensitivity, wide field of view, and higher information density of the visual data stream, and allowing the reduction of sensor and computational resources. Methods are proposed to optimize system performance and evaluated in extensive simulations advancing the state-of-the-art. Yet, the influence of more complex sensor models, image processing methods for feature extraction, and quantization effects on multi-focal perception is not addressed and will be subject of future research.

4 Multi-Focal Control of Robot Manipulators

In the preceding chapter the effects of multi-focal vision system configurations on measurement quality have been investigated. Methods have been proposed to increase the performance of perception by determining optimal system configurations at particular operating points. Camera and environment dynamics, i.e. controlled changes of operating points, are not taken into account. This chapter is concerned with the dynamical effects of multi-focal vision and multi-focal vision-based control (*visual servoing*).

Common visual servoing techniques suffer from several shortcomings. The visual controller degenerates with increasing distance to the observed reference object and decreasing focal-length of the vision device resulting in increased pose errors and pose error variances due to sensor noise and quantization or even rendering the whole system unstable. A certain control performance is only achievable by providing a sufficient focal-length, thereby, limiting the field of view. Yet, other concurrent conditions require a certain field of view, e.g. in order to assure visibility of a sufficient number and configuration of feature points of an observed object necessary in order to render the controller full rank, thereby, limiting the maximum focal-length. Thus, only a small operating range exists in which a desired control performance and stability can be assured.

Approaches towards visual servoing utilizing an adjustable focal-length have been proposed in order to overcome the limitation problems of the operating range. Yet, higher modeling and calibration complexity are introduced and the control dynamics are limited by focus adjustments. Other approaches consider features invariant to intrinsic parameters, subspace and geometrical methods which may be expected less susceptible to noise and, thus, to the degeneration of the controller. Yet, the operating range restrictions also apply.

The innovation of this chapter consists in multi-focal approaches to visual servoing. The primary goal is an improvement of visual servoing performance in terms of pose error variance and operating distance range. Novel concepts presented are a hybrid switching visual servoing strategy based on a dynamical sensor selection accounting for performance and field of view requirements and a multi-camera visual servoing strategy allocating high-sensitivity vision devices to selected features to be observed in addition to a wide-angle device. Key challenges are the formulation and performance assessment of multi-focal visual servoing strategies, the definition of switching conditions, and the stability analysis.

The remainder of this chapter is organized as follows: The basic assumptions and problem definition are given in Section 4.1. Preliminary investigations of the performance of conventional visual servoing with changing operating distances and focal-lengths are conducted in Section 4.2. The hybrid multi-focal visual servoing approach is introduced in Section 4.3, stability is proven, and performance is evaluated in comparison to the conventional approach. A multi-camera strategy is introduced and evaluated in Section 4.4.

4.1 Assumptions and Problem Definition

4.1.1 Problem Definition

Visual control of robot manipulators, commonly referred to as *visual servoing*, has been a research field of continued and increasing interest for the past three decades. The use of visual data within a feedback loop to position a robot has several benefits. The configurations of the robot effector and the environment are directly related via the visual perception providing accurate free-space motion control with even coarse knowledge of manipulator parameters. Only knowledge of local environmental geometry is needed allowing for application in weakly or unstructured environments.

Vision-based control problems covered by the known literature are commonly situated in small workspaces capable to be covered by a standard industrial robot manipulator. The main aspect strongly limiting the workspace is its distance dependent sensitivity resulting in a strong decrease of the control performance in terms of pose error, pose error variance, and stability with increasing distance to the observed reference structures. Larger focal-lengths can be chosen improving control performance, however, cannot solve the distance dependency problem. Moreover, large focal-lengths result in reduced fields of view. In consequence the number of visible features of the observed structure reduces potentially resulting in singularities of the visual controller.

Beside well known architectures, where the task is defined in different geometric spaces, i.e. image space or Cartesian space, e.g. cf. [61], respectively, several partitioned approaches exist, e.g. [23, 32, 89]. These address different shortcomings of the classical approaches as, e.g. the need of depth estimation and large translations at particular tasks. More recent research is concerned with invariance and geometric control to deal with different perspective transformations and transformation dependent features, e.g. [24, 26, 53, 87, 116]. The control performance can be assumed less effected by camera properties. However, all these approaches break down if conducted over a wide operating distance range. In order to overcome these shortcomings, works have been done on visual servoing based on variable intrinsic camera parameters, in particular, focal-length, e.g. [18, 57, 60, 88]. The focus can be controlled for an optimal focal-length, depth of focus, and field of view depending on the current situation and performance requirements. For this purpose camera models are proposed covering calibration of variable intrinsics, yet, adding modeling and preparation complexity. Some approaches consider the sensitivity of the visual Jacobian in order to control camera intrinsics and to plan optimal camera trajectories, e.g. [47, 55, 102, 118]. A methodical quantification of the impact on control performance in terms of pose error and variance is not known. Recent approaches consider switching controllers, e.g. to increase operating range and to overcome shortcomings as visibility of features and large movements, e.g. [34, 50, 56]. Only few systematic comparative evaluations of visual servoing performances are known, e.g. [51]. Performance in terms of pose error and pose error variance is scarcely considered and mainly evaluated in experimental validations. Manipulator dynamics are rarely taken into account.

A novel hybrid sensor switching visual servoing strategy will be presented in Section 4.3 in order to cope with the mentioned drawbacks of distance dependency and feature visibility resulting in improved control performance. Further improvements of control performance are achieved by a second approach presented in Section 4.4 dynamically allocating high-sensitivity vision devices to selected regions of the observed object.

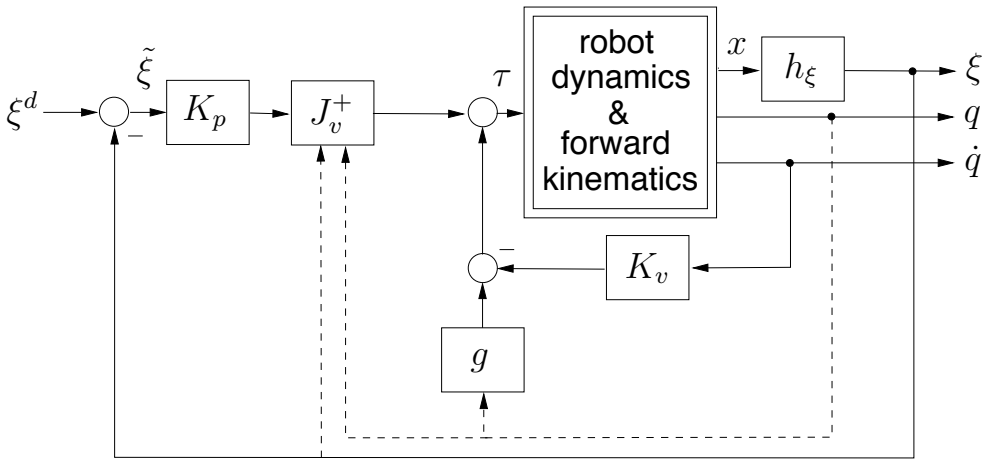

Figure 4.1: Conventional image-based visual servoing architecture.

4.1.2 Assumptions

In this chapter standard image-based visual servoing architectures considering manipulator dynamics form the basis of the investigations. A block diagram of such an architecture is shown in Figure 4.1. The manipulator dynamics are given by

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + g(q) = \tau, \tag{4.1}$$

with matrices M and C corresponding to manipulator inertia, centripetal and Coriolis torques $C\dot{q}$, gravitational torques g, and joint torques τ.

A cascade control approach is used based on a joint-level velocity controller and a visual controller computing commanded torques. The control law is

$$\tau = J_v(\xi, x(q), \dot{q})^+ K_p \tilde{\xi} - K_v \dot{q} + g(q), \tag{4.2}$$

with torques τ, the visual Jacobian J_v which has been described in Chapter 3 formulated with respect to joint coordinates q, positive-definite gain matrices K_p, K_v, feature error $\tilde{\xi} = \xi^d - \xi$ between the desired feature vector ξ^d and the current one $\xi = f(q)$ in image space, gravity compensation $g(q)$, joint angles q, and pose of the vision device $x(q)$. Only point features $\xi_i = [\xi_{u,i}\ \xi_{v,i}]^T$ of observed environmental structures are considered forming the feature vector $\xi = [\xi_1^T\ \xi_2^T\ \ldots\ \xi_m^T]^T$ not limiting the generality of the proposed multi-focal approaches.

The closed-loop system is obtained combining (4.1) and (4.2). In terms of the state-vector $[q^T\ \dot{q}^T]^T$ the system behavior can be written

$$\frac{d}{dt}\begin{bmatrix} q \\ \dot{q} \end{bmatrix} = \begin{bmatrix} \dot{q} \\ M(q)^{-1}(J_v(\xi, x(q), \dot{q})^+ K_p \tilde{\xi} - K_v \dot{q} - C(q,\dot{q})\dot{q} \end{bmatrix} = f_s(\xi, x, \dot{q}). \tag{4.3}$$

The multi-focal strategies proposed in Sections 4.3 and 4.4 are exemplarily instantiated and analyzed as extensions to the standard method, however, are not limited to a particular visual servoing approach.

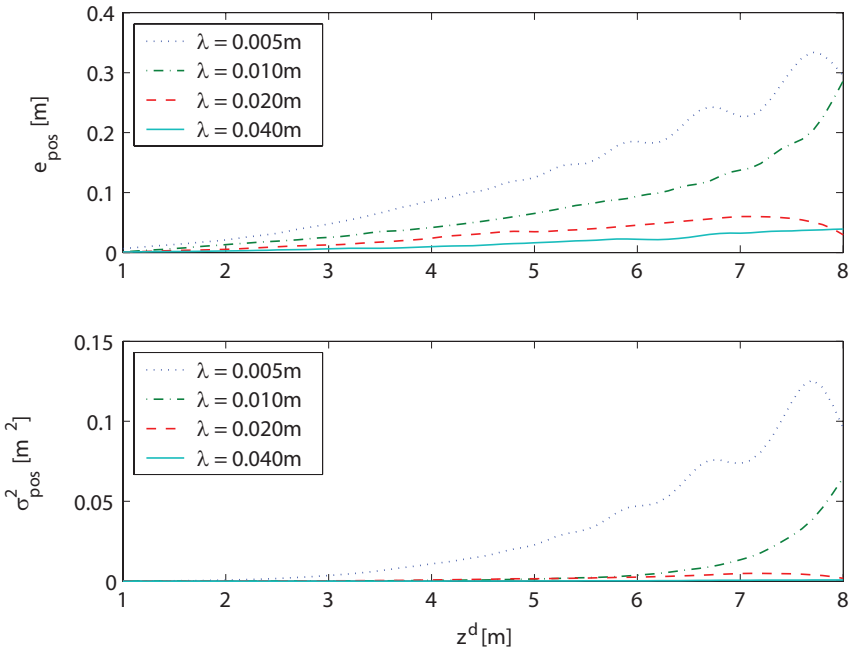

Figure 4.2: Remaining average translation error e_{pos} and error noise power σ^2_{pos} of visual servoing translation task along the optical axis versus desired distance to observed object (square with 0.5m edge lengths) at goal pose z^d with variable focal-length λ; inertia matrix $M = 0.05\text{diag}(1\text{kg}, 1\text{kg}, 1\text{kg}, 1\text{kgm}^2, 1\text{kgm}^2, 1\text{kgm}^2)$, damping $K_v + C = 0.2\text{diag}(1\text{kgs}^{-1}, 1\text{kgs}^{-1}, 1\text{kgs}^{-1}, 1\text{kgms}^{-1}, 1\text{kgms}^{-1}, 1\text{kgms}^{-1})$, feedback quantization 0.00001m, sensor noise power $\sigma^2_{meas} = 0.00001^2\text{m}^2$, control gain K_p tuned to converge system after approximately 2s.

4.2 Preliminaries on Conventional Visual Servoing Performance

In order to quantify the dependency of the pose and tracking error and error variances on distance and focal-length, preliminary investigations are conducted. The results will serve as a reference for the proposed visual servoing controller in Section 4.3.

In order to obtain comparable results, standard visual servoing tasks are performed: Firstly, a translation along the optical axis in order to reach a fixed desired pose and secondly a trajectory following task along the optical axis. These control aims are expressed in image coordinates ξ^d by transforming the Cartesian tasks into image space. For the first task the desired pose, i.e. the distance to the observed reference object, is varied. The desired trajectory of the second task is given by a sinusoidal translation away from and back to the observed object along the optical axis and a rotation about the optical axis

$$x^d(t) = \left[0 \quad 0 \quad \tfrac{7}{2}\sin\left(\tfrac{1}{5}t - \tfrac{\pi}{2}\right) - \tfrac{7}{2} \quad 0 \quad 0 \quad \tfrac{1}{5}t\right]^T.$$

For both tasks the focal-length is varied. The manipulator dynamics are modeled by a simple decoupled mass-damper-system. Manipulator geometry is neglected. Joint and Cartesian spaces are, thus, equivalent. Due to the stochastic nature of the process Monte-Carlo simulations with 50 trials are conducted. The control performance is evaluated in

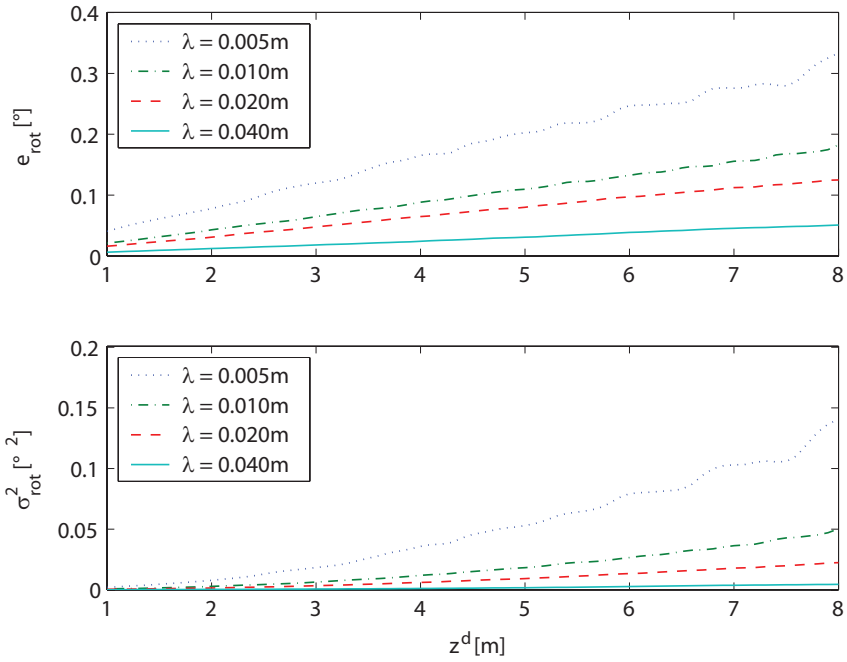

Figure 4.3: Remaining average rotation error e_{rot} and error noise power σ^2_{rot} of visual servoing translation task along the optical axis versus desired distance to observed object at goal pose z^d with variable focal-length λ; same parameter setting as in Figure 4.2.

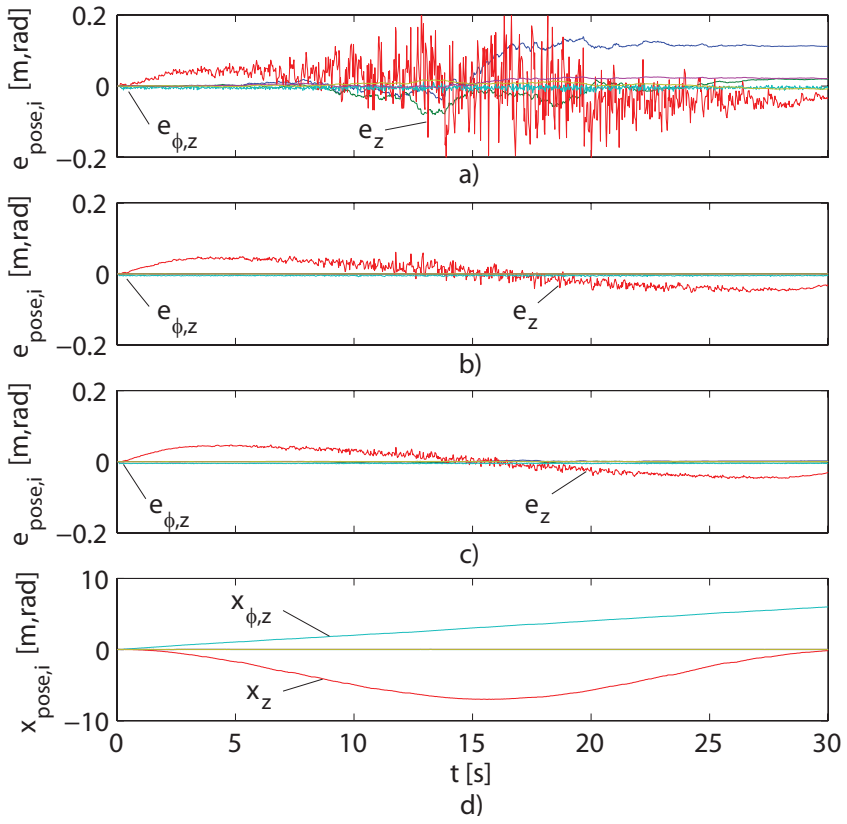

Figure 4.4: Tracking errors $e_{pose,i}$ and trajectory $x_{pose,i}$ of visual servoing trajectory following task over time t; focal-lengths a) $\lambda = 0.01$m, b) $\lambda = 0.02$m, c) $\lambda = 0.04$m; same parameter setting as in Figure 4.2.

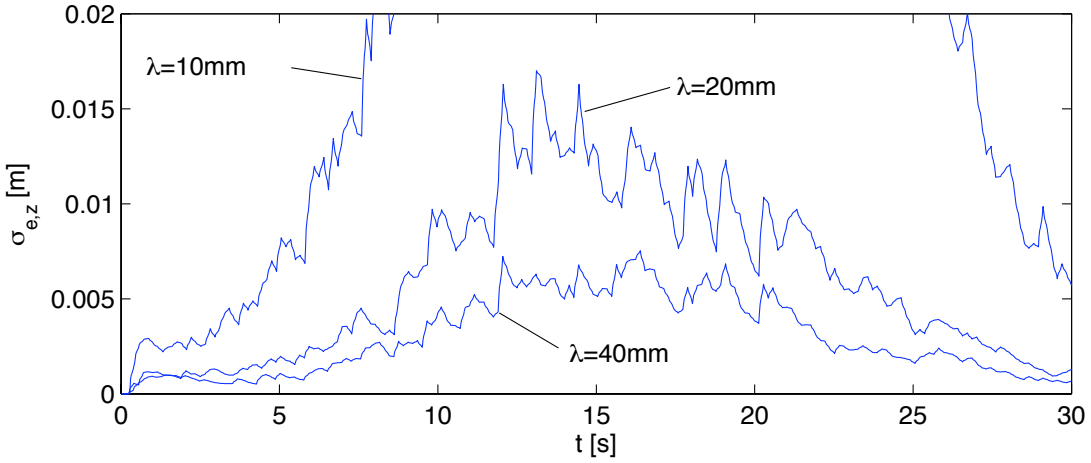

Figure 4.5: Corresponding short-time z-position error standard deviation estimates $\sigma_{e,z}$; same parameter setting as in Figure 4.2; time window $W = 3$.

terms of pose error, pose error variance and pixel error estimates utilizing the metrics defined in Appendix B.

4.2.1 Performance Evaluation

The simulation results of the two tasks are shown in Figure 4.2 to Figure 4.5. Figure 4.2 shows the translation errors and error standard deviations, Figure 4.3 the rotation error and error standard deviations. The tracking errors and error standard deviations of the trajectory following task are shown in Figure 4.4 and Figure 4.5, respectively.

As has been discussed in the preceding section the pose errors and pose error variances increase strongly with distance to the observed object. The increase is stronger than proportional. The dependency on focal-length is inverse proportional. For comparison the measured remaining pixel error and the pixel error variance are constant over the whole operating range amounting approximately 0.5 pixels and 0.01 pixels^2.

4.2.2 Discussion of the Results

In spite of the very low control errors in sensor space large errors in Cartesian space exist. The intrinsics and distance dependent sensitivity of the visual Jacobian and quantization effects result in varying pose error and pose error variance over the operating range caused by sensor noise. These effects remain a problem for wide range visual servoing rendering conventional visual servoing strategies unusable.

4.3 Hybrid Multi-Focal Visual Servoing

The preliminary results of the previous section show that conventional visual servoing techniques cannot cope with wide range applications due to the impact of quantization and noise on control performance with varying sensitivity of the visual controller. Constant errors and variances or at least upper bounds are desirable in order to make applications reasonable.

In this section a novel hybrid visual servoing approach based on multi-focal vision is proposed in order to overcome the investigated drawbacks of conventional techniques. The key idea is a dynamic camera switching strategy considering situational and performance parameters. In order to achieve a particular performance by dynamic switching, performance measures are necessary, which quantify the parameter dependent changes of the control performance. Therefore, measures based on internal properties of the visual controller are desirable in order to be independent of potentially erroneous and computationally expensive online performance assessments. In the following paragraphs the hybrid approach is introduced, analyzed, and evaluated. Several switching conditions are discussed considering online performance assessments and design tools proposed in Chapter 3.

4.3.1 Approach and Hybrid Model

Concluding from the drawbacks of conventional techniques the primary objective is to ensure a sufficient control performance over the operating range. A minimal upper bound for the pose error and pose error variance over the whole operating range is desirable. An important, yet, contradictory side condition is to provide a sufficiently large field of view, e.g. in order to keep all visual features in the field of view to assure a controller of full rank. The idea of the novel approach is a dynamical selection of vision sensors for a sufficiently small operating subrange in order to guarantee the desired performance.

A switching visual servoing strategy is proposed. A particular controller and sensor are selected dynamically from a set. The hybrid switching controller is defined as

$$\begin{aligned} \tau &= f^{\eta}(\xi, x(q)q, \dot{q}, \eta) = J_v^{\eta}(\xi, x(q), \dot{q}, \eta)^{+} K_p \tilde{\xi} - K_v \dot{q} + g(q), \\ J_v^{\eta} &\in \{J_{v,1},\, J_{v,2},\, \ldots,\, J_{v,n}\} = \mathcal{J}_v \subset \mathcal{J}^{m,single}, \end{aligned} \tag{4.4}$$

with vectorfield J_v^{η}, which can be switched, e.g., by conditions on ξ and $x(q)$, and a discrete control input $\eta \in \mathcal{P} = \{1, 2, \ldots, n\}$. The set $\mathcal{J}_v$ is a subset of the manifold $\mathcal{J}^{m,single}$ of all possible controllers of single-sensor configurations of rank m corresponding to the number of observed feature points. Within the scope of this work only the focal-length is considered to be a free parameter determining a particular $J_{v,i}$.

Sensor switching is expressed in a hybrid measurement equation

$$\xi = h_{\xi}^{\eta}(x(q), \lambda, \eta) + \nu, \qquad h_{\xi}^{\eta} \in \{h_{\xi,1},\, h_{\xi,2},\, \ldots,\, h_{\xi,n}\} = \mathcal{H}_{\xi} \subset \mathcal{H}^{m,single},$$

with vectorfield h_{ξ}^{η} of the set $\mathcal{H}_{\xi}$ capturing the sensor model dependent on sensor pose $x(q)$ and focal-length λ. Other intrinsics are omitted for better readability. Set $\mathcal{H}_{\xi}$ is a subset of the manifold $\mathcal{H}^{m,single}$ of all possible single-camera sensor configurations of rank m corresponding to the number of observed feature points. Vectorfield h_{ξ}^{η} can be switched by conditions on x and by discrete control input η. Sensor noise ν is considered having the same noise power for all sensors of the set.

In the following stability for this control method is proven based on Lyapunov's direct method. One challenge of this novel strategy is the definition of appropriate switching conditions to ensure the desired control quality. Therefore, various switching conditions are discussed in the following paragraphs based on performance measures and design tools considering dynamic system quantities and internal parameters.

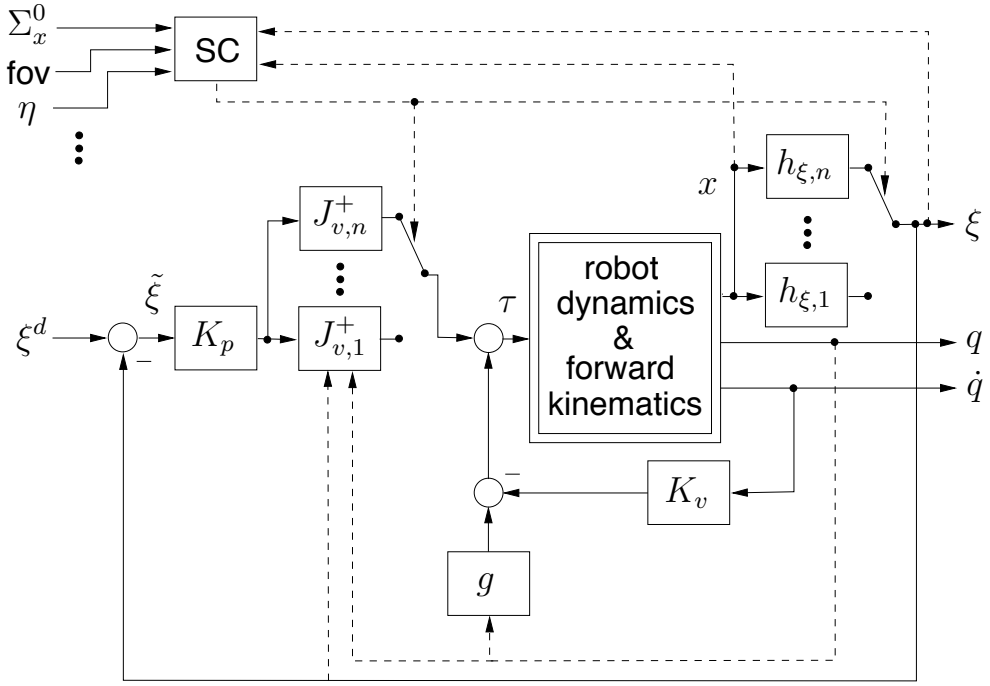

Figure 4.6: Multi-focal hybrid switching visual servoing architecture with switching condition SC.

4.3.2 Stability

A common method for proving stability of a hybrid system is Lyapunov's direct method, which requires a common Lyapunov function or a family of Lyapunov functions (multiple Lyapunov functions) under certain conditions.

Conventional Image-Based Visual Servoing. Stability of image-based visual servoing taking manipulator dynamics into account has been proven in [66]. Quantization effects are not considered. Consider again the manipulator dynamics

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + g(q) = \tau,$$

and the controller

$$\tau = J_v(\xi, x(q), \dot{q})^+ K_p \tilde{\xi} - K_v \dot{q} + g(q),$$

with visual Jacobian J_v, positive definite gain matrices K_p, K_v, feature error $\tilde{\xi} = \xi^d - \xi$, and joint angles q. Implying the existence of a joint configuration q^d where $\tilde{\xi}$ vanishes, an isolated equilibrium $[q^T \quad \dot{q}^T]^T = [q^{dT} \quad 0^T]^T$ can be concluded. Now consider the Lyapunov function candidate

$$V = \frac{1}{2}\dot{q}^T M \dot{q} + \frac{1}{2}\tilde{\xi}^T K_p \tilde{\xi},$$

and the time derivative yielding

$$\dot{V} = -\dot{q}^T K_v \dot{q}.$$

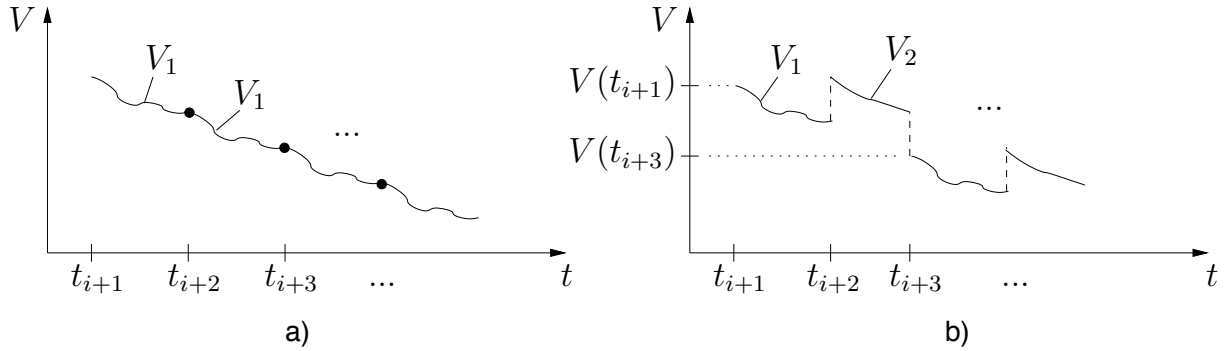

Figure 4.7: Propagation of Lyapunov functions of switched systems; a) common Lyapunov function, b) multiple Lyapunov functions.

Since $\dot{V}$ is a globally negative semidefinite function invoking Lyapunov's direct method a stable equilibrium can be concluded. Local asymptotic stability is proven using Krasovskii-LaSalle's theorem [66]. Due to ambiguities in $\xi(q)$ resulting in local minima global stability is only achievable by additional constraints on the controller as functions of the state-space. Therefore, state-of-the-art approaches from common control literature can be applied. A straight-forward approach is, e.g., the introduction of an additional potential function term $U(\xi, q)$ in the controller f^{η} [23]. However, this aspect does not limit the proposed approach and is, thus, out of scope of this work.

In case of quantized systems, e.g. discontinuous feedback systems typical for visual servoing, the effects of limit cycles and remaining control errors have to be taken into consideration. Methodical investigations in the field of visual servoing are not known in literature, however, recent approaches exist considering general nonlinear dynamics. These effects are not within the scope of this thesis and subject to future research directions.

Multi-focal Hybrid Image-Based Visual Servoing. The proposed visual servoing strategy is assumed to switch between image-based controllers

$$\tau = f^{\eta}(\xi, x(q), \dot{q}, \eta), \tag{4.5}$$

which are stable. If ideally

$$\hat{J}^{+}_{v,i} K_{p,i} J_{v,i} = \hat{J}^{+}_{v,j} K_{p,j} J_{v,j} \quad \forall q \in \mathbb{R}^n, \;\; K_{p,i} = K_{p,j},$$

with $(.)_i$ and $(.)_j$ denoting magnitudes before and after a switch and $\hat{(.)}$ denoting estimated magnitudes, holds, which is the case, if the estimates are exact, i.e. the vector fields of the switched system dynamics $f_{s,i}$ and $f_{s,j}$, cf. (4.3), are equal, then the energy of the control loop does not change during a switch. It is obvious that a common Lyapunov function exists having equal values just before and after a switch (see Figure 4.7a). However, the formulation in (4.3.2) cannot be utilized, as due to $\xi_i \neq \xi_j$ (a change of focal-length results in different $\xi(q)$) a jump of the potential energy occurs. Evaluating the Lie bracket simply yields

$$[f_{s,i}, f_{s,j}] = 0, \quad \forall q, \dot{q} \in \mathbb{R}^n, \text{ as } f_{s,i} = f_{s,j}, \tag{4.6}$$

thus, the switched system is locally asymptotically stable and a common Lyapunov function exists in a sufficiently small region around $[q^{d^T}\ 0^T]^T$ which can be determined, e.g., utilizing Lyapunov's indirect method under the assumption of exponential stability of the subsystems, e.g. cf. [82].

If otherwise

$$\hat{J}_i^+ K_{p,i} J_i \neq \hat{J}_j^+ K_{p,j} J_j,$$

e.g. due to parameter perturbations or different K_p, a jump of the potential energy occurs during a switch possibly increasing the total energy. If the products $\hat{J}_k^+ K_{p,k} J_k$ are equal up to a constant multiplier ϕ, i.e. $K_{p,j} = \phi K_{p,i}$, which is, e.g. the case if the estimated focal-lengths $\hat{\lambda}_{i,j}$ are not known exactly or the $K_{p,i,j}$ are in fact different, ϕ simply acts as an additional dc-gain. If either the focal-lengths or $K_{p,i,j}$ are free parameters then the perturbation can be compensated and a common Lyapunov function exists. Otherwise, for an arbitrary ϕ state-dependent switching conditions can be defined using multiple Lyapunov functions rendering the switched system stable, e.g. cf. [82]. Therefore, it must be assured that the values of the Lyapunov function V_i of subsystem i at the beginning of each time interval where i is active form a decreasing sequence (see Figure 4.7b), i.e.

$$V_i(t_{i+3}) - V_i(t_{i+1}) < 0. \tag{4.7}$$

A switching condition in terms of the state vector can be formulated: $V_i \rightarrow V_j$: $\tilde{\xi}_i(q) = 0$, arbitrary $\dot{q}$.

The value of V_i at time step t_{i+3} can be written

$$V_i(t_{i+3}) = V_i(t_{i+1}) + \int_{t_{i+1}}^{t_{i+2}} \frac{\partial V_i}{\partial q \partial \dot{q}} f_i dt + V_j(t_{i+2}) - V_i(t_{i+2}) + \int_{t_{i+2}}^{t_{i+3}} \frac{\partial V_j}{\partial q \partial \dot{q}} f_j dt + V_i(t_{i+3}) - V_j(t_{i+3}). \tag{4.8}$$

Evaluating (4.3.2) under the assumed condition it is $V_j(t_{i+2}) - V_i(t_{i+2}) = 0$ and $V_i(t_{i+3}) - V_j(t_{i+3}) = 0$. Matrix K_v is by design positive definite yielding

$$\int_{t_{i+k}}^{t_{i+k+1}} \frac{\partial V_i}{\partial q \partial \dot{q}} f_i dt = \int_{t_{i+k}}^{t_{i+k+1}} \dot{V}_i dt < 0, \tag{4.9}$$

and $\dot{V}_i = \dot{V}_j$. Thus, (4.7) holds. The analog can be shown for V_j. Thus, asymptotical stability of the switched system is assured.

This condition, however, is very restrictive. A relaxation can be achieved by observing the kinetic and potential energy of the systems and assure that the sum of the difference terms $V_j(t_{i+2}) - V_i(t_{i+2})$ and $V_i(t_{i+3}) - V_j(t_{i+3})$ does not exceed the energy decrease due to $\dot{V}_i$ and $\dot{V}_j$.

Asymptotical stability for more arbitrary switching and for perturbations of $\hat{J}$ not modeled by a multiplier ϕ can be achieved by the dwell-time approach. It is well known that a switched system is stable if all the individual subsystems are stable and the switching is sufficiently slow, so as to allow the transient effects to dissipate after each switch. This time period can be determined utilizing, e.g., the average dwell-time approach using multiple Lyapunov functions, e.g. cf. [82].

4.3.3 Switching Conditions and Performance Measures

Dynamic selection of an appropriate sensor in a given situation has to satisfy several constraints. A sufficient control performance has to be ensured as well as a sufficiently large field of view, e.g. to capture all feature points or a sufficiently high resolution, e.g. to examine particular parts of the scene. In the following a selection of possible performance measures to be considered in the formulation of switching criteria will be discussed.

- **Pose Variance.** The variance of the pose vector is the most direct method and an absolute measure for visual servoing performance. This measure requires a continuous sufficiently accurate online estimation. Recent works [81] investigate the propagation of measurement errors through the dynamic system. However, exact knowledge of system parameters is necessary. Yet, time delay and manipulator dynamics are not considered. Thus, this method is not yet suited for practical applications. In this thesis a different method is proposed based on direct estimation from manipulator motion, which can be measured. As the variance is time varying the estimate is based on a short time window.

  The pose variance is also the basis for relative measures proposed in the following, which predict a possible increase of performance by switching to a particular sensor evaluating internal system parameters and current variance estimate.

- **Sensitivity of the Controller.** The sensitivity of the controller, i.e. the singular values projected into Cartesian space, in the individual directions of Cartesian space and all other measures based on sensitivity can be used as relative measures. The change of performance in a particular direction by switching sensors can be predicted evaluating the sensitivity of the controller, e.g. the sensor Jacobian. This method is proposed in [102] and [60] in order to control the zoom of a camera introducing the term *resolvability* for sensitivity. In [55] and [47] dependency of sensitivity on feature configurations and the impact on control performance is investigated.

- **Smallest Singular Value** The smallest singular value is a measure for the worst case performance in the corresponding Cartesian direction. This measure is, e.g. used as a global non-directional measure in selection problems, e.g. active vision planning [45].

- **Perceptibility.** Perceptibility introduced by [118] in analogy to manipulability is a global measure mainly used in optimization problems as trajectory planning, e.g. [34, 118]. Perceptibility is a well suited measure for multi-camera configurations as pointed out in Chapter 3.

- **Condition of the Controller.** The condition number is a global measure and also used in optimization problems as above, e.g. [47]. However, in case of the proposed visual servoing strategy this measure turns out unsuited as it increases with distance, in many cases its value changes only very little during switching and does not capture the actual sensitivity of the controller to measurement noise.

- **Focal-length.** The focal-length is a straight forward relative measure to predict the change of performance during sensor switching. Control performance changes by the quotient of the next focal-length by the current one in case of single camera configurations.

- **Image Feature Position.** The distance of feature points to the borders of the visible image plane can be evaluated to prevent from feature points leaving the image plane and potentially rendering the system unstable due to resulting singularities of the visual controller. E.g., feature positions near the image center indicate weak control performance due to weak resolution and sensitivity.

- **Image Moments.** Image moments, e.g. to evaluate the projected area of the observed object, are another measure known from literature, e.g. [23], and a possible indicator for control performance.

- **Field of View / Resolution** The desired field of view and the resolution can be considered situational parameters determining whether a larger part of the scene or a particular part with higher resolution has to be observed. These conditions can be considered as discrete control input to the proposed controller.

Within the framework of the proposed hybrid multi-focal visual servoing strategy the focal-length and perceptibility are considered the most suited measures to quantify the expected change of performance due to sensor switching. The focal-length is a relative measure for sets of single camera configurations. Perceptibility is considered for multi-camera and mixed single- and multi-camera configurations due to its global character and quantification of the sensitivity of the controller to measurement noise. As both performance measures are relative measures the current pose variance is evaluated and the expected change of performance is computed relative to this estimate.

The switching condition for the proposed hybrid strategy is formulated based on a performance region Σ_x^0 expressed by a polyhedron in Cartesian space bounded by hyperplanes in the individual Cartesian directions. Special cases are given by the performance metrics in Appendix B, e.g. translation error variance or by the restriction to, e.g., the worst case or most sensitive Cartesian direction. An example is the pose error variance in the direction of the optical axis of the camera. In these cases the polyhedron is reduced to a simple performance band σ_x^0. Performance is either measured by pose error or pose error variance.

When the bounds of the polyhedron Σ_x^0 are met by the performance metrics Σ_x the expected performance $\Sigma_{x,j}^*$ for each sensor of the set $\mathcal{H}_\xi$ is evaluated and a switch to the one suited best under the side-conditions for the expected field of view and/or resolution Ψ^* is triggered

$$\begin{aligned} \mathcal{H}_\xi^\Sigma &= \left\{ \mathcal{H}_\xi^\Sigma \subset \mathcal{H}_\xi \,\middle|\, \|\Sigma_{x,j}^*(h_{\xi,j})\| < \|\Sigma_x^0\| \right\}, \\ < J_v^\eta,\, h_\xi^\eta > &= \left\{ < J_{v,j} \in \mathcal{J}_v, h_{\xi,j} \in \mathcal{H}_\xi > \,\middle|\, \|\Sigma_{x,j}^*\| < \|\Sigma_x^0\| \wedge \Psi^* = \underset{\Psi_j(h_{\xi,j}), h \in \mathcal{H}_\xi^\Sigma}{\arg\max} \sum_j \Psi_j \right\}. \end{aligned} \tag{4.10}$$

As $\Sigma_{x,j}^*$ depends on ξ and q the switching condition is state-dependent.

Focal-length and perceptibility are proposed in order to predict the expected performance after a switch. The corresponding relations are

$$\sigma_j^* = \frac{\lambda_i}{\lambda_j}\sigma_i,$$

$$\sigma_j^* \quad = \quad \sqrt[n]{\frac{w_i}{w_j}}\sigma_i,$$

where i refers to the current sensor configuration, j refers to a particular sensor configuration from the set, $\sigma_{(.)}$ are the corresponding standard deviations, $\lambda_{(.)}$ are the focal-lengths, $w_{(.)}$ are the perceptibilities, and n is the rank of the controller. If the controller is well conditioned, i.e. condition number $\kappa = 1$, the perceptibility relation holds. Otherwise, the change of performance is underestimated. This measure is, thus, rather coarse. However it is well suited for multi-camera configurations, where relations based on focal-length can be rather complex.

Example Consider the following example for better clarity. The translational pose error variance $\sigma^2_{trans,z}$ in z-direction (optical axis) is chosen as performance metrics. The polyhedron, thus, reduces to a variance band

$$\sigma_z^0 = v,$$

with an arbitrary threshold v. The side-condition is a maximum possible field of view, thus,

$$\Psi = \alpha,$$

with aperture-angle α. A set $\mathcal{H}_\xi$ of two sensors is considered. Then, the switching condition simplifies to

$$\mathcal{H}_\xi^\Sigma = \left\{ \mathcal{H}_\xi^\Sigma \subset \mathcal{H}_\xi \middle| \sigma^{2*}_{z,j}(h_{\xi,j}) < v \right\}, \quad \mathcal{H}_\xi = \{h_{\xi,1},\, h_{\xi,2}\}, \quad \mathcal{J}_v = \{J_{v,1}(h_{\xi,1}),\, J_{v,2}(h_{\xi,2})\},$$

$$< J_v^\eta,\, h_\xi^\eta >= \left\{ < J_{v,j} \in \mathcal{J}_v, h_{\xi,j} \in \mathcal{H}_\xi > \middle| \sigma^{2*}_{z,j}(h_{\xi,j}) < v \wedge \alpha^* = \underset{\alpha_j(h_{\xi,j}),h \in \mathcal{H}_\xi^\Sigma}{\arg\max} \sum_j \alpha_j \right\}.$$

The expected performance, i.e. variance or standard deviation, can be predicted based on the focal-lengths of the current sensor and the sensors of the set

$$\sigma^*_{z,j} = \frac{\lambda_i}{\lambda_j}\sigma_{z,i},$$

where i refers to the current sensor configuration, j refers to a particular sensor configuration from the set, and $\lambda_{(.)}$ are the corresponding focal-lengths.

The proposed switching condition can also be formulated vice versa for the field of view or the resolution criteria, so that as long as the minimum acceptable field of view or resolution is not violated the best performing sensor is chosen.

A third strategy is based on optimization. Each of the criteria is weighted, the weighted criteria are combined in an objective function and the particular sensor giving the maximum value of this objective function is chosen dynamically. Drawbacks of this strategy are that weights have to be chosen heuristically, control performance is not bounded and the variability of the performance is potentially high. In order to overcome these drawbacks, this strategy can be combined with one of the preceding.

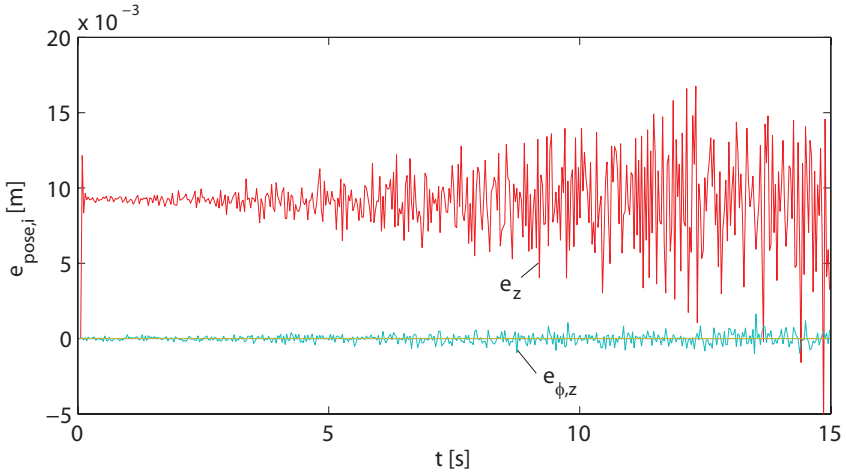

Figure 4.8: Tracking error $e_{pose,i}$ of mono-focal visual servoing trajectory following task along the optical axis over time; focal-length $\lambda = 10$mm; same parameter setting as in Figure 4.2.

4.3.4 Multi-Focal Visual Servoing Performance

The impact of the proposed visual servoing strategy on control performance is evaluated in simulations based on the system model defined in Section 4.1.2. The manipulator dynamics are modeled by a simple decoupled mass-damper-system. Manipulator geometry is neglected. Joint and Cartesian spaces are, thus, equivalent. Two trajectory following tasks are defined:

a) a desired trajectory given by a pure translation along the optical axis with respect to the observed object

$$x^d(t) = \begin{bmatrix} 0 & 0 & -0.4\frac{m}{s}t - 1m & 0 & 0 & 0 \end{bmatrix}^T,$$

A set $\mathcal{H}_\xi$ of three sensors with different focal-lengths of $\lambda \in \{10mm, 25mm, 40mm\}$ and corresponding visual controllers $\mathcal{J}_v$ based on the visual Jacobian are defined. For this task a simple distance-dependent switching strategy is used with switching points $z_1 = -2.6m$ and $z_2 = -4.2m$ from the object. The sensors are switched so that the focal-length increases with distance. For comparison the same task is performed mono-focal with only one camera and a focal-length $\lambda = 25mm$.

b) the desired trajectory defined in Section 4.2 consisting of translation along and rotation about the optical axis. A set $\mathcal{H}_\xi$ of three sensors with focal-lengths of $\lambda \in \{10\text{mm}, 20\text{mm}, 40\text{mm}\}$ and corresponding controllers $\mathcal{J}_v$ based on the visual Jacobian are defined. The switching condition (4.10) is used with a variance band of $\sigma_z^0 = 6.25 \cdot 10^{-6}\text{m}^2$ and a side-condition to provide a maximum field of view.

For both tasks the same parameter setup as in Section 4.1.2 is used. Simulations are conducted using MATLAB/SIMULINK.

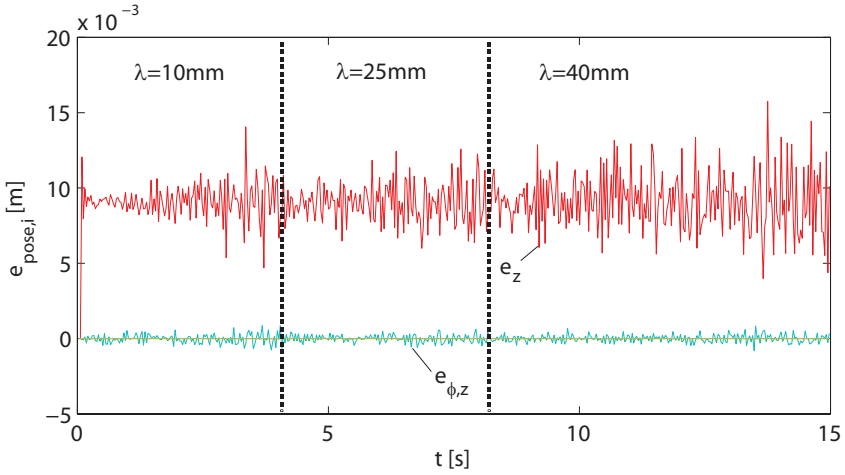

Figure 4.9: Tracking error $e_{pose,i}$ of multi-focal visual servoing trajectory following task along the optical axis over time; focal-lengths $\lambda_{0s\leq t<4s} = 10$mm, $\lambda_{4s\leq t<8s} = 25$mm, $\lambda_{8s\leq t\leq 15s} = 40$mm; same parameter setting as in Figure 4.2.

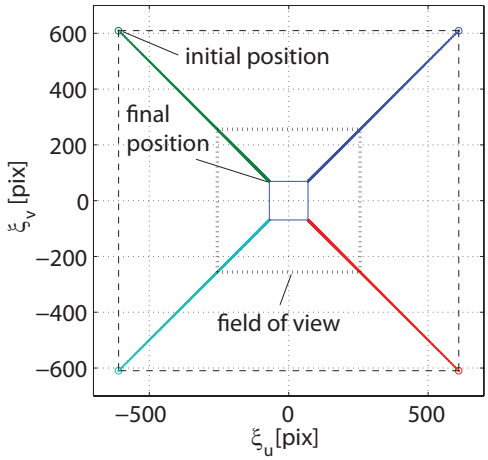

Figure 4.10: Feature point trajectories in image space for mono-focal visual servoing translation task along the optical axis corresponding to Figure 4.8.

The results are shown in Figure 4.8 to Figure 4.12. Figure 4.8 and Figure 4.9 show the tracking error in Cartesian coordinates for the mono-focal and the multi-focal task a). The variability of the tracking error increases with distance. The tracking error of the mono-focal setup shows a good performance in nearer distances up to about -4m, which corresponds to $t = 7.5$s, and particularly good performance in a range up to about -2.5m corresponding to $t = 4$s. Beyond -4m the variability and, thus, the variance becomes notably high rendering the overall visual servoing performance poor. It is emphasized that for the range of a very good performance up to about -1.2m the corresponding feature points in the image plane lie outside of the visible range due to the limited field of view as shown in Figure 4.10. Thus, this range is in fact useless as the real system would be unstable.

The tracking error of the multi-focal setup in Figure 4.9 shows significantly less variability than the mono-focal case. Up to a distance of about 6.5m at $t = 14s$ the variance is

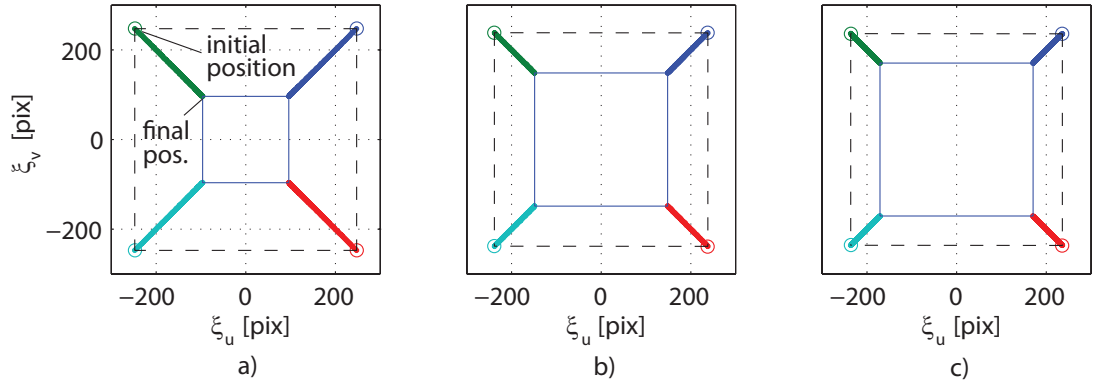

Figure 4.11: Feature point trajectories in image space for multi-focal visual servoing translation task along the optical axis corresponding to Figure 4.9; a) 0s $\leq t <$ 4s, b) 4s $\leq t <$ 8s, c) 8s $\leq t \leq$ 15s.

approximately constant. At nearer distances the variance is slightly higher, but as can be seen in Figure 4.11 all feature points lie in the visible image plane.

The integrated tracking errors for the translational components over the displayed time interval are

$$\begin{aligned} \int_{t=0s}^{t=15s} \|\hat{e}_{tran,mono}(t) - \hat{\bar{e}}_{tran,mono}\|_2 dt &= 0.12ms, \\ \int_{t=0s}^{t=15s} \|\hat{e}_{tran,multi}(t) - \hat{\bar{e}}_{tran,multi}\|_2 dt &= 0.08ms, \end{aligned}$$

with the mean $\hat{\bar{e}}_{tran}$ of the tracking error $\hat{e}_{tran}$. An improvement of tracking performance by 33% can be noted.

Figure 4.12 shows the results for the multi-focal trajectory following task b). The standard deviation (Figure 4.12b) is kept within a small band reaching from about 0.004m to 0.008m. The standard deviations of the multi-focal strategy reach approximately the values of the mono-focal task in Section 4.2.1 for the corresponding focal-lengths and intervals, but the overall variability is significantly lower. The spikes, which can be noted in the standard deviation diagram are caused by the switches. After a switch the desired feature value changes with the sensor, but the current value is still taken from the previous sensor due to time delay. Thus, the control error at this time instance is very high. This effect can be reduced by mapping the previous value of the feature vector to the image space of the new sensor or by definition of a narrower variance band as switching condition.

Figure 4.13 exemplarily illustrates the progression of the field of view over time for another mono-focal translation task and the corresponding multi-focal task for a motion along the optical axis. The field of view is defined by the visible part of the plane extending the surface of the observed object in x-direction. A simple pinhole camera model is assumed. For the multi-focal task a lower bound of about 1m resulting from the chosen switching condition can be noted. A higher lower bound is achievable only by accepting a larger variance band or different cameras, thereby, accepting higher pose error variances. The trade-off between control performance and field of view is obvious.

The effectiveness of the proposed multi-focal switching strategy has been shown successfully. The contributions of this novel approach are a guaranteed control performance

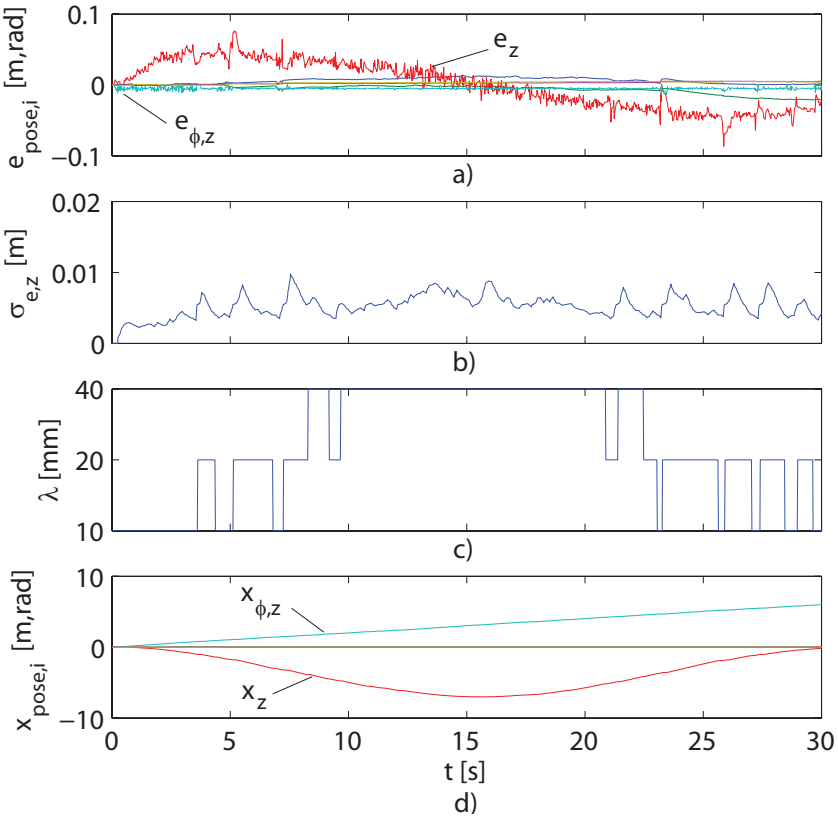

Figure 4.12: Multi-focal visual servoing trajectory following task results; a) tracking errors $e_{pose,i}$, b) short-time tracking error standard deviation estimates $\sigma_{e,z}$, c) current selected focal-length λ, and trajectory $x_{pose,i}$ over time t; same parameter setting as in Figure 4.2.

shown by means of a bounded pose error variance, a low variability of the performance over the whole operating range, and the consideration of situational side-conditions as, e.g., a maximum field of view.

4.3.5 Discussion

In this section a novel hybrid visual servoing strategy based on multi-focal vision has been proposed. An innovative parametrizable dynamic camera switching strategy has been introduced to overcome the investigated drawbacks of conventional visual servoing techniques in wide operating range scenarios. It has been shown how situational and performance issues are considered within the definition of proposed switching conditions. An essential means for the definition of switching conditions is the prediction of the expected change of control performance considering system states and internal parameters of the visual controller. Therefore, performance measures and design tools are proposed considering findings from Chapter 3.

The impact of the proposed approach has been evaluated, compared with mono-focal setups, and successfully demonstrated. The strategy contributes to ensure high control performance over large operating ranges and is suited for combination with many of the visual servoing approaches from known literature. Moreover, it facilitates asymptotically stable visual servoing if an overall mission or the current situation requires dynamical

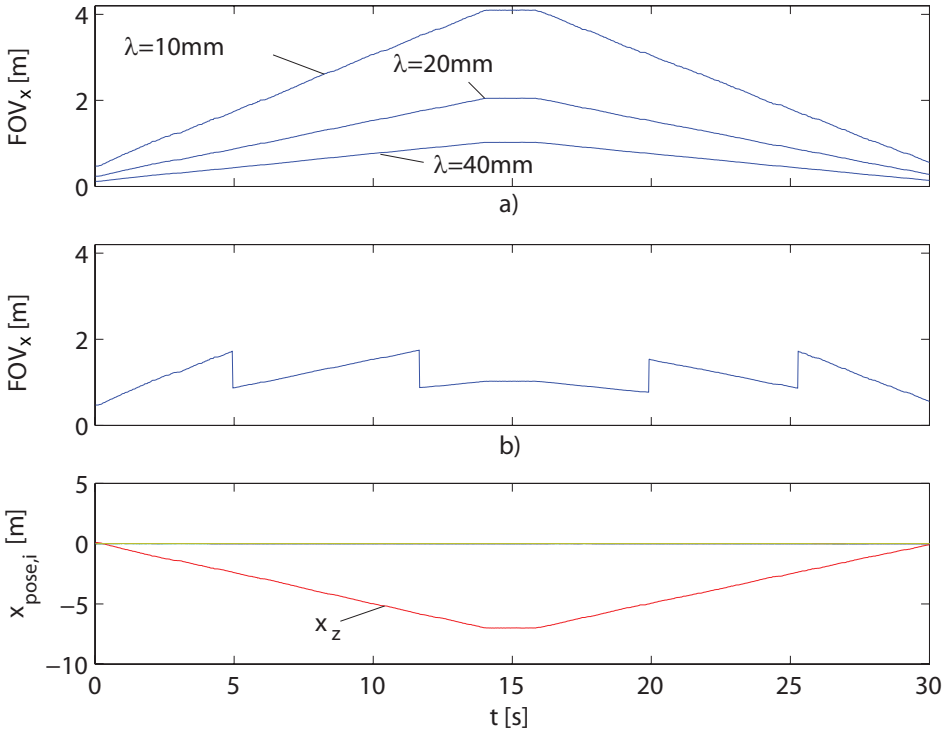

Figure 4.13: Progression of the extension of the field of view FOV_x in x-direction orthogonal to the optical axis at distance x_z from the vision sensor for a a) single-camera visual servoing task and the b) proposed switched camera visual servoing task with pose trajectory x_z.

changes of the sensor configuration. Examples are simultaneous scene observation dynamically allocating individual sensors to particular tasks and partial sensor breakdown where sensors of different characteristics are taking over.

4.4 Multi-Camera Strategies

The proposed hybrid controller dynamically changes the configuration of the vision system. In the foregoing section the focal-length has been used as a free parameter in order to achieve a desired control performance of the switched system. Thereby, the focal-lengths of all sensors have been varied equally and simultaneously. In many situations it is not required or not possible to switch the focal-length of the whole vision system instantaneously, e.g. if an available high-sensitivity sensor does not cover all the feature points to be observed in order to make the controller full rank due to its limited field of view. Therefore, in this section it is proposed that the focal-lengths of all simultaneously used sensors may change independently. An example is the observation of a selected feature point with a high-sensitivity sensor and the remaining feature points with a large field of view, but a low-sensitivity sensor. Utilizing the performance criterion proposed in Section 3.4.2 the possible configurations of the available set of vision sensors can be evaluated towards sensitivity in particular task-relevant Cartesian directions. Finally, a selection of sensor configurations is made based on task-requirements on the control performance. In

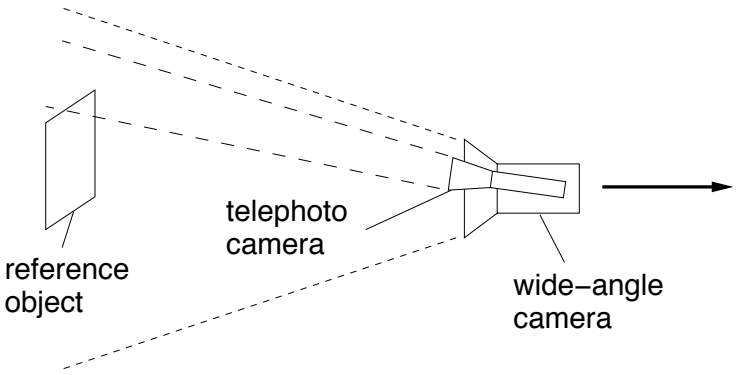

Figure 4.14: Multi-focal visual servoing task with wide-angle and telephoto camera simultaneously observing a reference object; note the different fields of view marked by dashed lines.

the following sections, the selection criterion for the switching condition is defined and the performance is exemplarily evaluated in a translational visual servoing task.

4.4.1 Switching Condition

In Section 4.3.3 the condition for the dynamical selection of the optimal sensor for a multi-focal visual servoing task has been formulated for a single sensor with particular performance characteristics. In this section this condition is generalized towards configurations of multiple sensors with individual characteristics. Thereby, from a set of available vision sensors a particular combination of sensors is selected dynamically in order to satisfy some performance criteria for the visual servoing task.

The switching condition (4.10) is generalized to the multi-sensor formulation in a straight forward manner formally allowing the combination of several sensors in the measurement equation (4.3.1) for $h_{\xi,i}(x(q),\lambda)$. The selectable sensor is an element of a subset of the manifold $\mathcal{H}^{m,multi}$ of all possible multi-sensor configurations of rank m corresponding to the number of combined sensors in the vision device

$$h_{\xi}^{\eta}(x(q),\lambda,\eta) = \{h_{\xi,1}, h_{\xi,2}, \ldots, h_{\xi,n}\} = H_{\xi} \subset \mathcal{H}^{m,multi}, \tag{4.11}$$

with corresponding Jacobian of the visual controller

$$J_{v}^{\eta}(\xi(q),\dot{q},\lambda,\eta) = \{J_{v,1}, J_{v,2}, \ldots, J_{v,n}\} = J_{\xi} \subset \mathcal{J}^{m,multi}, \tag{4.12}$$

where λ is henceforth a vector containing the focal-lengths of the individual sensors

$$\lambda = [\lambda_1\ \lambda_2\ \ldots\ \lambda_n]^T. \tag{4.13}$$

In order to define a performance region Σ_x^0 for switching condition (4.10), the predicted performance after a switch has to be computed. The change of focal-length as proposed in Section 4.3.3 cannot be utilized anymore as several and potentially different focal-lengths of the individual sensors exist. Therefore, the evaluation of the performance in particular Cartesian directions is proposed as suggested in the design considerations for a general

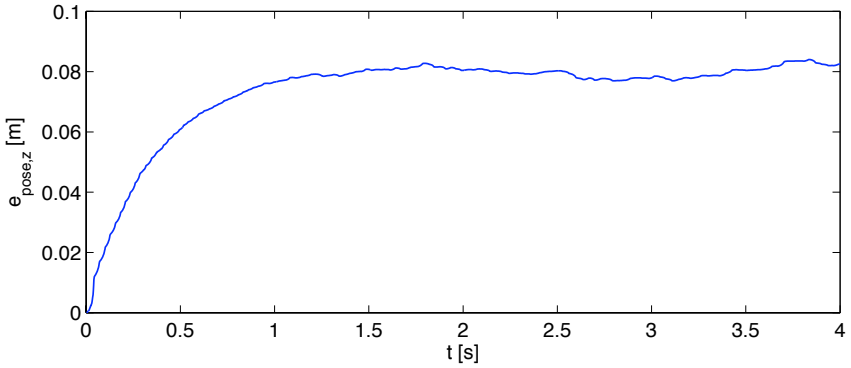

Figure 4.15: Tracking error $e_{pose,z}$ of multi-focal switched visual servoing trajectory following task along the optical axis; desired trajectory $x_z^d(t) = -0.2ms^{-1}t - 1\text{m}$; focal-lengths $\lambda_{0s \le t < 2.6s} = 0.005\text{m}$, $\lambda_{2.6s \le t \le 4s} = 0.040\text{m}$; observed object: square with 0.5m edge lengths at x_z^d; inertia matrix $M = 0.5\text{diag}(1\text{kg}, 1\text{kg}, 1\text{kg}, 1\text{kgm}^2, 1\text{kgm}^2, 1\text{kgm}^2)$, damping $K_v + C = 200\text{diag}(1\text{kgs}^{-1}, 1\text{kgs}^{-1}, 1\text{kgs}^{-1}, 1\text{kgms}^{-1}, 1\text{kgms}^{-1}, 1\text{kgms}^{-1})$, feedback quantization 0.000001m, sensor noise power $\sigma_{meas}^2 = 0.000001^2\text{m}^2$, control gain K_p tuned to converge system after approximately 2s.

multi-focal vision system in Section 3.4.2. Utilizing (3.14) the polyhedron Σ_x^0 is, thus, bounded in the individual Cartesian directions by the corresponding directional sensitivities $\|_r r_i^*\|$, $i = 1, 2, \ldots, n$ of the potential multi-sensor controllers $J_{v,j}^{\eta}$, and comparing these with the current ones.

4.4.2 Example Multi-Camera Task

In order to demonstrate the benefits of multi-focal multi-camera visual servoing, consider again a trajectory following task along the optical axis using a vision system observing a squarish object of four feature points. The vision system consists of two cameras, one wide-angle camera observing three of the features and a switchable camera observing the remaining feature point comprising either a wide-angle or a telephoto characteristic. To simplify matters both cameras are assumed coaxial.

As the task-relevant control performance is the variance of the tracking error in direction of the optical axis, the sensitivity of the visual controller in this direction is considered. The sensitivity of the current controller, i.e. two wide-angle cameras, and of the predicted controller, i.e. after switching the second camera to telephoto characteristic, is computed continually. Once the tracking error variance of the controller exceeds a threshold of $\sigma_{e,z} > 0.00004\text{m}$ the controller and the second camera are switched to telephoto. In this example switches are only allowed after a time of 2s when the system has settled to a constant tracking error.

Figure 4.15 shows the tracking error along the optical axis of the switched system. The standard deviations of the tracking errors of the wide-angle system, of the multi-focal system, and of the multi-focal switched system are shown in Figure 4.16. At $t = 2.6\text{s}$ the controller and sensor are switched resulting in a reduced tracking error standard deviation with lower variability compared to the wide-angle case. The unswitched multi-focal system, i.e. where the second camera is constantly used in telephoto mode observing only one feature point of the object, shows an even lower standard deviation. The corresponding

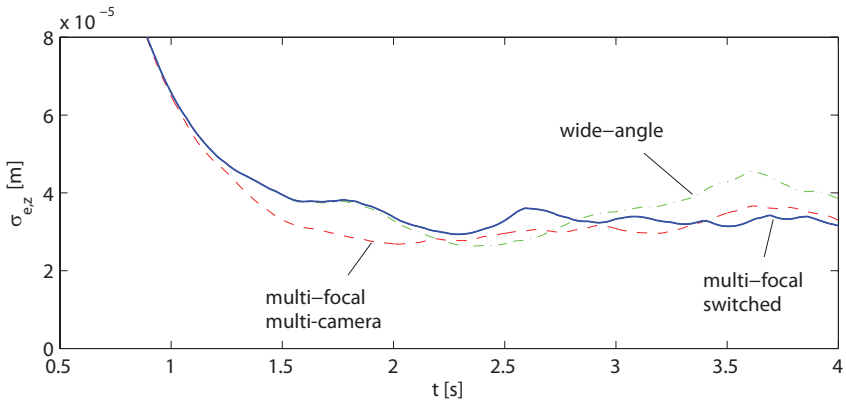

Figure 4.16: Short-time standard deviation estimate $\sigma_{e,z}$ of tracking error in Figure 4.15 of *multi-focal switched* visual servoing task, of corresponding unswitched mono-focal (*wide-angle*) task with $\lambda = 0.005$m, and of unswitched *multi-focal* task where one corner of the reference square object is observed with $\lambda = 0.04$m and the other features with $\lambda = 0.005$.

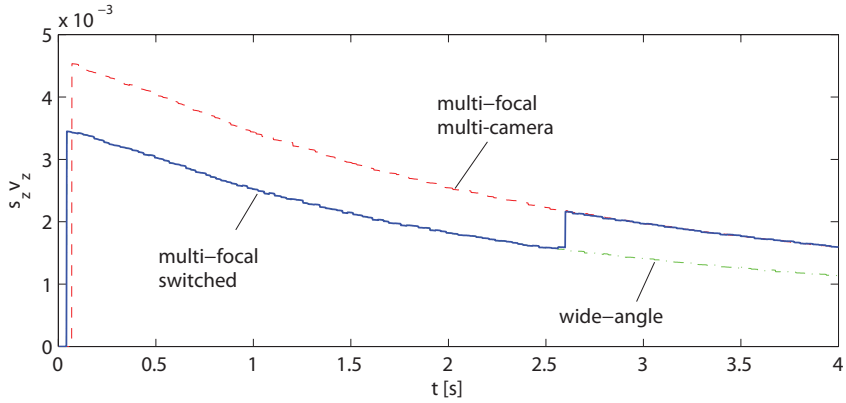

Figure 4.17: Corresponding sensitivities $s_z v_z$ of the visual servoing controller in task (z-)direction; corresponding singular value s_z of J_v and element v_z of matrix V of $J_v = U\Sigma V^T$.

sensitivities of the controllers are shown in Figure 4.17 by the products of the singular value σ_z of J_v and the corresponding element of the eigenvector of V of $J_v = U\Sigma V^T$ in the direction of the optical axis.

This example demonstrates several benefits of multi-focal vision in visual servoing. In addition to the strategy in the previous section where only one sensor at a time is active, control performance can also be improved by using a multi-focal vision system with several sensors simultaneously and switching the individual sensors independently. Secondly, a wide field of view is provided and a desired control performance can be assured not only by switching between vision sensors, but by observing only one or more selected feature points of the observed object with a high-sensitivity sensor. These observed features can, e.g., be selected similarly to the example in Section 3.4.1 by determining the feature for which the controller yields the highest sensitivity.

4.5 Discussion

Visual servoing suffers from a distance-dependent pose error and pose error variance. Using conventional visual servoing techniques this drawback cannot be overcome due to the trade-off between the field of view and the sensitivity of the vision system. A minimum field of view has to be guaranteed in order to be able to observe a reference object at the nearest distance. However, this constraint results in a weaker sensitivity at larger distances to the observed object and increased pose errors and variances.

In this chapter novel visual servoing strategies have been proposed based on multi-focal vision which overcome these drawbacks guaranteeing a desired control performance over a whole operating distance. Several possibilities to exploit the benefits of multi-focal vision have been proposed and evaluated in extensive simulations: Serial switching between vision sensors of different characteristics based on performance-dependent switching conditions, usage of several vision sensors of different characteristics at the same time, and individual switching of one or more of these simultaneously used sensors. Stability has been discussed utilizing common and multiple Lyapunov functions, respectively.

It has been shown that each of the proposed strategies significantly improved the visual servoing performance by reduction of the pose error variance. Depending on the application scenario several guidelines for using multi-focal vision can be given. If only one vision sensor at a time is selectable then a dynamical sensor selection satisfying desired performance constraints and side-conditions is proposed. If several vision sensors can be used simultaneously selected features of a reference object can be observed with higher-resolution sensors while a large field of view sensor ensures observation of a sufficient number of features in order to render the visual controller full rank. The higher-resolution sensors should preferably be focused on those feature points causing the highest sensitivity of the controller.

In this work, quantization effects have been considered in the simulations. However, the proof of stability is pending. The phenomenon of limit cycles due to quantization and parameter perturbations has been mentioned. The characterization and quantification dependent on quantization levels and parameter uncertainties is still an open problem in visual servoing. Another aspect not yet considered is the influence of feature extraction methods on servoing performance which is particularly relevant when sensors with significantly different quantizations typical for multi-focal vision are used together. These aspects are subject to future research directions.

5 Multi-Focal View Direction Planning for Mobile Robots

In the preceding chapters static and dynamic performance characteristics of multi-focal vision have been investigated. This chapter focuses on multi-focal vision from a higher-level information-specific perspective considering the information gained by the individual vision sensors and the mission to be accomplished by a robot. Given a set of independent active vision sensors with different performance characteristics and a robot's mission the scientific question being answered in this chapter is: *what* to observe *when* with *which* sensor.

This objective is commonly known as sensor planning problem, e.g. in a navigation context of autonomous systems. Different types of sensors are fused whereas vision sensors are the most common sensors to be controlled actively. Works considering mono- or stereo-vision setups are known. Approaches to determine an optimal view direction for the current situation are mostly based on maximization of some information measure over some limited time horizon, i.e. an optimal solution for a several steps ahead planning. Works considering several independent active vision sensors with different performance characteristics are not known to the best of the author's knowledge.

The innovation of this chapter consists in a multi-focal approach towards active vision for the navigation of mobile robots with two or more independent active vision devices of different types, i.e. field of view and accuracy. The main goal is an improvement of the robot's mission performance by multi-focal vision sensor planning defined as the accuracy of localization and perception of the environment. Contributions are higher performance, higher efficiency, and flexible sensor resources allocation compared to mono-focal embodiments. Key challenges are conditions for view direction planning considering accuracy and field of view, mission-specific multi-focal multi-camera view direction planning strategies, and the comparison with mono-focal state-of-the-art approaches. The planning methods are exemplarily applied to a humanoid locomotion task considering a global path to be covered and a number of reference objects to be observed within the environment.

The remainder of this chapter is organized as follows: The basic framework and assumptions are defined in Section 5.1. The considered perception, robot, and environment models as well as the data fusion approach using vision sensors and odometry are described in Section 5.2. Section 5.3 is concerned with conditions for view direction selection and a multi-focal view direction planning strategy. The planning strategy is evaluated in extensive simulations. Concluding remarks towards multi-focal planning architectures considering several competing tasks are given in Section 5.4.

5.1 Problem Definition

The purpose of this work is the investigation of multi-focal camera coordination mechanisms for the navigation of mobile robots. In this section framework and approach are defined.

A mobile robot equipped with internal sensors for the estimation of its velocity and gear angle is capable of estimating its position and orientation with respect to its pose in a previous time step. This is called dead-reckoning or odometry. In order to estimate the robot pose with respect to a reference coordinate frame the whole chain of relative homogeneous transformations from a known initial pose within this frame has to be considered. Due to measurement errors and slippage the pose estimations, i.e. relative transformations, are erroneous. The errors accumulate with time causing a drift of the estimated robot pose. In order to overcome the drift problem absolute measurements can be taken, e.g. evaluating visual information which is the focus of this work. These absolute measurements can be used to simply reset the robot pose when available or to be fused dynamically with odometry data. Common fusion approaches are based on Kalman- and Particle-Filters or set-based methods. Thereby, relative and absolute measurements complement one another combining their advantages of high bandwidth and high absolute accuracy, respectively.

The use of active vision systems for navigation is state-of-the-art and has significant advantages over passive systems as, e.g., omnidirectional systems which are the most common passive embodiment: A selective allocation of sensor resources with a higher measurement accuracy than could be achieved with omnidirectional systems with equal pixel matrix of the sensor. Prominent works in this field are, e.g., simultaneous localization and mapping (SLAM) with active vision, e.g. [29, 30, 130] and visual guidance of humanoid robots [45]. The objective of controlling the camera view direction in order to satisfy one or more tasks, e.g. robot localization, leads to a selection problem. Thereby, the "most appropriate" among several possible view directions has to be selected in order to meet the relevant task requirements "sufficiently". Works in this field are manifold. Prominent works cover, e.g., foundations of human overt attention, computational neuroscience approaches, and technical application-oriented approaches, cf. e.g. [45, 62, 68, 136]. Most approaches in mobile robotics use optimization or decision-theoretic techniques, e.g. [29, 45].

Active vision systems comprising only one type of vision sensors face a tradeoff between accuracy and field of view due to limited computational resources. Within the context of robot navigation this implies a tradeoff between localization accuracy and keeping a large part of the scene in view. If a map of reference objects and the current robot pose are known with sufficient accuracy then a vision sensor of the highest possible accuracy and a field of view matching the size of the largest object could be chosen. However, potentially interesting or dangerous objects and events in the local environment could not be detected. Within the scope of navigation for exploration a sufficiently large field of view is required in order to make out a next possible reference object along the robot's path while fixating a current one for localization. Therefore, accuracy is strongly limited. Another drawback is the increasing uncertainty of the position of a reference object while not being observed. Thus, the probability of the object being actually located outside the field of view rises and the object might not be seen even if the camera is directed towards its believed position. Another aspect is the fact that in many situations it is not necessary to provide high accuracy and a wide field of view simultaneously. For example if only a point features have to be measured for localization with high accuracy and the rest of

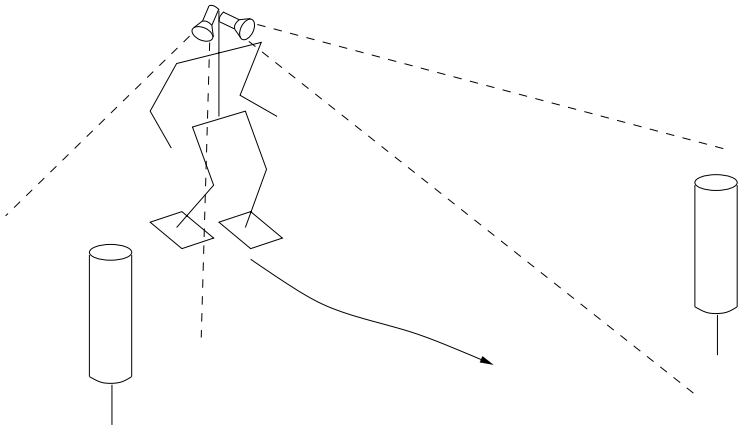

Figure 5.1: Humanoid robot navigation scenario with multi-focal vision.

the scene is only assessed towards the presence of some stimuli then observing the whole scene with high accuracy would cause unnecessary high computational and perceptual costs. Thus, one sensor providing high accuracy *and* field of view would not only require extensive computational resources, but produce a vast amount of unnecessary data with low usable information density.

A multi-focal approach to active vision is proposed in this chapter in order to overcome these drawbacks. Thereby, high accuracy and wide field of view are provided simultaneously and independent of one another. Embodiments of a multi-focal vision system with two stereo-cameras with different accuracies and fields of view are investigated. In the following sections the approach is outlined and criteria for the coordination of the individual sensors are given.

5.1.1 Assumptions, Scenario, and Approach

Considered is a locomotion task of a mobile robot where the robot moves along a preplanned path. It has visual and odometrical capabilities such that it is able to localize itself and other objects within the environment. The robot is equipped with a multi-focal vision system consisting of two stereo-camera devices with independently controllable pan- and tilt-angles, different focal-lengths, and different fields of view. The robot's mission is to follow the desired path. Therefore, it has to localize itself continually evaluating odometry data and visual information. Given a particular environmental situation, i.e. configuration of observable objects and robot pose, the objective is to dynamically select appropriate view directions for both vision devices. Figure 5.1 exemplarily shows a situation in the considered navigation scenario where a humanoid robot fixates two landmarks with two vision devices of its multi-focal vision system in order to localize itself in the world.

The proposed approach consists in an optimization of the pan-/tilt-angles of both vision devices with respect to mission-relevant tasks. Within the scope of the investigations conducted in this thesis a minimization of the uncertainty of the robot's pose error is considered in order to achieve the mission - accurately following the planned path - optimally. Therefore, the current robot state (pose) in the environment and the environment (object positions) are considered. This approach extends the method of [45] to the general multi-

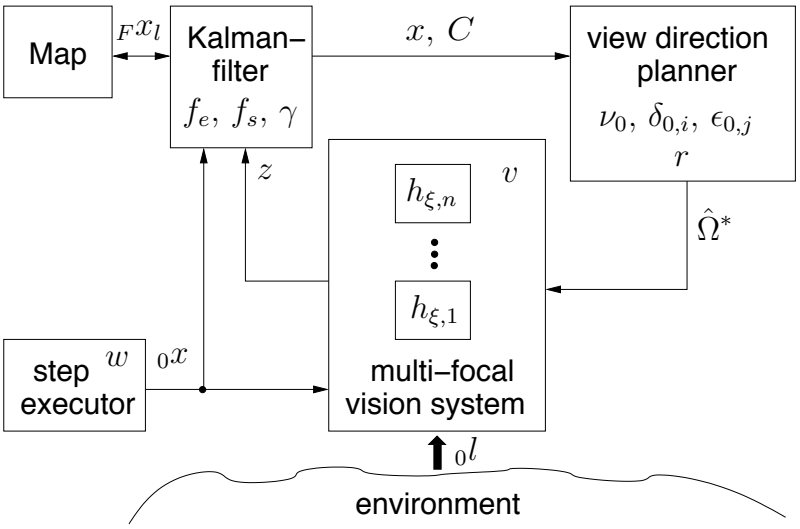

Figure 5.2: Multi-focal view direction planning architecture and simulation layout.

focal case. Several conditions for determining view directions are defined in Section 5.3.1 taking into account the robot pose uncertainty and the fields of view of the cameras.

The pose of the robot and positions of objects in the environment are estimated by an extended Kalman-filter. The measurement equation captures the two stereo-cameras whereas for an object visible for the camera with higher resolution only this camera is considered for the position measurement of the object. Several configurations are investigated which differ in the optical characteristics of the cameras and in whether the relative poses of two camera pairs are fixed.

Figure 5.2 shows the proposed system architecture composed of a view direction planner computing the view directions for each of the vision devices of the multi-focal vision system. Depending on whether or not landmarks are visible for the vision system the current robot pose is visually measured or predicted based on a robot locomotion model, respectively. This data fusion is done by a Kalman-filter. In the following sections these components are explained in detail.

5.1.2 Considerations and Conditions for Camera Coordination

Various aspects have to be considered in order to derive mathematical constraints for a task-dependent control of the camera view directions of a multi-focal vision system. These considerations define requirements on accuracy and field of view of the individual sensors.

Robot Pose. If the mission includes some particular path of the robot with certain constraints regarding locomotion accuracy the error of the robot pose is an important aspect to be considered. Due to odometry errors the absolute error of the estimated robot pose increases as the robot moves. By taking position measurements of objects the positions of which are known sufficiently accurate the absolute robot pose can be estimated and the robot pose error can be reduced. An approach of view direction planning for a single vision sensor is to select that view direction in the next time step at which the robot pose error is maximally reduced. The basic principle is a prediction of the reduced pose uncertainty corresponding to the information gained in the next time step for each

possible view direction. An optimization problem is then solved searching for minimum pose uncertainty. This approach has successfully been applied to vision guided humanoid walking [45]. An extension proposed in this thesis is the consideration of more than one sensor. The predicted reduction of the robot pose uncertainty in the next time step is computed for each possible set of camera view directions. Accordingly, the optimal solution of view direction selection is that set of directions with which the robot pose uncertainty is minimized. A possible drawback of this approach are the high computational costs as the search space is extended by one dimension per sensor. The computational costs increase further if the optimization problem is not only solved considering the information gained in the next time step, but by several steps ahead planning. In order to reduce the computational costs the search space can be narrowed accepting a suboptimal solution by optimizing the view directions of the individual sensors independently of one another.

Object Pose. Also the poses of observed objects may be of relevance to the robot's mission, e.g. for manipulation, interaction, or collision avoidance tasks. The planning problem of appropriate view directions can be solved analogously to the approach in the previous section. The predicted reduction of object uncertainties is computed for each object and each possible configuration of camera view directions and that configuration is chosen which reduces the object uncertainties maximally. Dependent on the mission context a task-relevant formulation can be chosen. An example is the approach to a collision avoidance task proposed in [45] evaluating the uncertainties of obstacle positions orthogonal to the robot's path.

Visibility: If a camera is directed towards an object to be observed, knowledge of the object position is necessary. If the assumed object position is erroneous, the object may be located outside the visible field of the camera. This is of particular relevance to sensors with small fields of view, e.g. telephoto cameras. Two conditions can be derived for a camera control strategy: 1. The position of the object is too uncertain such that a camera shift towards the object might be useless as the object may most probably not be seen. 2. The position uncertainty is just small enough such that the object will most probably be located within the visible field, but would increase in the next time step so that it would exceed the limits of the visible field; thus, if the camera is not directed towards the object in this time step the object will probably be lost in the next. The second criterion has to be traded versus other view direction determining conditions.

Discrete Events: In dynamically changing real-world environments events also occur which are discrete in nature and which might be relevant to the primary mission or to secondary tasks to be accomplished. Examples are signs and signals, moving, appearing, or changing objects, human behavior and communication, etc. A possible approach to consider such events in a planning architecture is the representation by binary variables weighted according to the individual importance.

These criteria for view direction selection can be considered simultaneously. This leads to a typical decision problem which can be solved by applying common methods from decision-theory. A problem in this context is an appropriate dynamical selection of weighting factors for the different competing taks in order to solve the decision problem optimally. A common approach is the application of learning schemes in order to obtain (sub)optimal weights.

In either case additional constraints or a more generic formulation of the decision problem are necessary. Multi-focal view direction planning with multi-task competition is discussed in Section 5.4.

5.2 Localization of a Humanoid Robot

The approach to the considered robot localization problem is based on an extended Kalman-filter whereas the state vector is composed of the robot pose and object positions in the environment. The robot model is formulated according to [45] capturing the propagation of the foot step placements of a humanoid robot. The measurement equation is based on a multi-focal sensor model described in Section 3.1.

5.2.1 Planning of Perceptual Resources and SLAM

Active vision view direction planning with respect to locomotion tasks is considered mainly in works in the field of human and humanoid walking as well as simultaneous localization and mapping (SLAM).

SLAM

In order to navigate a robot through an environment estimations of the robot pose and of object positions in the environment are necessary. As the measurement uncertainties of both are correlated only a simultaneous estimation is possible which is the main focus of SLAM [39, 124]. This simultaneous estimation is commonly done by using probabilistic filters. Kalman- and Particle-Filters and various derivations are the most common tools. Among the estimation of states, i.e. robot pose and object positions, an explicit representation of the uncertainties and ambiguities associated with these states is necessary.

Bayesian analysis is specifying a prior distribution that is sophisticated enough to incorporate a priori knowledge but simple enough to make the problem algebraically tractable. The filtering problem consists of a recursive estimation based on a set of noisy observations. At least the first two moments of the state vector are considered. A dynamic nonlinear state space model capturing the locomotion of the robot is used. Most realizations in SLAM are discrete time. A discrete time state space model consists of a stochastic propagation (prediction or dynamic) equation that links the observation given the current state. If the dynamic and observation equations are linear and the associated noises are Gaussian, the optimal recursive filtering solution is the Kalman filter [64]. The most widely used filter for nonlinear systems with Gaussian additive noise is the well known extended Kalman filter (EKF). The EKF approximates the nonlinearities of the system and observation models by a Taylor series expansion about the current estimate, which is usually truncated after the first term. The estimation accuracy of the EKF depends on how well the system is approximated by the linearization. If the nonlinearities are significant or the noise is non-Gaussian, the EKF gives poor performance. Improvements to EKF-based SLAM have been made proposing a variety of extensions as, e.g., the iterated Kalman filter, multiple hypothesis Kalman Filter, and others. Several other approaches to recursive nonlinear filtering have appeared in the literature. These include grid-based methods, Monte-Carlo methods, Gauss quadrature methods, unscented filter and particle filter methods. Most of these filtering methods have their basis in computationally intensive numerical integration

techniques that have become tractable due to the increase in computational power over the last decade.

Sensor Planning

Allocation of perceptual resources commonly referred to as sensor planning requires active sensors or sensors with an adjustable focus of attention. In the context of SLAM, vision sensors are the most common active sensors considered. The objective is a selection of the sensor's main direction of perception in the next time step. Therefore, measures of entropy or information content are commonly defined, e.g. [29, 30, 67, 123, 130]. Common measures consider the robot pose and object positions as mentioned in Section 5.1. These are usually formulated evaluating the covariance matrix (in case of a Kalman-filter), other probabilistic moments, or probability distributions of the probabilistic filter. The predicted resulting (co)variances in the next time step corresponding to the states of interest of the filter, e.g. robot and object positions, are then compared for each of the possible sensor alignments. If no other concurrent tasks exist these sensor alignments win for which the (co)variances are reduced optimally according to the formulated information measures, e.g. maximal reduction of the robot pose variance. This has also been accomplished in state-of-the-art works considering a certain planning time horizon, i.e. optimizing the sensor alignments for several steps ahead based on the current available information, e.g. [45]. However, most approaches are based on greedy methods, i.e. only considering the gain of information in the next time step. Commonly, sensor planning and path planning are performed separately, i.e. sensor planning considers an already planned path, e.g. [29, 30, 130].

5.2.2 Robot Model, Perception Model, and Environment Model

In the following the assumed robot, perception, and environment models for the localization of a humanoid robot are defined. These models form the basis for the multi-focal view direction planning strategy in Section 5.3. The humanoid robot model represents the propagation of footstep placements in the environment if dead-reckoning errors are present. The perception model is composed of two or more sensors representing cameras with various focal-lengths. It projects Cartesian points to sensor space adding Gaussian sensor noise. The environment model is basically a map of Cartesian points which are considered landmarks. Each time a new landmark is detected by the vision system the map is extended. The landmarks in the map are computed with respect to the robot. These models are based on [45] extending the perception model to multi-focal vision.

Simplified Humanoid Robot Model

The mobile robot is considered a humanoid walking robot which is capable of placing its footsteps with respect to a robot centered reference frame S_F. The robot model gives the resulting distorted footstep placements in a world frame S_0 in response to a commanded step with position ${}_F x_s$, ${}_F y_s$ and orientation ${}_F \theta_{s,k}$ in open loop control. The commanded footsteps are considered results of a higher-level path planner which is not within the scope of this work.

The foot poses are distorted by dead-reckoning errors $w = [\Delta x_{s,k}\ \Delta y_{s,k}\ \Delta\theta_{s,k}]^T$ representing the difference between a commanded step and the corresponding measured one. The reference frame S_F is placed at that foot currently in contact with the ground during the single-support phase. Double support phases with both feet on the ground are ne-

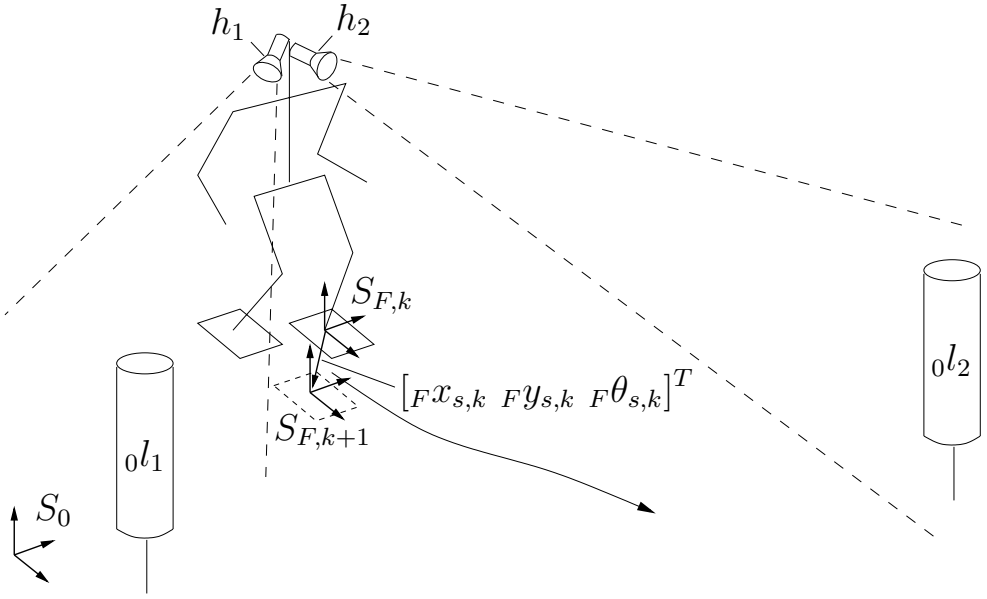

Figure 5.3: Humanoid robot navigation scenario with multi-focal vision.

glected. As the single-support foot changes from step to step the reference frame S_F must also change. This is considered by a binary variable $\gamma_k \in \{0; 1\}$. A simplified formulation is chosen considering only the time steps at the beginning, respectively, the end of a footstep of the humanoid robot.

Following these assumptions the discrete time robot model can be expressed by the nonlinear model

$$ {}_0x_{k+1} = {}_0x_k(1 - \gamma_{k+1}) + f_s({}_0x_k, u_{k+1}, w_k)\gamma_{k+1}, $$

giving the next foot pose ${}_0x$ at a time and footstep $k+1$ with control vector $u_{k+1} = [{}_F x_{s,k}\ {}_F y_{s,k}\ {}_F\theta_{s,k}\ \gamma_{k+1}]^T$ containing the commanded footstep in frame S_F, the dead-reckoning error w, and

$$ f_s({}_0x_k, u_{k+1}, w_k) = \begin{bmatrix} \begin{bmatrix} {}_0x_{0,k} \\ {}_0y_{0,k} \end{bmatrix} + Rot_{2x2}({}_F z, {}_0\theta_k) \begin{bmatrix} {}_F x_{s,k} + \Delta y_{s,k} \\ {}_F y_{s,k} + \Delta y_{s,k} \end{bmatrix} \\ {}_0\theta_k + {}_F\theta_k + \Delta\theta_{s,k} \end{bmatrix}, $$

with $[{}_0x_{0,k}\ {}_0y_{0,k}\ {}_0\theta_{0,k}]^T = {}_0x_k$ the starting pose for the next foot step defined by the foot placement from the previous time step, ${}_F z$ the vector perpendicular to the ground plane, and the commanded step $[{}_F x_{s,k}\ {}_F y_{s,k}\ {}_F\theta_{s,k}]^T$.

Multi-Focal Perception Model

A general perception model is considered combining two or more vision devices with independent pose control. These devices may differ in their performance properties focal-length and field of view. Utilizing the sensor model described in Section 3.1 and depicted in the schematic Figure 3.1 the general nonlinear multi-focal perception model is written

$$ \xi_k = h_\xi({}_0x_k, {}_0l, {}_F P_k, v_{\xi,k}) = \begin{bmatrix} h_{\xi,1}({}_0x_k, {}_0l, {}_F P_{1,k}) \\ h_{\xi,2}({}_0x_k, {}_0l, {}_F P_{2,k}) \\ \vdots \\ h_{\xi,n}({}_0x_k, {}_0l, {}_F P_{n,k}) \end{bmatrix} + v_{\xi,k}, \tag{5.1} $$

with feature point vector in image space $\xi = [\xi_1^T \dots \xi_n^T]^T$, $\xi_i = [\xi_{u,i}\ \xi_{v,i}]^T$, robot pose ${}_0x_k$, a vector containing the positions of all point landmarks in the environment ${}_0l$, and the perspective projection matrices ${}_FP_i$ containing the transformation matrices ${}^FT_{ci}$ of frame S_{ci} of camera i with respect to the robot foot frame S_F. Individual rotational or translational components of ${}^FT_{ci}$ may be controllable. Vector $v_{\xi,k}$ represents sensor noise in image space. Sensors $h_{\xi,i}$ may be single-cameras, stereo-cameras, or multi-camera configurations. In the remainder of this chapter only stereo-cameras are considered without loss of generality.

In order to build a map and to localize the robot within the environment, the 3D positions of the landmarks have to be reconstructed. This can be done utilizing the procedure described in Section 3.1.1 finding a weighted least squares solution

$$ {}_0\hat{l}^* = \left(\hat{F}^T E^{-1} \hat{F}^T\right)^{-1} \hat{F}^T E^{-1} \hat{b}, \tag{5.2} $$

with estimated landmark positions ${}_0\hat{l}^*$, E the sensor noise covariance matrix expressed in Cartesian space (cf. Section 3.1.1), and $\hat{F}$, $\hat{b}$ as defined in Section 3.1.1 perturbed with sensor noise.

Environment Model

The mobile robot moves along a planned path through an environment with a finite number of observable point objects in the following referred to as landmarks. The position of these landmarks is not known a priori. By moving through the environment and evaluating visual data the robot, thus, has to explore the environment and successively build a map containing estimated landmark positions.

Landmark positions can be estimated solving the 3D reconstruction problem utilizing (5.2) whenever individual landmarks are visible in at least one of the vision sensors yielding a nonlinear measurement equation

$$ z_k = f_m(\xi_k, {}_FP_k, v_{\xi,k}), $$

where the measurements z_k are a nonlinear funtion of the projections of the landmark positions in image space ξ_k, the perspective projection matrices ${}_FP_i$, and the sensor noise $v_{\xi,k}$. These measurements are relative to the current robot reference frame S_F. Due to sensor noise the reconstructed landmarks positions in the map are erroneous.

Those landmark positions which are not visible at a time step have to be computed from previous measurements, i.e. the map has to be translated and rotated according to the robot locomotion. However, due to dead-reckoning errors the predicted landmark positions are also erroneous. The transformation of landmark positions accounting for the robot locomotion and the changing robot frame S_F according to the foot in contact with the ground as defined above can be written

$$ {}_Fx_{l,k+1} = {}_Fx_{l,k}(1 - \gamma_{k+1}) + f_e({}_Fx_{l,k}, u_{k+1}, w_k)\gamma_{k+1}, $$

with

$$f_e({}_F x_{l,k}, u_{k+1}, w_k) = \begin{bmatrix} Rot_{2x2}({}_F z, -({}_F\theta_{s,k} + \Delta\theta_{s,k})) \left[\begin{bmatrix} {}_F x_{l,1,k} \\ {}_F y_{l,1,k} \\ {}_F z_{l,1,k} \end{bmatrix} - \begin{bmatrix} {}_F x_{s,k} + \Delta x_{s,k} \\ {}_F y_{s,k} + \Delta y_{s,k} \\ 0 \end{bmatrix} \right] \\ \vdots \\ Rot_{2x2}({}_F z, -({}_F\theta_{s,k} + \Delta\theta_{s,k})) \left[\begin{bmatrix} {}_F x_{l,n,k} \\ {}_F y_{l,n,k} \\ {}_F z_{l,n,k} \end{bmatrix} - \begin{bmatrix} {}_F x_{s,k} + \Delta x_{s,k} \\ {}_F y_{s,k} + \Delta y_{s,k} \\ 0 \end{bmatrix} \right] \end{bmatrix},$$

with the vector containing the landmark positions, i.e the map, ${}_F x_l = [{}_F x_{l,1}\ {}_F y_{l,1}\ {}_F z_{l,1}\ \cdots\ {}_F x_{l,n}\ {}_F y_{l,n}\ {}_F z_{l,n}]^T$, control vector $u_{k+1} = [{}_F x_{s,k}\ {}_F y_{s,k}\ {}_F\theta_{s,k}\ \gamma_{k+1}]^T$ containing the commanded step as defined above, and the dead-reckoning errors $w = [\Delta x_{s,k}\ \Delta y_{s,k}\ \Delta\theta_{s,k}]^T$.

5.2.3 Robot Localization

The robot's mission is to follow a path as well as possible. Thus, the robot has to estimate its pose in the environment continually. However, the robot dead-reckoning model and the measurements are erroneous. Moreover, if no absolute measurements of the map are taken and the robot is only localized evaluating odometry data the dead-reckoning errors accumulate non-recoverably resulting in a drift of the estimated robot pose. Therefore, a fusion of the two localization principles – relative odometry and absolute vision-based – is necessary. Even better results can be obtained if information about the error signals is taken into account. This can be achieved utilizing an extended Kalman-filter as an optimal recursive state estimator of nonlinear systems. In the following the basic principle applied to the assumed scenario is described in brief. This description of basic steps of the general localization and mapping procedure applied to the problem formulation is only intended to serve as a brief summary for the definition of the general framework of this chapter. For a detailed explanation of the Kalman-filter and SLAM foundations the kind reader may refer to common literature.

As common in localization and mapping, the system state is composed of the robot pose ${}_0 x$ in the world frame and the vector of landmarks positions ${}_F x_l$ with respect to the robot frame

$$x_k = \begin{bmatrix} {}_0 x_k^T & {}_F x_{l,k}^T \end{bmatrix}^T. \tag{5.3}$$

The nonlinear state space model is composed of the dead-reckoning model of the robot locomotion (5.2.2), the dead-reckoning model of the map (5.2.2), and the vision-based measurements (5.2.2)

$$\begin{aligned} x_{k+1} &= f({}_0 x_k, u_{k+1}, w_k) = x_k(1 - \gamma_{k+1}) + \begin{bmatrix} f_s({}_0 x_k, u_{k+1}, w_k) \\ f_e({}_0 x_k, u_{k+1}, w_k) \end{bmatrix} \gamma_{k+1}, \\ z_{k+1} &= f_m(\xi_{k+1}, {}_F P_{k+1}, v_{\xi,k+1}). \end{aligned}$$

The random variables w and v represent process and measurement noise, respectively. They are assumed white, zero mean with diagonal non-zero covariance matrices Q_k and R_k, respectively. Linearization of the model yields

$$x_{k+1}^{lin} = A_k x_k + B_{k+1} u_{k+1} + W_k w_k,$$

$$z_{k+1}^{lin} = H_{k+1}x_{k+1} + V_{k+1}v_{k+1},$$

with

$$A_k = \frac{\partial f}{\partial x_k},\ B_{k+1} = \frac{\partial f}{\partial u_{k+1}},\ W_k = \frac{\partial f}{\partial w_k},\ H_{k+1} = \frac{\partial h}{\partial x_{k+1}},\ V_{k+1} = \frac{\partial h}{\partial v_{k+1}}.$$

The evaluation of the linearization can be found in [45] concluding validity for a sequence of approximately ten steps. As long as no vision-based measurements of the robot pose and landmark positions are available the robot pose and landmark positions have to be predicted evaluating the dynamics

$$\begin{aligned} \hat{x}_{k+1|k}^{lin} &= f(\hat{x}_{k|k}, u_{k+1}, 0), \\ \hat{z}_{k+1|k}^{lin} &= f_m(\hat{\xi}_{k+1|k}, {}_F P_{k+1}, 0), \end{aligned}$$

the a priori state and measurement estimates with noises assumed zero and estimated values $\hat{(.)}$. For the a priori covariance estimate the changing robot frame has to be taken into account. Thus, the covariance estimate depends on γ

$$\begin{aligned} C_{k+1|k} &= C_{k|k}, && \text{if } \gamma_{k+1} = 0, \\ C_{k+1|k} &= A_k C_{k|k} A_k^T + W_k Q_k W_k^T, && \text{if } \gamma_{k+1} = 1, \end{aligned}$$

where

$$Q_k = \begin{bmatrix} \sigma_{\Delta x_s}^2 & 0 & 0 \\ 0 & \sigma_{\Delta y_s}^2 & 0 \\ 0 & 0 & \sigma_{\Delta \theta_s}^2 \end{bmatrix},$$

represents the dead-reckoning noise covariance matrix for a single step.

When measurements are available the estimated values are corrected. The update equations are expressed

$$\begin{aligned} \hat{x}_{k+1|k+1} &= \hat{x}_{k+1|k} + K_{k+1}(z_{k+1} - \hat{z}_{k+1|k}), \\ C_{k+1|k+1} &= C_{k+1|k} - K_{k+1} S_{k+1} K_{k+1}^T, \end{aligned}$$

with the Kalman gain K defined as

$$K_{k+1} = C_{k+1|k} H_{k+1}^T S_{k+1}^{-1},$$

with

$$S_{k+1} = H_{k+1} C_{k+1|k} H_{k+1}^T + V_{k+1} R_{k+1} V_{k+1}^T,$$

and the measurement noise covariance matrix

$$R_{k+1} = \text{diag}(\sigma_{\xi,u,1}^2,\ \sigma_{\xi,v,1}^2,\ \ldots\ , \sigma_{\xi,u,n}^2,\ \sigma_{\xi,v,n}^2),$$

for a multi-focal vision system comprising n sensors. Within the scope of this work two stereo-cameras are considered resulting in $R \in \mathbb{R}^{8\times 8}$.

This method predicts the robot footstep poses and landmark positions based on the robot and environment models defined in the previous section when no measurements are available. When measurements are taken utilizing the multi-focal sensor model from the previous section the footstep poses and landmark positions are corrected. In the following section this method is used in order to investigate the impact of different focal-lengths, fields of view, and view direction changes on the localization accuracy of the humanoid robot.

5.3 Multi-Focal View Direction Planning

In Section 5.1 the shortcomings of conventional view direction planning with mono-focal vision systems have been discussed in exemplary settings. An exploration scenario of a mobile robot has served to show challenges for improvement of mission relevant parameters as, e.g., localization accuracy and probability of collision. The existence of problems which cannot be solved based on mono-focal vision has also been pointed out, e.g., achieving good localization accuracy *and* scene overview in environments with sparsely distributed landmarks. The concept of multi-focal view direction planning has been introduced in order to overcome the shortcomings of the state-of-the-art.

This section is concerned with the definition of mathematical conditions for multi-focal view direction planning, the formulation of planning strategies, and the evaluation based on a humanoid robot locomotion scenario which has been defined in the previous section.

5.3.1 Criteria and Information Measures for Camera Coordination

As summarized in Section 5.1 several criteria have to be considered for a goal-oriented task-specific coordination of the camera view directions. Within the scope of multi-focal view direction planning in the defined robot locomotion scenario with point landmarks a sufficient accuracy of the estimated robot pose is the main objective. Without sufficient localization accuracy the mission - following a global path - cannot be completed successfully. A second criterion is given by potential mission-relevant activities in the environment which are considered being discrete. Examples are signals or events of interest detected by wide-angle vision sensors which have to be observed by a higher-resolution sensor. Another criterion which is of particular interest for telephoto cameras with extremely small aperture angles is the visibility of objects to be observed. If the uncertainty of the assumed object position the camera is directed at is too high then the object might eventually be located outside the field of view. The second and third criteria are important aspects in dynamical environment, e.g., if moving objects are present changing positions between two observations. In the following conditions are defined capturing these criteria.

Primary Mission - Predicted Uncertainty of the Robot Pose

If the mission of the robot is to follow a path as well as possible a measure assessing the robot pose error with respect to a world coordinate frame is necessary. In terms of the assumed humanoid robot model a possible approach is to evaluate the robot pose covariance matrix, e.g., computing the volume of the covariance ellipsoid. In [45] the mean of the main axes of the covariance ellipsoid given by the square roots of the eigenvalues of the robot pose covariance matrix is considered. Based on this the task-related information measure *incertitude* for a current foot step s in world frame S_0 is defined as

$$\nu_0^s = \frac{1}{2} \sum_{j=1}^{2} \sqrt{e_j^s}, \tag{5.4}$$

with e_1^s and e_2^s the eigenvalues of the 2×2 robot pose covariance matrix for x- and y-direction of world frame S_0. In [45] it is argued that the orientation error is low whenever the position error is low. Thus, the eigenvalue corresponding to the orientation uncertainty is neglected.

This measure is nondirectional in nature. The differences of the eigenvalues representing different uncertainties in various Cartesian directions are not considered. Thus, large pose errors in particular directions can be unnoticed. A possible approach in order to overcome this shortcoming is an evaluation of the condition number of the pose covariance matrix.

Potential Object or Event of Interest

In real-world environments the mission of the robot may require reactions to events or objects of interest in order to be successful. Examples are signs and signals to be observed, moving objects to be tracked, or a human initiating a communication process.

As these aspects are of discrete nature their presence or activity can be expressed as a binary variable

$$\varepsilon_{0,j} = \begin{cases} 0, & \text{if interesting object/event present,} \\ 1, & \text{otherwise,} \end{cases} \tag{5.5}$$

representing an interest operator signaling a potentially interesting object or event to be focused.

Predicted Visibility of an Object of Interest

If a camera has to be focused at an object of interest an estimation of the object position is necessary. This estimation may be erroneous due to sensor and dead-reckoning errors. If these errors exceed the field of view of the camera the object may not be observable though assumed focused.

Figure 5.4 schematically shows the top view of a point object observed by a vision sensor with a limited field of view defined by its aperture angle α_{pan}. The aperture angle defines a point set $T_{\alpha,pan}$ covering the triangular field of view. The assumed object position is disturbed by gaussian noise in two dimensions. This noise is visualized by a p-confidence ellipse of the corresponding covariance matrix of the object position estimate defined by a point set C_p covering the ellipse area.

In the depicted case the confidence ellipse touches the edges of the field of view triangle in at least one point. Thus, the probability of the object to be located within the field of view is at least p.

Considering this basic example two causes of action for camera coordination can be considered:

(i) **Re-Observe:** A given probability threshold is exceeded. As a consequence the camera is refocused at the object in order to lower the position uncertainty to an acceptable level.

(ii) **Discard:** A given probability threshold is exceeded. As a consequence the object position is considered too uncertain to be captured by the camera and the object is discarded from current view direction planning.

Given these options there are several ways to define conditions for view direction planning:

Binary Constraints, Level of Confidence. An acceptable confidence level p for the position uncertainty is defined, e.g., for a 90% confidence ellipsoid. The current configuration of view direction, aperture angles, and object position covariance matrix is evaluated. If the confidence ellipsoid is fully located within the field of view then the object is considered visible at an acceptable confidence level.

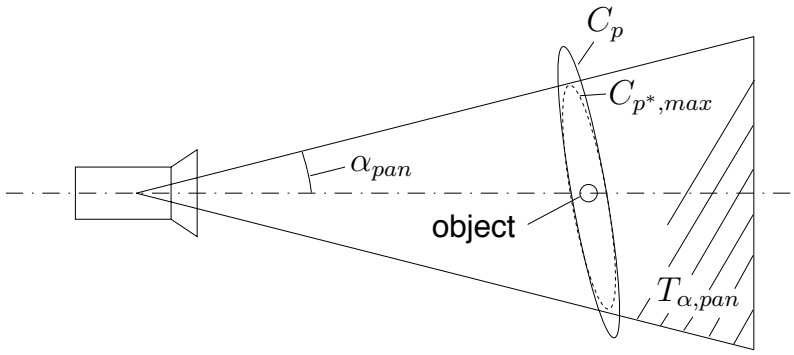

Figure 5.4: Visibility of an object observed by a vision device with aperture angle α_{pan}; p-confidence ellipse C_p of object position covariance matrix; field of view $T_{\alpha,pan}$; maximum confidence ellipse $C_{p^*,max}$ within $T_{\alpha,pan}$.

Therefore, the intersection of the point sets of field of view $T_{\alpha,pan,tilt}$ and confidence ellipsoid C_p has to satisfy

$$C_p \cap T_{\alpha,pan,tilt} = C_p, \tag{5.6}$$

where

$$C_p = \left\{ {}_r x \in \mathbb{R}^3 \middle| \, ({}_r x - {}_r\hat{x}_i)^T C_i^{-1} ({}_r x - {}_r\hat{x}_i) \leq \chi_p^2, \; {}_r\hat{x}_i \in \mathbb{R}^3 \right\}, \tag{5.7}$$

with covariance matrix C_i of the object position estimate, confidence level p, ${}_r(.)$ denoting the vision sensor reference frame, estimated object position ${}_r\hat{x}_i$ with respect to the vision sensor, and

$$T_{\alpha,pan,tilt} = \left\{ {}_r x \in \mathbb{R}^3 \middle| \, X \leq Z \tan \alpha_{pan}, \, Y \leq Z \tan \alpha_{tilt} \right\}, \tag{5.8}$$

where α_{pan}, α_{tilt} are the aperture angles of the vision system in x- and y-direction with respect to the vision system reference frame and ${}_r x = (X, Y, Z)^T$.

A binary variable δ_0 can be defined for each object i of the map

$$\delta_{0,i}(C_i, {}_r\hat{x}_i, \alpha_{pan}, \alpha_{tilt}) = \begin{cases} 0, & \text{if } \; C_p \cap T_{\alpha,pan,tilt} \neq C_p, \\ 1, & \text{if } \; C_p \cap T_{\alpha,pan,tilt} = C_p, \end{cases} \tag{5.9}$$

which can be integrated in a view direction planning strategy. State $\delta_{0,i} = 0$ then represents a signal to *re-observe*, respectively, *discard* object i, depending on the chosen action.

This constraint is an approximation giving the worst case probability of the observed object being located in the field of view.

Another possibility is to compute the current maximum confidence level p^* determining the maximum confidence ellipsoid $C_{p^*,max}$, $p^* \in \{0, 100\}$, with tangential planes defined by the aperture angles (see Figure 5.4). The current confidence level is then compared with a predefined threshold $p^* < p$. As long as this constraint is not violated the object is considered visible at an acceptable confidence level.

Thereby, an approximation of the worst case probability of the object being visible is also given.

Continuous Constraints, Probability of Visibility. The previous constraints are of binary nature, i.e. a decision is made when a condition is satisfied. If two or more objects of interest exist and both objects have to be re-observed in order to lower their position uncertainty it is not clear which of both to observe first. Therefore, a continuous measure for the predicted visibility of an object to be observed is necessary for priorization of possible view directions. There are a number of possible definitions.

A straight forward formulation of a continuous constraint is based on the previous constraint evaluating the current maximum level of confidence p^*. However, this confidence level is only an approximation for the worst case probability which could be overpessimistic.

Another possibility is the computation of the true probability of visibility computing the volume integral of the probability distribution of the object position uncertainty over the Cartesian space from infinity to the limits defined by the aperture angles of the vision system. However, as unmodeled nonlinearities and linearization errors of robot and sensor models exist, the covariance matrix and field of view are both coarse approximations. The evaluation of the approximation error is nontrivial and, thus, an assessment of this measure is difficult.

5.3.2 Planning Strategies for Robot Localization

In the previous section constraints and information measures have been defined which are evaluated in the following in order to plan view directions for the individual cameras of a multi-focal vision system. The presumed objective for view direction planning is to gather the largest possible amount of information with respect to the mission to be accomplished.

The assumed mission of the robot is to follow a path as well as possible. As a consequence the estimation error of the robot pose within the environment during its motion has to be minimal in order to complete the mission optimally. In order to minimize this error appropriate view directions of the individual cameras of the multi-focal vision system have to be chosen. Following this, an optimal configuration of view directions for the current time step satisfies the condition of minimizing the robot pose estimation error.

Primary Mission

Based on these considerations a novel multi-focal approach to gaze control for mobile robots is proposed. This approach constitutes a generalization of the method of [45] extending it to multi-camera and multi-focal vision. The basic principle is an information maximization over a set of possible view directions of independent vision devices. As defined in Section 5.2 two or more sensors can be grouped to one independent device with a main view direction and certain performance characteristics which have been discussed in Chapter 3, e.g. forming mono- or multi-focal stereo-pairs.

In terms of robot localization the view direction planning strategy can be expressed as a maximum reduction of the predicted robot position uncertainty $\hat{\nu}_0$ at the next time step $s+1$ over all possible predicted view directions $\hat{\Omega} = (\hat{\Omega}_1\, \hat{\Omega}_2\, \ldots\, \hat{\Omega}_n)^T$ of the n individual vision devices writing

$$\hat{\Omega}^{*,s+1|s} = \arg\min_{\hat{\Omega}} \hat{\nu}_0^{s+1|s+1}, \tag{5.10}$$

where $\hat{\Omega}_j = (pan_j\, tilt_j)^T$ are the pan- and tilt-angles of vision device j.

This formulation is called greedy as only the next time step is considered giving a suboptimal solution with respect to the defined mission. An optimal solution can only be obtained evaluating the complete time horizon from start to goal position of the robot which is impractical. However, [45] shows that the difference between optimal and suboptimal solution is negligible for a time horizon of ten steps in the considered scenario.

As pointed out in Section 5.1 this optimization may be computationally expensive depending on the number of individual vision devices and the resolution of commandable pan- and tilt-angles. The search space can be significantly reduced optimizing the view directions $\hat{\Omega}_j$ of each vision device independently accepting a suboptimal solution.

Secondary Tasks

Additional tasks which are not directly relevant for achieving the global mission have also to be considered in order to provide a robust, flexible, and safe behavior of the robot in dynamically changing real-world environments. Two other constraints have been proposed which are particularly relevant in multi-focal vision: the occurrence of potentially interesting events and the predicted visibility of objects.

As proposed in the previous section discrete events can be represented by a binary variable. For a particular vision device of the muti-focal system this condition can be considered combining primary mission and secondary task writing

$$\hat{\Omega}_j^{*,s+1|s} = \arg\min_{\hat{\Omega}_j, n_{ev}} \left((1-r)\varepsilon_{n_{ev}} + r(1-\varepsilon_{n_{ev}})\hat{\nu}_0^{s+1|s+1} \right), \tag{5.11}$$

where $\hat{\Omega}_j^*$ is the view direction of a particular vision device j, $\hat{\Omega}_j$ is the set of possible view directions for j, n_{ev} the set of events of interest, $\varepsilon_{n_{ev}}$ is a binary variable, cf. (5.5), with $\varepsilon_{n_{ev}} = 0$ signaling the presence of an interesting event, and $\hat{\nu}_0$ is a measure for the robot position uncertainty. Multiplier $r \in [0;1]$ is an interest operator which can be used in order to deliberately decide between both tasks, the primary mission and the secondary task. In the next section r is used in order to limit the position uncertainty of the robot by setting $r = 0$ if ν_0 exceeds a certain threshold and $r = 0.5$ otherwise.

So far the fields of view of the vision sensors have only been considered in the sensor model of the update equation of the Kalman-filter in Section 5.2. Landmarks which are not visible in the fields of view of the individual vision sensors are, thus, not used for pose estimation. As pointed out in the previous section due to estimation errors objects may be located outside the visible field though their predicted position falls within the field of view.

The multi-focal approach to view direction planning is, therefore, extended accounting for the predicted visibility of observable objects in the next step $s+1$. This is of particular importance in dynamically changing environments. The idea is an introduction of a re-observation term in order to direct a vision device towards an object when its predicted position uncertainty exceeds the limits of the field of view of a particular vision device j, e.g. a telephoto camera. Similarly to the previous formulation this strategy can be expressed by

$$\hat{\Omega}_j^{*,s+1|s} = \arg\min_{\hat{\Omega}_j, n_{obj}} \left((1-r)\delta_{n_{obj}} + r(1-\delta_{n_{obj}})\hat{\nu}_0^{s+1|s+1} \right), \tag{5.12}$$

where n_{obj} is the set of objects of interest, e.g. landmarks, and $\delta_{n_{obj}}$ is a binary variable with $\delta_{n_{obj}} = 0$ signaling that a certain confidence level of the covariance ellipsoid of the corresponding object exceeds the field of view of vision device j.

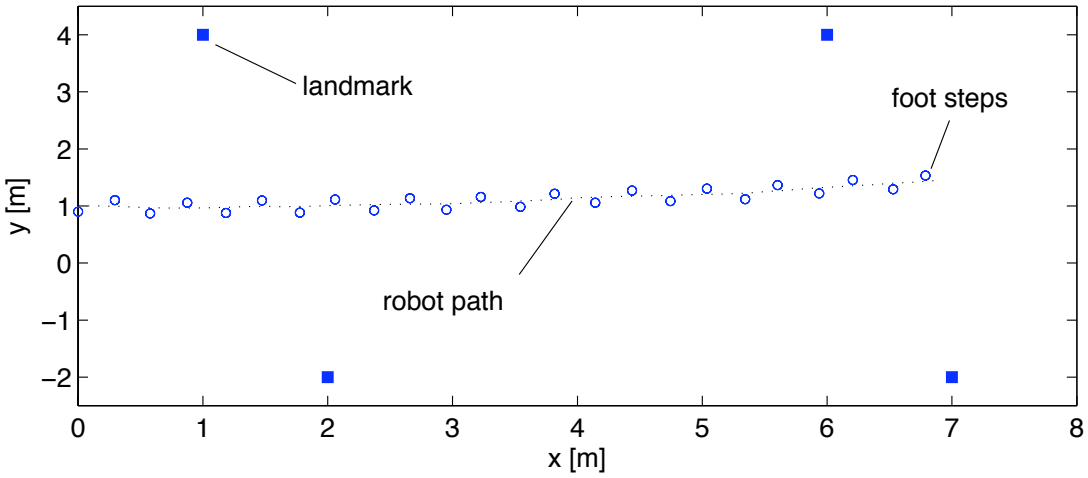

Figure 5.5: Top-view of the humanoid robot navigation scenario.

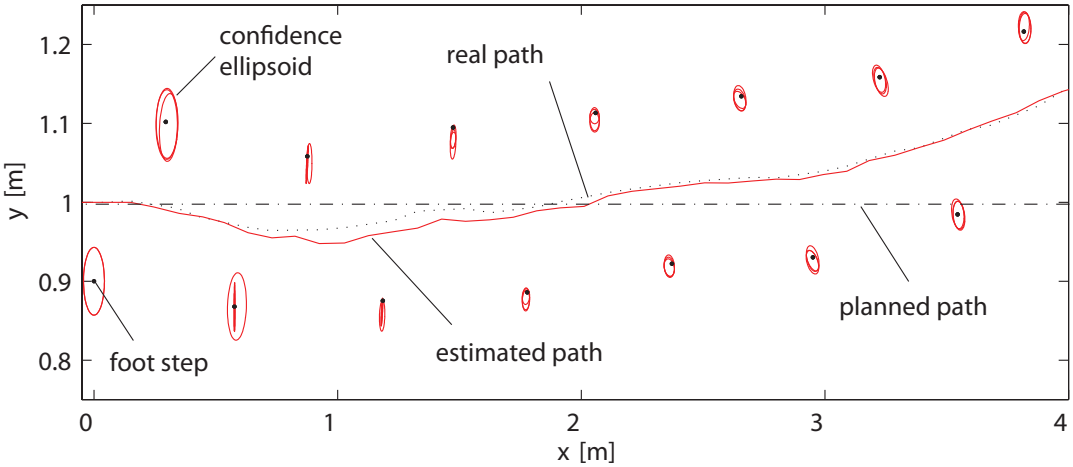

Figure 5.6: Real, estimated, and planned paths and foot steps.

5.3.3 Simulations and Evaluation

In order to demonstrate the benefits of multi-focal view direction planning the proposed approach is now evaluated in a structured humanoid robot navigation scenario. Several mono- and multi-focal vision system configurations are evaluated by comparison of the achieved navigation performances.

The basic scenario is shown in Figure 5.5. Four landmarks are distributed within a rectangular environment. The mission of the robot is to follow a straight path in x-direction. In order to complete the mission successfully the robot has to localize itself within the environment evaluating available visual information on the positions of the identified landmarks.

The robot pose is estimated dynamically utilizing the Kalman-filter approach described in Section 5.2.3 based on the robot, environment, and perception models defined in Section 5.2.2.

In order to maximize the information gain optimal view directions of the individual vision devices are selected dynamically based on the proposed approach in Section 5.3. The positions of the landmarks are not known a priori nor are the number of landmarks. Configurations of the vision system of the considered scenarios to be compared are:

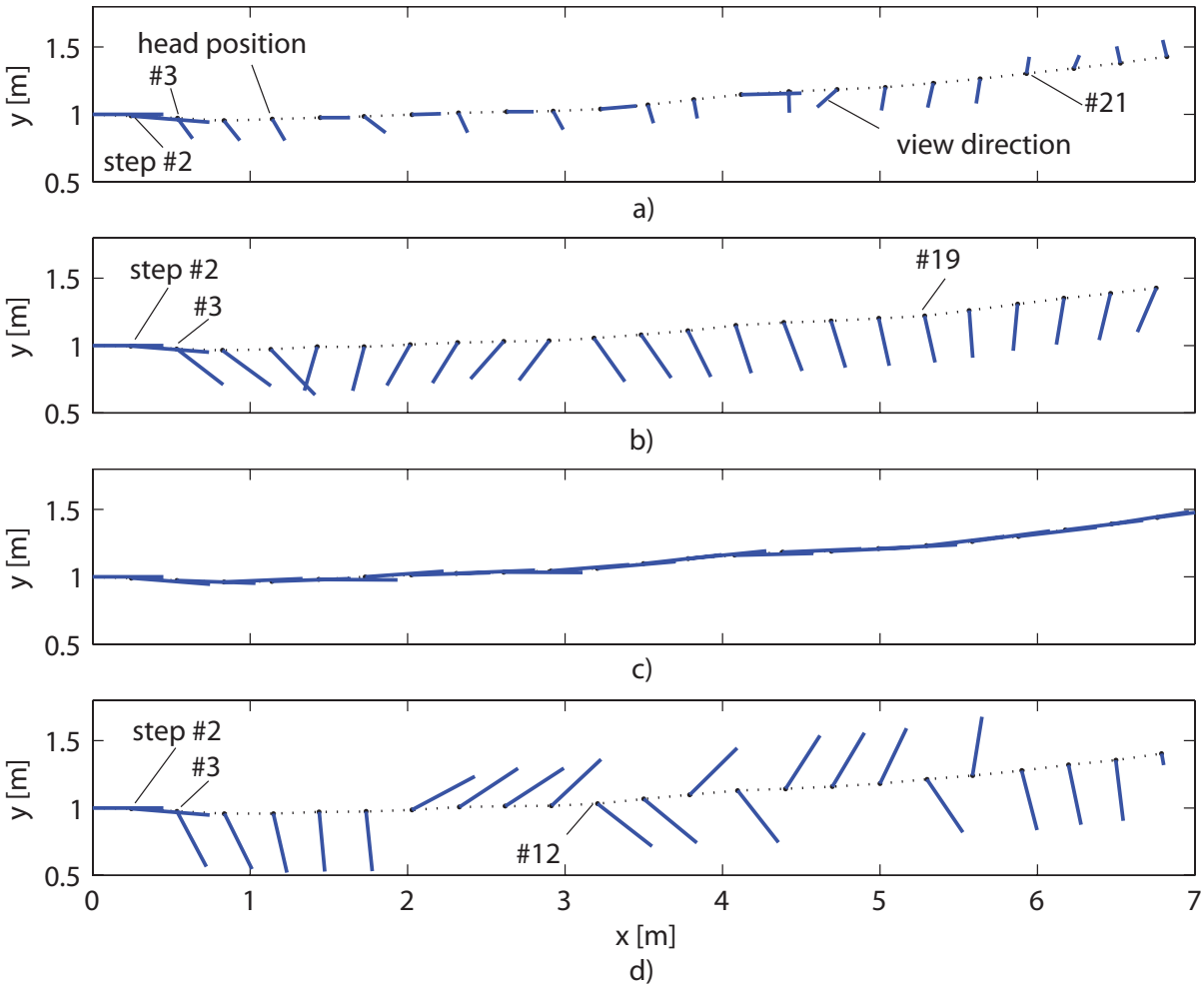

Figure 5.7: Comparison of resulting planned pan-angles for a) wide-angle (focal-length $\lambda = 2$mm, aperture-angles $\alpha_{pan,tilt} = [60\ 60]^{\circ T}$), b) conventional ($\lambda = 20$mm, $\alpha_{pan,tilt} = [30\ 30]^{\circ T}$), c) telephoto ($\lambda = 40$mm, $\alpha_{pan,tilt} = [10\ 10]^{\circ T}$), and d) foveated ($\lambda_w = 2$mm, $\lambda_t = 40$mm, $\alpha_{w,pan,tilt} = [60\ 60]^{\circ T}$, $\alpha_{t,pan,tilt} = [10\ 10]^{\circ T}$) vision devices.

a) **wide-angle:** a single wide-angle stereo-camera, focal-lengths $\lambda = 2$mm, aperture angles $\alpha_{pan,tilt} = [60\ 60]^{\circ T}$,

b) **conventional:** a single conventional stereo-camera, focal-lengths $\lambda = 20$mm, aperture angles $\alpha_{pan,tilt} = [30\ 30]^{\circ T}$,

c) **telephoto:** a single telephoto stereo-camera, focal-lengths $\lambda = 40$mm, aperture angles $\alpha_{pan,tilt} = [10\ 10]^{\circ T}$,

d) **foveated:** a foveated vision system (multi-focal system with fixed relative poses of the cameras) with one wide-angle stereo-camera, $\lambda_w = 2$mm, $\alpha_{w,pan,tilt} = [60\ 60]^{\circ T}$, and one telephoto stereo-camera, $\lambda_t = 40$mm, $\alpha_{t,pan,tilt} = [10\ 10]^{\circ T}$; additionally, this scenario is investigated with a larger field of view for the wide-angle camera of $\alpha_{w,pan,tilt} = [80\ 80]^{\circ T}$,

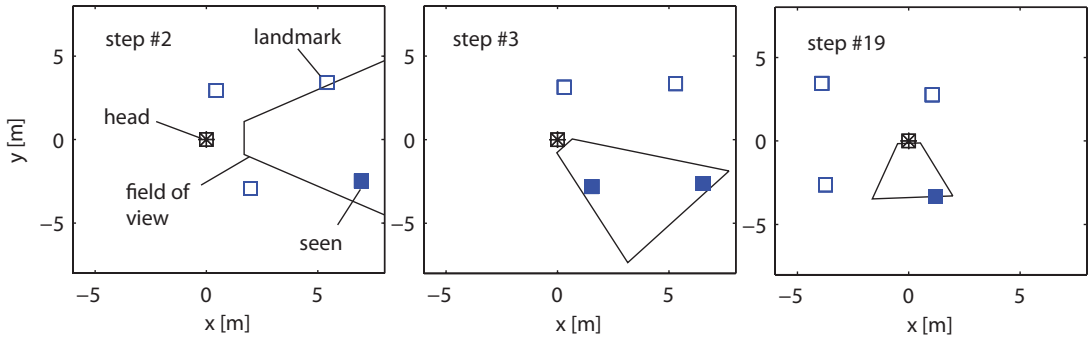

Figure 5.8: Projections of the field of view of a conventional vision device of footsteps #2, #3, and #19 of Figure 5.7b.

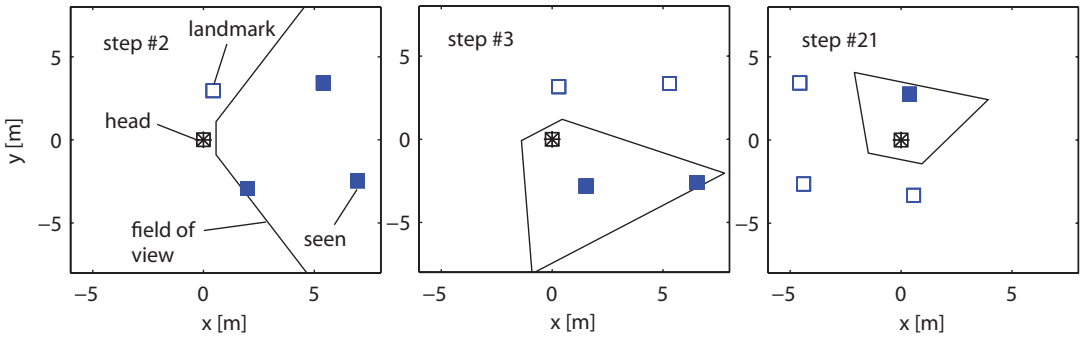

Figure 5.9: Projections of the field of view of a wide-angle vision device of footsteps #2, #3, and #21 of Figure 5.7a.

e) **multi-focal:** a multi-focal system with a wide-angle stereo-camera, $\lambda_w = 2\text{mm}$, $\alpha_{w,pan,tilt} = [60\ 60]^{\circ T}$, and a telephoto stereo-camera, $\lambda_t = 40\text{mm}$, $\alpha_{t,pan,tilt} = [10\ 10]^{\circ T}$.

All cameras are ideal, based on the pinhole camera model neglecting lens distortion and quantization effects.

The navigation performance is rated assessing the localization accuracy. Therefore, the covariance matrix of the robot position is evaluated computing the areas of the 90%-confidence ellipses of the footstep positions of the humanoid robot. Figure 5.6 contains a cut-out of Figure 5.5 showing the planned and real paths, the path estimated by the Kalman-filter, the foot step positions, and their covariance ellipses. It is noted that due to dead-reckoning errors the real path deviates increasingly from the planned path as locomotion control is open loop. However, the estimated path follows the real path well. Figure 5.7 and Figure 5.11 show the resulting view directions for each step of the robot and for each of the vision devices. The resulting fields of view on the ground plane within the environment are depicted in Figures 5.8, 5.9, 5.10, and 5.12. The propagations of the areas of the confidence ellipses are shown in Figure 5.13. Table 5.1 shows a comparison of the means of the confidence ellipse areas and the average number of landmarks visible for the vision systems for all scenarios. These results are discussed in the following.

Mono-Focal Localization Performance

Mono-focal vision systems, i.e. systems comprising only one sensor type, suffer from a trade-off of accuracy versus field of view. In robot localization not only measurement

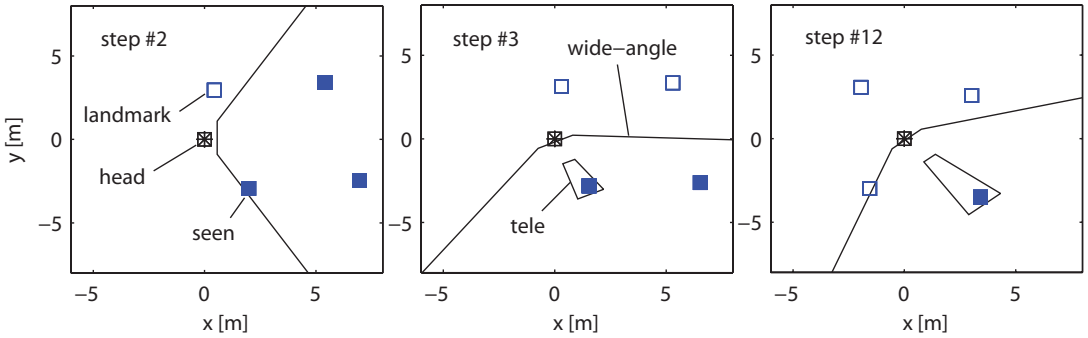

Figure 5.10: Projections of the field of view of a foveated vision device of footsteps #2, #3, and #12 of Figure 5.7d.

accuracy, but also the number of visible landmarks is an important factor in order to determine the current robot position. Depending on the distribution of landmarks and the current situation it may be better to observe more landmarks with lower accuracy in one situation and fewer landmarks with higher accuracy in another. Thus, mono-focal systems are always a compromise working well only for a very limited class of environmental conditions and situations, however, failing in others. This problem is reflected by the results shown in Figures 5.7 to 5.14.

The extreme case vision system comprising only one telephoto camera with very narrow field of view fails in most cases. If no a priori knowledge on landmark positions is available and no landmark is located within the field of view no information for the view direction planner is available. The only possibility is a continuous scan commanding the camera to rotate over the whole range trusting to accidental detection of the one or other landmark. However, this strategy is highly unefficient and unflexible. Figure 5.7c shows the case where no landmark is detected. In consequence view direction planning fails.

A conventional camera with medium accuracy and focal-length has a better chance to capture landmarks. Figure 5.7b shows the resulting optimal view directions for this scenario. It can be noted that in this case only one landmark is detected at the starting position (see Figure 5.8) and only accidentally a second one is seen after shifting the view direction (step #3). Thus, the planner only considers the lower landmarks within the environment. However, most of the time only one landmark is visible. In consequence moving further through the environment at some position the gaze has to be shifted towards another landmark (due to joint limits) resulting in a sudden increase of the robot position covariances (see Figure 5.13, $x = 3\text{m}$).

Interestingly, the wide-angle system shows a significantly better performance which can be noted by rapid decrease of the area of the foot step confidence ellipse area shown in Figure 5.13. Three landmarks are visible at the initial position (see Figure 5.9) and can, thus, be considered for optimal view direction planning. In most cases two landmarks are visible at a time.

Multi-Focal Localization Performance

The main contribution and benefit of multi-focal vision is the possibility to allocate sensor resources flexibly depending on environmental conditions and current situation. As can be seen in Figure 5.12 a variety of different configurations of view directions are selected by the planner in order to satisfy the current situational requirements optimally. At each step

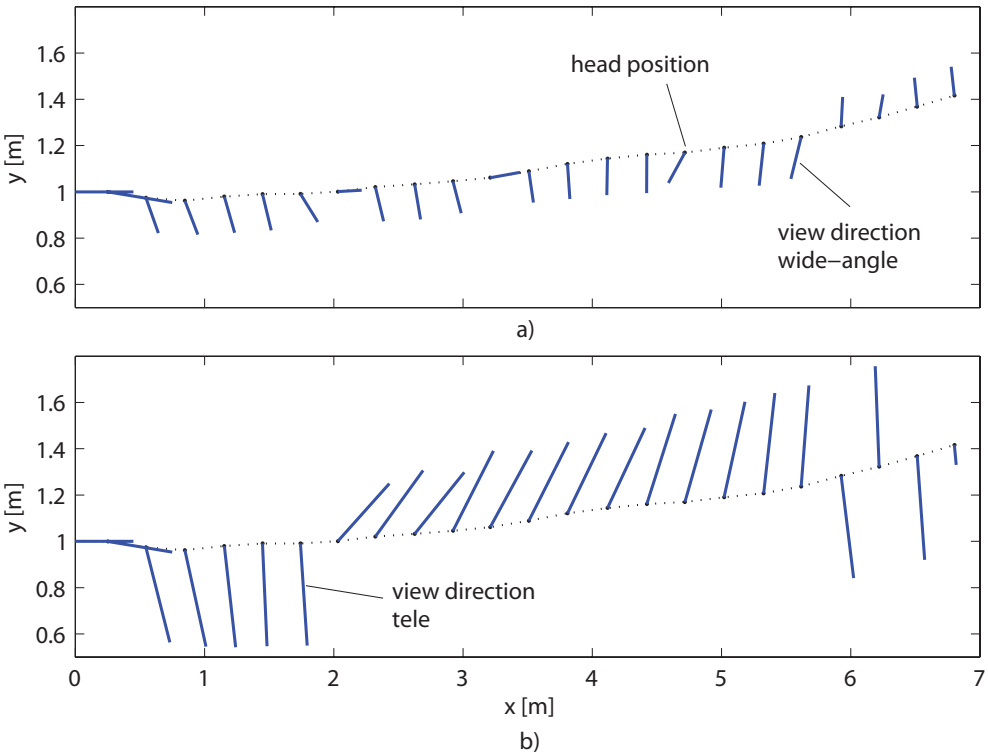

Figure 5.11: Planned pan-angles of a a) wide-angle (focal-length $\lambda_w = 2$mm, aperture-angles $\alpha_{w,pan,tilt} = [60\ 60]^{\circ T}$) and a b) telephoto vision device ($\lambda_t = 40$mm, $\alpha_{t,pan,tilt} = [10\ 10]^{\circ T}$) of a multi-focal vision system.

measurements of two or three landmarks are taken. The significantly better performance is obvious which is noted by the much lower areas of the footstep confidence ellipses compared to all other vision system configurations, see Figure 5.13 and Figure 5.14.

The foveated vision systems, i.e. systems with large field of view and small central high-resolution region, which are used in many state-of-the-art robot heads, however, yield very different performances. The considered foveated scenarios vary in the field of view of the wide-angle camera. Looking at the performance measures in Table 5.1 the average number of visible landmarks differs by more than one landmark at each step. This results in a much higher localization uncertainty for the vision system with slightly smaller wide-angle field of view shown in Table 5.1. The confidence ellipse areas for this foveated system only reach the values of a conventional camera (see Figure 5.13) while the one with the larger field of view performs almost equally to the multi-focal system with independent cameras, see also Table 5.1. This can be explained by the fact that the wide-angle region is always shifted with the telephoto region. There is no possibility to adjust the wide-angle pose such that a sufficient number of landmarks is captured. Thus, the number of visible landmarks depends strongly on the field of view of the wide-angle camera. In the scenario with smaller wide-angle field of view only occasionally the wide-angle camera captures more than one landmark which can be seen in Figure 5.10. At most steps a configuration as seen in step #12 is given. So at many steps merely one landmark is visible.

Assessing these results, the advantages of multi-focal active vision in mobile robot navigation are obvious. Localization accuracy is strongly improved and in case of multi-focal vision with independent cameras sensor resources can be flexibly allocated depending on

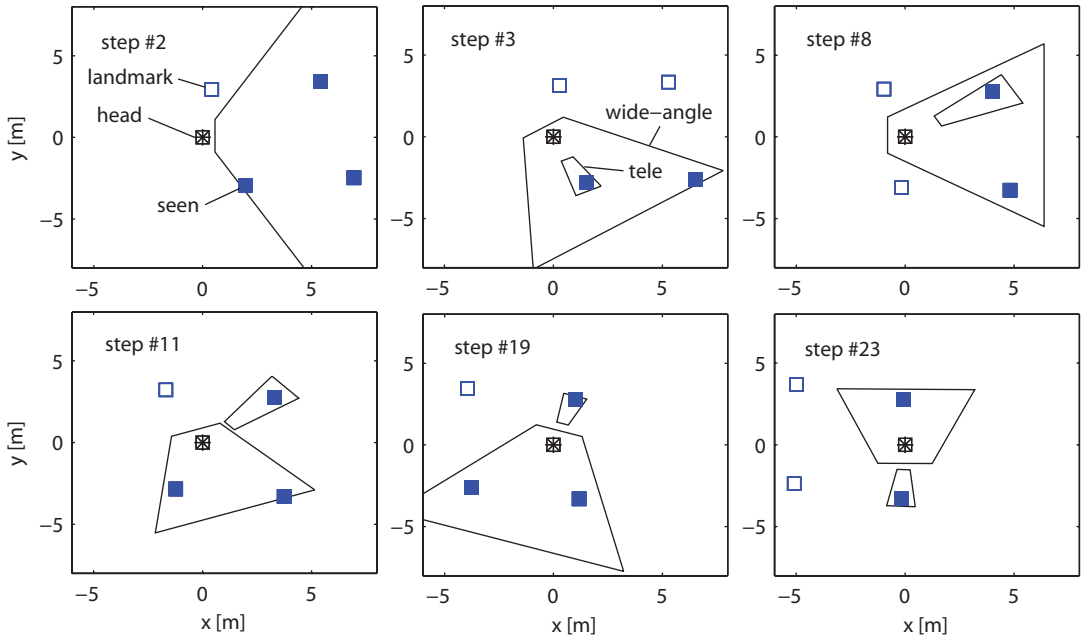

Figure 5.12: Projections of the field of view of the wide-angle and telephoto device of a multi-focal vision system of footsteps #2, #3, #8, #11, #19, and #23 of Figure 5.11.

the current task and situational requirements. In case of foveated active vision a certain field of view has to be provided depending on the environmental setting in order to achieve similar performance.

5.3.4 Discussion

The concept of multi-focal view direction planning for mobile robot navigation has been introduced in order to overcome various drawbacks of conventional strategies comprising only one type of or kinematically coupled vision devices. The performance of various vision system embodiments has been investigated based on a standard navigation scenario localizing a humanoid robot walking on a straight path through a structured environment. It has been demonstrated that multi-focal vision outperforms all other vision system embodiments due to its flexibility in resource allocation and additional high-accuracy sensors

Table 5.1: Mean of the Areas A_{90} of the 90%-confidence ellipses of the footstep covariance matrices and average number $\bar{N}_{vis}$ of visible landmarks for mono- and multi-focal robot localization scenarios.

scenario	$A_{90}[10^{-4}m^2]$	N_{vis}
telephoto	7700	0
conventional	2.7	1.1
foveated (60°)	2.4	1.3
foveated (80°)	1.7	2.3
wide-angle	2.1	1.8
multi-focal	1.6	2.4

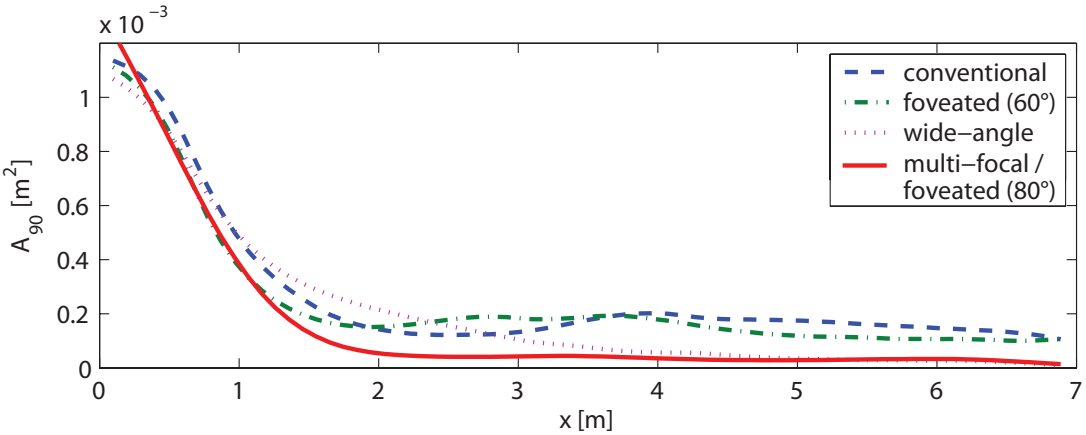

Figure 5.13: Areas A_{90} of the 90%-confidence ellipses of the footstep position estimates using a conventional, foveated, wide-angle, and multi-focal vision system, respectively.

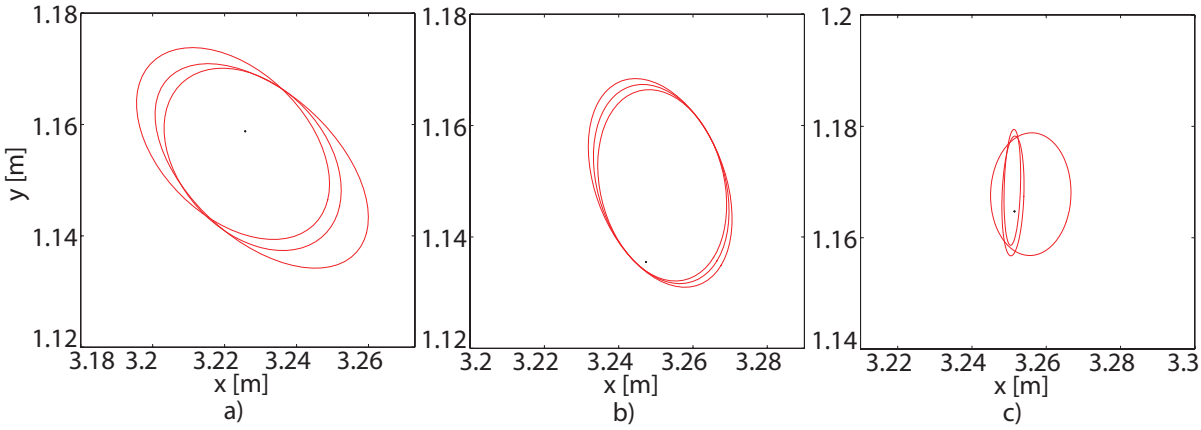

Figure 5.14: Comparison of the 90%-confidence ellipses of the footstep position estimates at step #12 using a a) conventional, b) wide-angle, and c) multi-focal vision system with three measurements per footstep.

resulting in significantly improved mission performance, i.e. localization accuracy, of the mobile robot. Foveated devices which are increasingly used perform similarly only if a sufficient field of view of the wide-angle camera is provided.

The main reason for the weaker performance of mono-focal system configurations is the trade-off between field of view and accuracy which substantially reduces the number of reference objects to be observed or the accuracy of object position measurements. A certain field of view is needed in order to continually make out next reference objects along the robot's path strongly limiting accuracy. In the extreme case of a telephoto camera with very narrow aperture angles no objects are detected at all. The foveated version, i.e. a multi-focal system with relative sensor poses fixed, suffers from the shortcoming that the view direction has to be determined based on the foveated region which potentially prevents objects to be detected by the wider-angle region if its aperture angles are too small.

Even though a relatively straight forward perception model has been used generalization to more complex models with distortions, quantization, etc., is possible. However, also in terms of nonlinear distortions a better performance of multi-focal vision can be expected as distortions in higher-resolution devices are comparably small due to narrow aperture

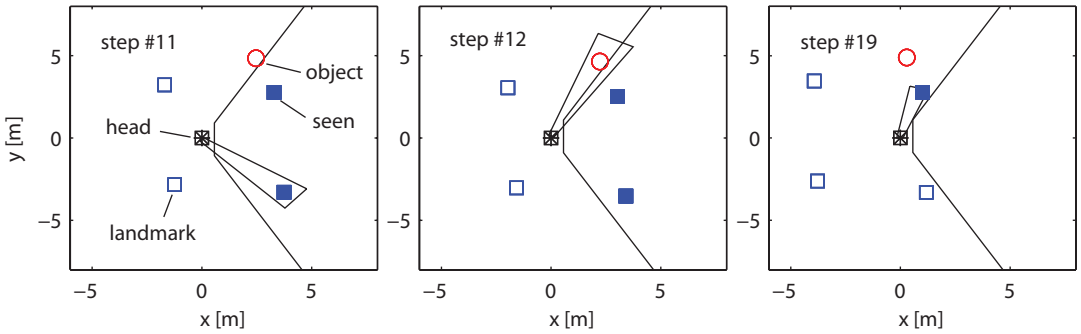

Figure 5.15: Planned pan-angles of a multi-focal vision system reacting to an event of interest trading it versus robot position uncertainty.

angles. In this work only point landmarks have been considered. In order to evaluate the impact of quantization effects on multi-focal vision performance also the impact of image processing algorithms has to be considered.

5.4 Secondary Tasks - Towards Multi-Focal Multi-Task Architectures

Beside the main objective - following the path - other tasks could be relevant to view direction planning, e.g. in order to guarantee safety or to perform secondary tasks, which could even have an impact on the overall mission. These tasks can be discrete or continuous in nature. In the previous section two tasks have been defined which are of particular relevance to multi-focal view direction planning: the reaction to events and the consideration of the predicted visibility. In the following the integration of both is discussed based on the navigation scenario from the previous section.

Event or Object of Interest The multi-focal view direction planning strategy is extended to consider potential events or objects of interest utilizing (5.11). The defined objective is to follow the global path and observe an interesting event whenever it does not interfere with the primary mission. The interest operator r is, therefore, defined such that the planning strategy switches to robot localization when a threshold for the position uncertainty is exceeded. This threshold is set yielding the mean of the position covariance ellipse main axes $\nu_0 = 0.005\text{m}$. Operator $r = f(\nu_0)$ is set to zero if this threshold is exceeded and to one, otherwise. The multi-focal setting from the previous section is used. The view direction of the telephoto device is controlled by the two-task strategy and that of the wide-angle device is constantly kept straight ahead. An event of interest is placed next to the upper right landmark in the environment shown in Figure 5.5.

The resulting fields of view corresponding to the planned view directions are shown in Figure 5.15. The switching between the two tasks is also depicted in Figure 5.16. At step #12 the telephoto device switches from localization to event tracking and back to localization at step #13. The propagation of the mean ν_0 of the main axes of the robot position covariance ellipse is depicted in Figure 5.17 showing the oscillation around the set threshold.

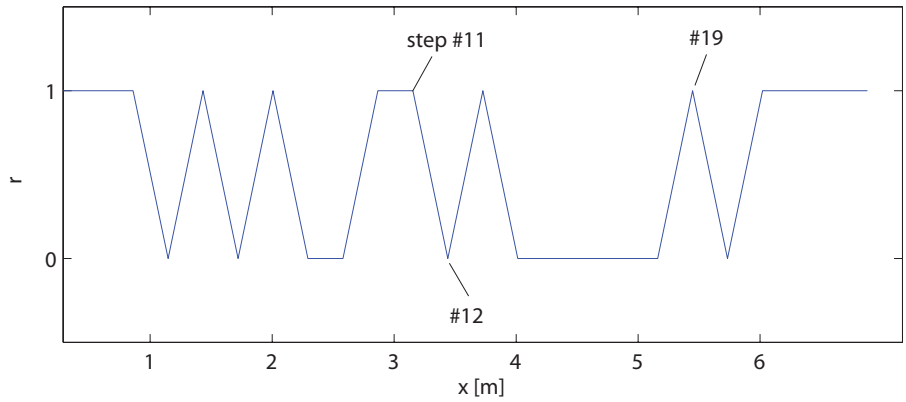

Figure 5.16: Selected state of interest operator r corresponding to $r = 0$: object of interest, $r = 1$: robot localization.

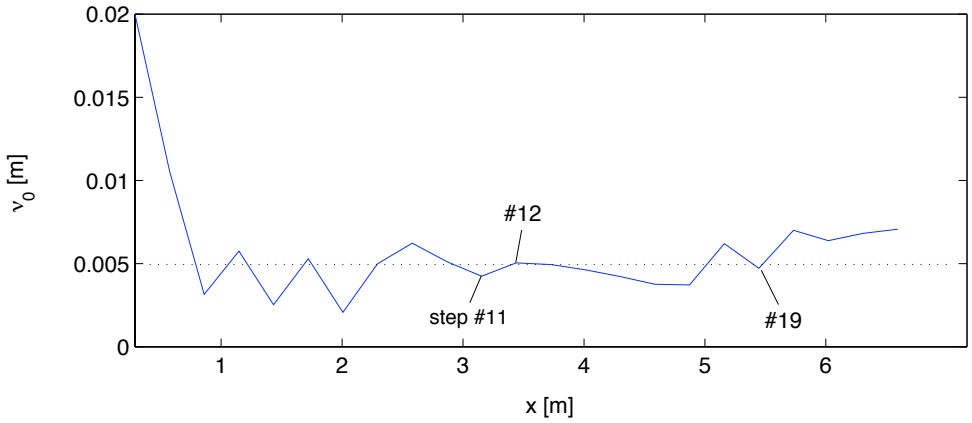

Figure 5.17: Means of the main axes of the footstep covariance ellipses; threshold for robot localization at $\nu_0 = 0.005$m.

Predicted Visibility In order to demonstrate the impact of the predicted visibility of an object, an object of interest is placed at the same position of the event from the previous scenario. Utilizing the planning strategy in Eq. 5.12 the telephoto device now switches to observe the object whenever its position covariance ellipse exceeds the limits of the field of view if focused. A result is depicted in Figure 5.18 showing the predicted covariance ellipse before a view direction shift (Figure 5.18a) clearly exceeding the field of view of the camera when observed which is shown in Figure 5.18b. As a result from view direction shifting the covariance ellipse is clearly located within the field of view (Figure 5.18b).

A straight forward way of integrating secondary tasks relevant to multi-focal view direction planning has been shown. In arbitrary and changing real-world scenarios, however, a basic thresholding strategy with fixed task weightings cannot provide the flexibility needed in order to facilitate robust, adaptive, and safe behavior of the robot. An architecture with units for higher-level situation assessment is necessary evaluating the trade-off of all mission- and task-related aspects. This generally leads to a decision problem which could be managed with state-of-the-art decision-theoretic or attentional models. Moreover, a comprehensive environmental model is essential capturing the task-relevance of structures and events, e.g. utilizing probabilistic measures. These aspects open up a variety of re-

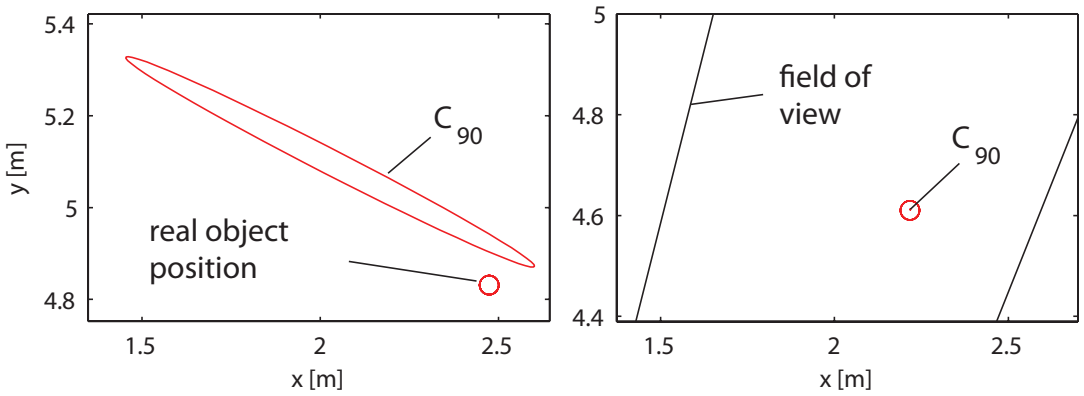

Figure 5.18: Visibility of an observed object; 90%-confidence ellipses of estimated object position a) before and b) after observation with a telephoto vision device (after camera shift).

search directions in multi-focal view direction planning which go far beyond the scope of this thesis and are subject to future research.

5.5 Discussion

Active vision for robot navigation has become an essential tool in order to provide flexible orientation in complex environmental settings. The selection of the vision system and the coordination of the view direction of active vision face a trade-off between field of view and accuracy. Mono-focal vision system embodiments suffer from several shortcomings impairing localization accuracy and, thus, navigation performance. In order to explore the world a certain field of view has to be provided. The more reference objects for orientation are visible the better are the estimates of the robot pose. However, the larger the field of view the worse is the measurement accuracy. High-accuracy *and* large field of view sensors require extensive computational resources and produce a vast amount of data with low usable information density.

In this chapter a novel multi-focal approach to navigation with active vision has been proposed in order to overcome these drawbacks. The approach consists in the utilization of a multi-focal vision system with coupled or independent vision devices and a multi-focal view direction planning strategy. Several planning conditions have been defined considering the localization accuracy, events or objects of interest to be observed, and the predicted visibility of objects. An extended strategy has been proposed facilitating the observation of events or other objects of interest, e.g. with higher accuracy, which is only possible with multi-focal systems. The approach has been evaluated in a humanoid robot navigation scenario and the performance has been compared to mono-focal active vision approaches.

In this thesis foveated and multi-focal vision systems have been used for robot navigation for the first time. The results show a significant improvement of the localization accuracy of the robot with multi-focal active vision in comparison to all other investigated approaches. Moreover, utilization of multi-focal vision and the proposed planning strategy make an individual flexible observation of selected locations of interest in robot navigation possible. These findings open up promising research directions in active vision-based robot navigation in the context of attentional models for multi-focal active vision accounting for

higher-level scene assessment and environmental modeling, which are subject to future work.

6 Conclusions and Future Directions

6.1 Concluding Remarks

The work presented focuses on the methodical investigation of multi-focal vision for measurement and robotics at various abtraction levels. Individual research areas cover statical and dynamical performance parameters as well as information-based planning issues for camera coordination considering variations of geometrical configuration and sensor devices. Multi-focal vision systems are known since several decades, however, only very few methodical works are known exploiting the advantages of multi-focal vision. In an extensive framework, this work sheds light on the beneficial impact of multi-focal vision on measurement quality, control performance, and flexibility in resources allocation demonstrated in a manifold of examples and applications. Conceptual innovative are multi-focal approaches to vision-based manipulator control and active vision for mobile robots significantly improving performance compared to conventional approaches on the one hand and on the other hand facilitate application scenarios which are not realizable using mono-focal approaches. The main approaches along with the major results are highlighted in the following.

The performance characteristics of vision systems comprising only one type of vision sensors are well covered in known literature. Investigations reach from single to multi-sensor configurations and even the fusion of different sensor types is known. Yet, methodical investigations on how configurations of different sensor types influence the measurement quality are not covered by common literature. In Chapter 3 a fundamental investigation of the perception performance of multi-focal vision systems is conducted. It is shown that the sensitivity of the vision system can be improved in selected regions by adding higher-sensitivity sensors. Thereby, the benefit of a large field of view can be maintained. Methods are proposed in order to determine appropriate foci of attention in order to optimize sensitivity. Multi-focal systems have a shortcoming of a weaker condition number than comparable mono-focal systems. Rotating the sensitivity ellipsoid and, thereby, moving its larger main axes towards particular Cartesian directions of interest by adjusting the focal-lengths and optional rotation of the vision devices, this drawback is turned into an advantage: Sensor resources can be reduced by certain combined configurations of low-sensitivity and high-sensitivity sensors compared to mono-focal high-sensitivity configurations. These results contribute to facilitate situational adaptation of system configuration and sensor resources to the current requirements in order to optimize measurement performance, sensor resources, and computational cost.

Common vision-based control approaches suffer from several drawbacks: A certain number and configuration of reference features has to be observed in order to assure a certain control performance and the performance varies with the observation distance. The former requires a sufficient field of view and the latter a sufficient focal-length of the vision device. Both are limited by the region spanned by the features and the minimum distance to the

vision system in order to assure their visibility. In Chapter 4 two innovative multi-focal approaches to vision-based manipulator control are proposed which overcome these drawbacks. A novel hybrid control strategy switches between several vision devices according to the current performance and field of view requirements. The contribution is a guaranteed control performance within a defined bounded performance region. The second proposed strategy allocates high-sensitivity vision devices to certain selected features in order to improve sensitivity. The remaining features are observed with a low-sensitivity large field of view vision device in order to render the controller full rank. Thereby, control performance is significantly improved. This aspect is only realizable with multi-focal systems. Using only high-sensitivity sensors in a mono-focal system setup would require a sufficient number of sensors to observe all necessary features and, thus, additional resources. Moreover, the two strategies are combined exploiting the benefits of both. The innovative concept of multi-focal vision-based control provides a flexible method to improve control performance and reduce perceptual resources.

From a higher-level perspective control of active vision has to consider the current mission and situational requirements in order to select appropriate view directions for the individual vision devices. While Chapter 4 considers dynamical issues, in Chapter 5 the information gain of multi-focal vision system configurations is assessed in order to derive optimal view directions for visual guidance of a mobile robot in the current situation. The novel concept of multi-focal active vision for robot navigation is applied to a humanoid walking robot scenario considering a foveated system and a system comprising two independent stereo cameras. The contribution of this flexible perceptual resource allocation method manifests itself in a significantly improved localization accuracy and improved reactive behavior. Multi-focal vision with independent sensors turns out to outperform all other considered systems in performance and reactivity. The novel perceptual and active vision concepts consitute a promising signpost for future research in mobile systems with high perceptual capabilities and autonomy.

Summarizing, the ideas, concepts, and approaches developed in this thesis significantly advance the state of the art in design and control of vision, active vision, and visual servoing systems. The results are expected to have a high impact on the development of future systems and applications in many research areas as, e.g., robotics, transportation, and surveillance as well as perception for autonomous and cognitive systems.

6.2 Outlook

Vision is one of the most powerful information sources. Current and future advances in computing power and bus systems go hand in hand with advances in sensor technology and increasing demands on information processing. Thus, increasing computational resources have to process increasing amounts of data and the need for information pre-filtering and resource optimization will always be essential. Multi-focal vision as a flexible, efficient, and economical concept is an elementary tool to keep up with these increasing demands and to improve the performance of vision-based applications. The fundamental research on design, control, and planning aspects of multi-focal vision presented in this thesis constitutes the basis for future developments in this area. There is a number of exciting research directions directly emerging from this thesis, some of which are:

- *Quantization effects and error propagation* - Yet, effects of quantization have only been considered in a few works, e.g. quantized feedback stabilization of nonlinear systems and image processing. In order to assess the effects of quantization and probabilistic errors the whole processing chain has to be evaluated covering sensor characteristics, feature extraction, data fusion and processing algorithms, transformation of the dynamical system, etc. The propagation of quantization errors and sensor noise through the control system is an important next step in multi-focal systems research. Some interesting effects pending to be quantified are steady-state errors and limit cycles as well as stability issues.

- *Multi-focal multi-level data fusion* - Only some works are known considering the fusion of multi-focal visual information yet, e.g. multi-resolution disparity maps. However, known approaches are limited to certain abstraction levels. A future challenge in multi-focal vision is the fusion of multi-resolution data on statical, dynamical, and higher levels such as knowledge representation.

- *Higher-level situation assessment and multi-focal attention models* - In order to make use of multi-focal efficiency and flexibility with respect to mission and situational demands an assessment of the current situation is necessary. Higher-level information extraction, learning, knowledge representation, and decision issues have to be considered in order to cope with complex and real-world environments and to continually adapt the configuration and view direction of the multi-focal system. This will be an essential aspect in multi-focal perception and active vision for future autonomous and cognitive systems research.

Many aspects of multi-focal vision and of the research presented in this thesis are not limited to vision systems and are basically applicable to any multi-resolution sensor setting exploiting the mentioned benefits. Research on multi-focal vision will have a large impact on integrated concepts of multi-sensor systems and sensor networks and vice versa. Promising will be the joint research on multi-focal resource allocation and technological and biologically inspired attentional mechanisms where significant synergies are expected which will highly advance the state of the art with high impact on future technology and applications.

A Experimental Vision System

Within the framework of the development of LOLA, the successor model of the humanoid robot JOHNNIE, a multi-focal vision system has been developed [71]. The system comprises four cameras mounted on a 6 DOF platform. An extreme wide-angle stereo-camera pair is mounted on a central pan/tilt-platform. Two additional cameras forming a second stereo pair with high resolution are gimbal-mounted on the central platform. The system is designed for very high velocities and accelerations of the gimbal-mounted cameras in order to outperform human capabilities. An embedded motion controller with CAN interface provides an easy integration into a top-level control architecture.

Intended for real world outdoor environments, the main design considerations for this vision system are a selective high measurement accuracy of object and robot positions, a simultaneous detection of objects of interest within a wide field of view, the recognition of distant objects, and fast and flexible reaction to complex dynamic scenes. The application on mobile autonomous platforms requires lightweight design and low power consumption.

System Overview

The multi-focal vision system is shown in Figure A.1. The system comprises four digital IEEE1394-cameras with lenses of various focal-lengths. A commercially available wide-angle stereo pair (Bumblebee, Point Grey Inc.) is mounted on a central pan-tilt platform. Its baseline is 12cm and the focal-lengths are 2mm each. The aperture angles are approximately 85° from the optical axis for each camera and 50° for the range of stereo vision. Two cameras (Dragonfly, Point Grey Inc.) equipped with telephoto lenses with focal-lengths of 25mm each are gimbal-mounted on the pan-tilt platform. Their aperture angles are approximately 5°. This results in a maximum angular resolution of approximately 0.02°. The maximum baseline of the telephoto cameras is 31cm if both cameras are oriented straight ahead. The maximum video framerate of all cameras is 30fps.

Brushless DC motor direct drives are used to drive the 4 DOF of the gimbal-mounted cameras ensuring minimum friction and high accelerations. The 2 DOF of the pan-tilt platform are driven by DC motors with harmonic drive gears. All position sensors are single-turn conductive plastic precision potentiometers with ball-bearings.

The whole system body is made of aluminum alloy. The centers of mass of all sub-components are placed approximately in the corresponding joint axis and the moments of inertia are kept as small as possible resulting in pose independent and minimum drive torques. Maximum dimensions of the vision system are 37x30x5cm. The weight is $\sim$ 2.2kg.

Kinematics

The kinematic structure of the vision system is shown in Figure A.2. The forward kinematics are derived based on the coordinate frames shown. The reference frame S_r is placed in the intersection point of the joint axes of the central pan-tilt platform. The effector frames correspond to the camera frames for the gimbal-mounted (telephoto) cameras and the frame placed in the plane of symmetry between both of the cameras of the central stereo pair, respectively. The link lengths are shown in Table A.1.

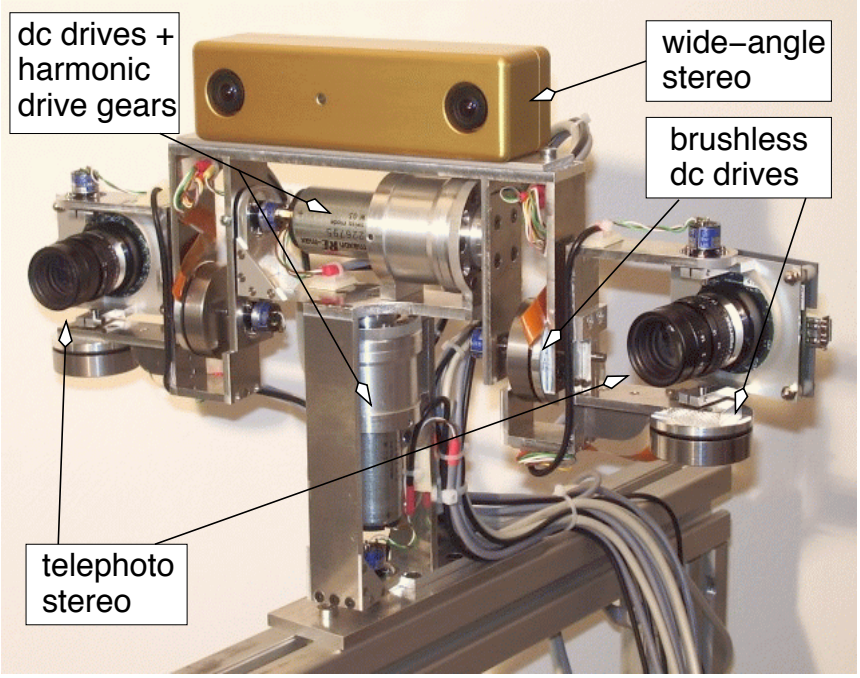

Figure A.1: Multi-focal high-performance vision system.

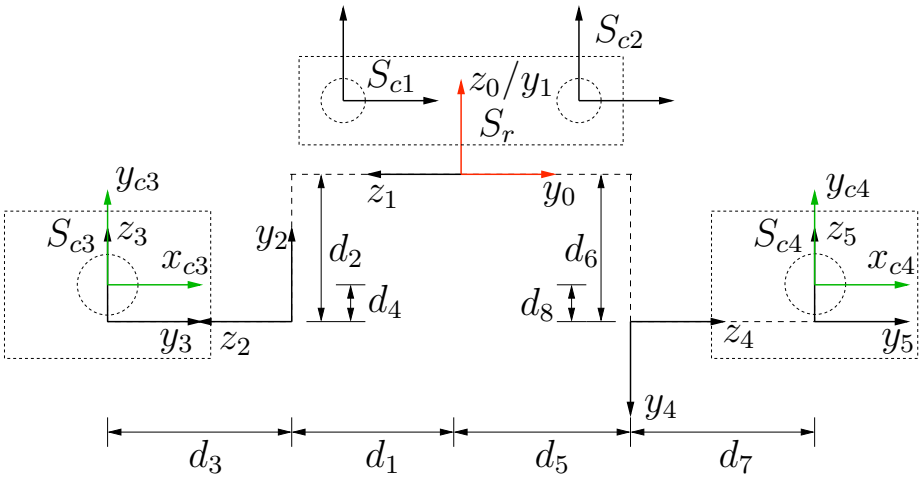

Figure A.2: Kinematic structure of the vision system.

Table A.1: Link legths of the multi-focal vision system

link	d_i [m]
d_1, d_5	0.065
d_2, d_6	0.045
d_3, d_7	0.089
d_4, d_8	0.015

Dynamics

The maximum torques of the central dc drives and the outer direct drives, and the moments of inertia with respect to the individual joint axes are shown in Table A.2. The moments of inertia are determined using the modeling software SolidWorks (SolidWorks Corp.) based on a detailed model of the vision system and measured parameters as mass and dimensions. The expected maximum accelerations and velocities are coarsely estimated based on maximum drive torques and moments of inertia

$$\tau_{i,max} = I_i \ddot{\theta}_{i,max},$$

where $\tau_{i,max}$ is the stall torque, I_i the estimated moment of inertia, and $\ddot{\theta}_{i,max}$ the maximum angular acceleration for the joint axis i. The velocity after a 360°-turn is taken for maximum value

$$\dot{\theta}_{i,360^\circ} = \sqrt{720^\circ \ddot{\theta}_{i,max}}.$$

Table A.2: Maximum drive torques and moments of inertia

axis	$\tau_{i,max}$ [Nm]	I_i [kgm^2]	$\dot{\theta}_{i,360^\circ}$ [°/s]	$\ddot{\theta}_{i,max}$ [°/s^2]
1	4.8	$1.63{\cdot}10^{-2}$	2470	16900
2	4.8	$3.77{\cdot}10^{-3}$	5120	72900
3; 5	$260{\cdot}10^{-3}$	$149{\cdot}10^{-6}$	8490	100000
4; 6	$260{\cdot}10^{-3}$	$200{\cdot}10^{-6}$	7320	74500

Mechatronics and Control Architecture

The general control architecture is shown in figure A.3. The control algorithm is implemented on an MPC555 (Motorola), a 32-bit PowerPC RISC microcontroller with a core performance of 52.7kmips at 40MHz. The chip comprises a dual CAN 2.0B serial communication protocol controller and two time processor units (TPU) controlling the 8 channel PWM sub-modules.

Using the on-chip PWM-modules and six I/O pins the power electronics module is controlled and galvanically isolated. The power electronics module consists of two H-bridges and four brushless dc driver ICs controlling the drive section.

Each of the gimbal-mounted telephoto cameras is driven by two brushless dc direct drives (30W, EC45, Maxon). The central pan-tilt platform is driven by dc drives (22W, REmax29, Maxon) with Harmonic Drive gears (100:1).

Single-turn conductive plastic precision potentiometers with ball-bearings (MCP05, 1kΩ, Megatron) are used as position sensors. The sensor signal is converted by a six-channel 16-bit analog-digital converter with simultaneous sampling capability. An anti-aliasing-filter of second order is integrated. The sensor data are memory mapped to the controller.

Via CAN serial communication protocol the communication to a host PC is managed exchanging the current and desired angular orientations. The overall architecture provides a modular and easy integration into a top-level decision and control architecture.

The no-load power consumption is approximately 1W and 27W for the signal and power parts, respectively, and 5W and 164W in full-load.

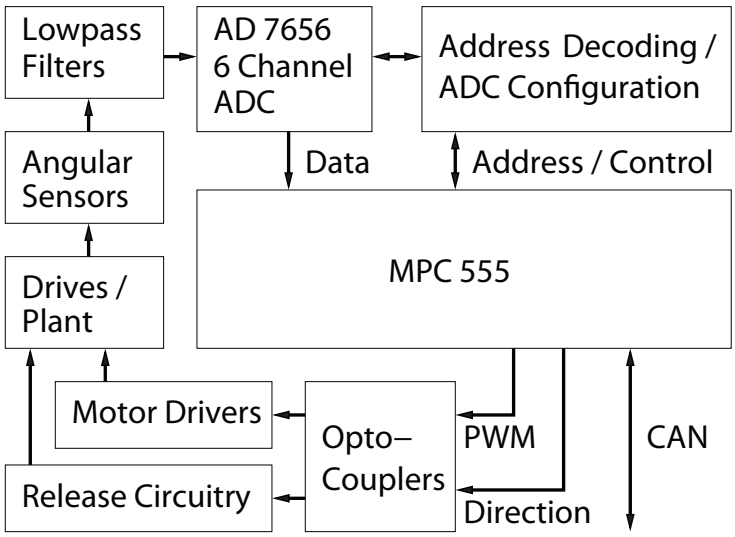

Figure A.3: Architectural structure of the mechatronical system.

Currently, a simple cascade control strategy is used with outer position control and inner velocity control.

Dynamic Performance

The vision system is designed to reach very high velocities and accelerations. In order to evaluate the dynamic performance of the vision system, step responses are recorded, which are shown in Figures A.4, A.5 and A.6.

Figure A.4 shows the step response of the position controlled right gimbal-mounted camera to an orientation step of the pitch-angle of 54° corresponding to a sensor output change of 0.75V. A rise time of $t_{90\%} = 76$ms can be noted.

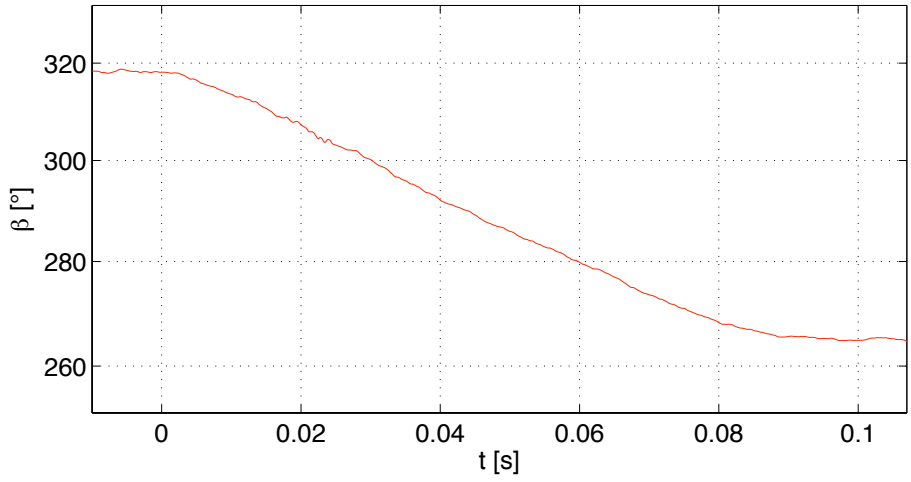

Figure A.4: Pitch-angle step response of right gimbal; desired step $\Delta\beta = 54°$.

The step response to an orientation step of the yaw-angle of the position controlled right gimbal-mounted camera system is shown in Figure A.5. An orientation step of 72°

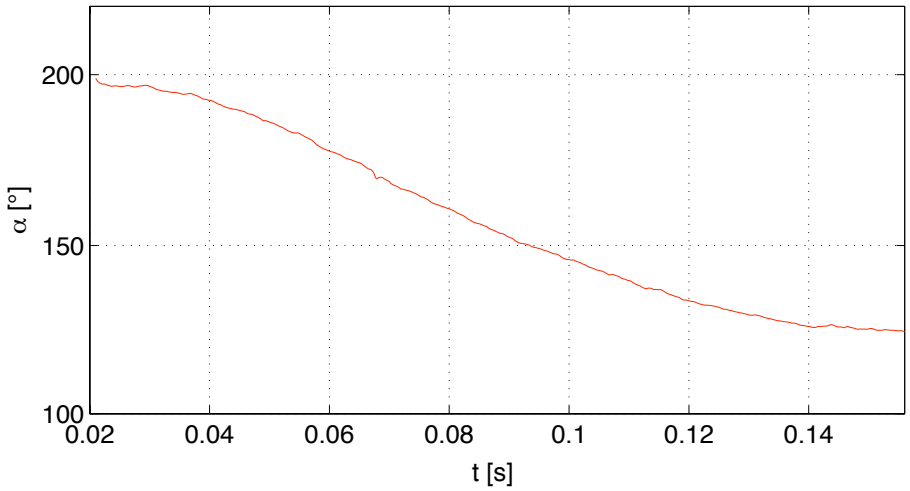

Figure A.5: Yaw-angle step response of right gimbal; desired step $\Delta\alpha = 72°$.

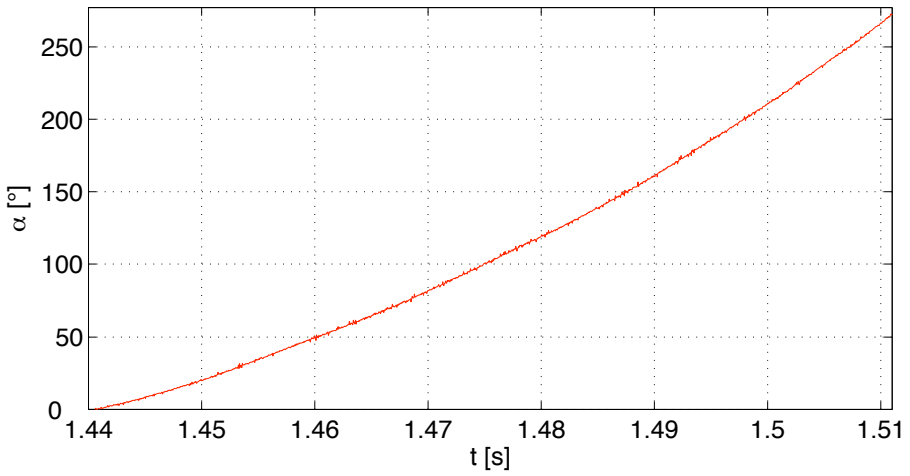

Figure A.6: Yaw-angle step response to maximum set value of right gimbal starting from rest pose.

corresponding to a sensor voltage output change of 1V is commanded. The rise time is approximately $t_{90\%} = 90$ms.

The response of the uncontrolled right gimbal-mounted camera system to maximum manipulated value is shown in Figure A.6. A whole rotation of 360° around its yaw-axis is performed in approximately 86ms. Evaluating this, a maximum angular velocity after a 360°-turn of approximately 8400°/s and a maximum angular acceleration of about 97300°/s^2 are reached. For comparison, the peak velocities and accelerations of the human eye [16] are estimated to about 900°/s and 20000°/s^2.

B Visual Servoing Performance Metrics

The performance of visual servoing is commonly evaluated in terms of a feature error norm. The pose of the robot manipulator is less considered and mainly a matter of experimental validations. In this work the error and variance of the vision sensor pose are considered. The translational and rotational parts of the average remaining pose error are computed separately after the assumed convergence of the system. These are defined as

$$\begin{aligned} \hat{e}_{tran} &= \frac{1}{N-k_1}\sum_{k=k_1}^{N-1} \|x^t(kT) - x^{t,d}\|_2, \\ \hat{e}_{rot} &= \frac{1}{N-k_1}\sum_{k=k_1}^{N-1} \|x^r(kT) - x^{r,d}\|_1, \end{aligned}$$

with Cartesian pose vector x, $(.)^t$ and $(.)^r$ denoting the translational and rotational components, respectively, $(.)^d$ denoting the desired pose, N maximum number of samples, sampletime T, and $k_1 T$ the time step after which the system is considered converged.

A good estimate for the translation error variance is given by

$$\hat{\sigma}^2_{tran} = \frac{1}{N-k_1}\sum_{k=k_1}^{N-1} \|x^t(kT) - x^{t,d}\|_2^2 - \hat{e}^2_{tran},$$

and similarly for the rotation error the variance estimate is defined as

$$\hat{\sigma}^2_{rot} = \frac{1}{N-k_1}\sum_{k=k_1}^{N-1} \|x^r(kT) - x^{r,d}\|_1^2 - \hat{e}^2_{rot}.$$

In case of trajectory following tasks, the tracking error is computed as

$$\hat{e}_{tran}(kT) = \|x^t(kT) - x^{t,d}(kT)\|_2,$$

with desired trajectory $x^{t,d}(kT)$ and similarly for the rotational part. Regarding the variance estimates a small time window of W steps is evaluated giving

$$\hat{\sigma}^2_{tran}(kT) = \frac{1}{W}\sum_{n=k-W+1}^{k} \|x^r(nT) - x^{n,d}\|_2^2 - \hat{e}^2_{tran,W}(kT),$$

with $e_{tran,W}(kT)$ the mean of the W consecutive translation error values, and similarly for the rotational part. This definition gives rather coarse, but comparable results.

The error in image space is defined straight forward as the mean of the offset vectors between the current and desired feature positions

$$\hat{e}_{pix} = \frac{1}{I(N-k_1)}\sum_{k=k_1}^{N-1}\sum_{i=1}^{I} \|\xi_i(kT) - \xi_i^d\|_2,$$

with ξ_i, ξ_i^d the current and desired component vectors of feature point i and the number of feature points I. The variance estimate of the feature error vector is defined as

$$\hat{\sigma}_{pix}^2 = \frac{1}{I^2(N-k_1)} \sum_{k=k_1}^{N-1} \left[\sum_{i=1}^{I} \|\xi_i(kT) - \xi_i^d\|_2 \right]^2 - e_{pix}^2,$$

In case of trajectory following tasks, the feature error vector is evaluated according to the definitions regarding the pose error.

Bibliography

[1] P. Allen, A. Timcenko, B. Yoshimi, and P. Michelman. Automated tracking and grasping of a moving object with a robotics hand-eye system. *IEEE Transactions on Robotics and Automation*, 9(2):152–165, 1993.

[2] M. Alpern. The position of the eyes during prism vergence. A.M.A. Arch. Ophthalmol., Vol. 57, pp. 345–353, 1957.

[3] B. Ambler and D. Finklea. The influence of selective attention in peripheral and foveal vision. *Perception and Psuchophysics*, 19(6):518–524, 1976.

[4] R. Andersson. Dynamic sensing in ping-pong playing robot. *IEEE Transactions on Robotics and Automation*, 5(6):728–739, 1989.

[5] N. Apostoloff and A. Zelinsky. Vision in and out of vehicles: Integrated driver and road scene monitoring. In *Proc. of the 8th International Symposium on Experimental Robotics (ISER 2002)*, Sant Angelo d'Ischia, Italy, 2002.

[6] M. Arlotti and M. Granieri. A perception technique for a 3D robotic stereo eye-in-hand vision system. In *Proceedings of the 5th International Conference on Advanced Robotics (ICAR 1991)*, pages 1626–1629, 1991.

[7] A. Arsenio. M4 Project: Adding an Active Vision Head to the M4 Robot, March 2000. Report to DARPA.

[8] R. Bajcsy, J. Kosecka, and H.I. Christensen. Discrete event modeling of navigation and gaze control. *International Journal of Computer Vision*, 14(2):179–191, 1995.

[9] A. Baumberg. Reliable feature matching across widely separated views. In *Proceedings on the IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[10] J. Beck and B. Ambler. The effects of concentrated and distributed attention on peripheral acuity. *Perception and Psychophysics*, 14:225–230, 1973.

[11] R. Bodor, R. Morlok, and N. Papanikolopoulos. Dual-camera system for multi-level activity recognition. In *Proc. of the 2004 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2004)*, pages 643–648, Sendai, Japan, 2004.

[12] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robots*, 1(2):315–348, 2004.

[13] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scasselati. Active vision for sociable robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 31(5):443–453, 2001.

[14] R. A. Brooks, C. Breazeal, M. Marjanovic, B. Scasselati, and M. M. Williamson. *The Cog Project: Building a Humanoid Robot*, pages 52–87. Springer, Heidelberg, Berlin, Germany, 1999.

[15] T. Buschmann, S. Lohmeier, **K. Kühnlenz**, M. Buss, F. Pfeiffer, and H. Ulbrich. Lola – a performance enhanced humanoid robot. *it – Information Technology*. to appear.

[16] R.H.S. Carpenter. *Movements of the Eyes*. Pion, London, 1988.

[17] F. Chaumette, K. Hashimoto, E. Malis, and P. Martinet. Ttp4 : Tutorial on advanced visual servoing. Tutorial Notes, IEEE/RSJ IROS 2004, 2004.

[18] X. Clady, F. Collange, F. Jurie, and P. Martinet. Object tracking with a pan-tilt-zoom camera : application to car driving assistance. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2001)*, pages 1653–1658, Seoul, Korea, 2001.

[19] J.J. Clark and N.J. Ferrier. Modal control of an attentive vision system. In *Proceedings of the International Conference on Computer Vision*, pages 514–523, 1988.

[20] P.I. Corke. *High-Performance Visual Closed-Loop Robot Control*. PhD thesis, University of Melbourne, 1994.

[21] P.I. Corke. *Visual Control of Robot Manipulators – A review*, pages 1–32. World Scientific, 1994.

[22] P.I. Corke and M. Good. Dynamic effects in visual closed-loop systems. *IEEE Transactions on Robotics and Automation*, 12(5):671–696, 1996.

[23] P.I. Corke and S.A. Hutchinson. A new partitioned approach to image-based visual servo control. In *Proceedings of the 31st International Symposium on Robotics*, pages 30–35, Montreal, Canada, 2000.

[24] N. Correll. 6-DOF visual servoing using the Lie group of affine transformations. Technical Report ISRN LUTFD2/TFRT--5690--SE, Department of Automatic Control, Lund Institute of Technology, Sweden, 2002.

[25] N. Cowan. Binocular visual servoing with a limited field of view. In *Mathematical Theory of Networks and Systems*, Notre Dame, Indiana, 2002.

[26] N.J. Cowan and D.E. Chang. Geometric visual servoing. *IEEE Transactions on Robotics*, 21(6):1128–1138, 2005.

[27] T. Darrell. Reinforcement learning of active recognition behaviors. Interval Research Technical Report 1997-045. http://www.interval.com/papers/1997-045, 1997.

[28] J. Davis and X. Chen. Foveated observation of shape and motion. In *Proc. of the 2003 IEEE Int. Conf. on Robotics and Automation (ICRA 2003)*, pages 1001–1005, Taipei, Taiwan, 2003.

[29] A.J. Davison. *Mobile Robot Navigation Using Active Vision*. PhD thesis, Robotics Research Group, Department of Engineering Science, University of Oxford, UK, 1999.

[30] A.J. Davison. SLAM with a single camera. In *Workshop on Concurrent Mapping and Localization for Autonomous Mobile Robots at ICRA 2002*, Washington, DC, USA, 2002.

[31] A.J. Davison, O. Stasse, and K. Yokoi. Vision based SLAM for a humanoid robot. In *SLAM Workshop, International Conference on Robotics and Automation (ICRA 2005)*, Barcelona, Spain, 2005.

[32] K. Deguchi. Optimal motion control for image-based visual servoing by decoupling translation and rotation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 1998)*, pages 705–711, 1998.

[33] K. Deguchi. A direct interpretation of dynamic images with camera and object motions for vision guided robot control. *International Journal of Computer Vision*, 37(1):7–20, 2000.

[34] L. Deng, F. Janabi-sharifi, and W.J. Wilson. Hybrid motion control and planning strategies for visual servoing. *IEEE Transactions on Industrial Electronics*, 52(4):1024–1038, 2005.

[35] R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

[36] E. D. Dickmanns. Expectation-Based, Multi-Focal, Saccadic (EMS) Vision for Dynamic Scene Understanding. In *Nordic Signal Processing Symposium*, Norway, October 2002.

[37] E. D. Dickmanns. An Advanced Vision System for Ground Vehicles. In *International Workshop on In-Vehicle Cognitive Computer Vision Systems (IVC2VS)*, Graz, Austria, April 2003.

[38] R.L. Didday and M.A. Arbib. Eye movements and visual perception: A "two visual system" model. *International Journal of Man-Machine Studies*, 7:547–569, 1975.

[39] G. Dissanayake, P. Newman, S. Clark, H. Durrant-Whyte, and M. Csorba. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotics and Automation*, 17(3), 2001.

[40] R.W. Ditchburn. *Eye Movements and Visual Perception*. Oxford:Clarendon Press, 1973.

[41] T. Drummond and R. Cipolla. Application of Lie algebras to visual servoing. *International Journal of Computer Vision*, 37(1):21–41, 2000.

[42] J. H. Elder, F. Dornaika, B. Hou, and R. Goldstein. *Attentive wide-field sensing for visual telepresence and surveillance*. Academic Press, Elsevier, 2004.

[43] M. Elena, M. Christiano, F. Damiano, and M. Bonfe. Variable structure pid controller for cooperative eye-in-hand/eye-to-hand visual servoing. In *Proceedings of the IEEE International Conference on Control Applications (CCA 2003)*, pages 989–994, 2003.

[44] C.W. Erlksen and J. St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics*, 40(4):225–240, 1986.

[45] J. F. Seara. *Intelligent Gaze Control for Vision-Guided Humanoid Walking*. PhD thesis, Institue of Automatic Control Enginnering, Technische Universität München, 2004.

[46] J. F. Seara, K. H. Strobl, E. Martin, and G. Schmidt. Task-oriented and Situation-dependent Gaze Control for Vision Guided Autonomous Walking. In *Proceedings of the IEEE/RAS International Conference on Humanoid Robots (Humanoids 2003)*, München and Karlsruhe, Germany, 2003.

[47] J. Feddema, C. Lee, and O. Mitchell. Automatic selection of image features for visual servoing of a robot manipulator. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 1989*, pages 832–837, Scottsdale, AZ, USA, 1989.

[48] J. C. Fiala, R. Lumia, K. J. Roberts, and A. J. Wavering. Triclops: A tool for studying active vision. *International Journal of Computer Vision*, 12(2–3):231–250, 1994.

[49] G. Flandin, F. Chaumette, and E. Marchand. Eye-in-hand/eye-to-hand cooperation for visual servoing. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2000)*, pages 2741–2746, 2000.

[50] N.R. Gans and S. Hutchinson. An asymptotically stable switched system visual controller for eye-in-hand robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)*, 2003.

[51] N.R. Gans, S.A. Hutchinson, and P.I. Corke. Performance tests for visual servo control systems, with application to partitioned approaches to visual servo control. *The International Journal of Robotics Research*, 22(10-11):955–981, 2003.

[52] V. Gengenbach, H.-H. Nagel, M. Tonko, and K. Schäfer. Automatic dismantling integrating optical flow into a machine-vision controlled robot system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 1996)*, pages 1320–1325, 1996.

[53] G.D. Hager. Calibration-free visual control using projective invariance. In *Proceedings of the 5th International Conference on Computer Vision (ICCV 1995)*, pages 1009–1015, 1995.

[54] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, New York, NY, USA, 2000.

[55] K. Hashimoto and T. Noritsugu. Performance and sensitivity in visual servoing. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 1998)*, pages 2321–2326, 1998.

[56] K. Hashimoto and T. Noritsugu. Potential problems and switching control for visual servoing. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000)*, pages 423–428, 2000.

[57] E. Hayman. *The use of zoom within active vision.* PhD thesis, University of Oxford, 2000.

[58] N. Hollighurst and R. Cipolla. Uncalibrated stereo hand-eye coordination. *Image and Vision Computing*, 12(3):187–192, 1994.

[59] R. Horaud, D. Knossow, and M. Michaelis. Camera cooperation for achieving visual attention. *Machine Vision and Applications*, 16(6):331–342, 2006.

[60] K. Hosoda, H. Moriyama, and M. Asada. Visual servoing utilizing zoom mechanism. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 1995)*, pages 178–183, 1995.

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

[62] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3), 2001.

[63] N. D. Jankovic and M. D. Naish. Developing a modular spherical vision system. In *Proc. of the 2005 IEEE Int. Conf. on Robotics and Automation (ICRA 2005)*, pages 1246–1251, Barcelona, Spain, 2005.

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

[65] R. Kelly. Robust asymptotically stable visual servoing of planar robots. *IEEE Transactions on Robotics and Automation*, 12:759–766, 1996.

[66] R. Kelly, R. Carelli, O. Nasisi, B. Kuchen, and F. Reyes. Stable visual servoing of camera-in-hand robotic systems. *IEEE Transactions on Mechatronics*, 5(1):39–48, 2000.

[67] N. Kita. Intelligent plant inspection by using foveated active vision sensor. In *Proceedings of the International Conference on Human-Computer Interaction*, München, Germany, 1999.

[68] C. Koch and S. Ullmann. Selecting one among the many: A simple network implementing shifts in visual attention. MIT AI Memo No. 770, 1984.

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

[70] D. Kragic and H.I. Christensen. Using a redundant coarsely calibrated vision system for 3D grasping. In *Proceedings of Computational Intelligence for Modeling, Control, and Automation (CIMCA 1999)*, pages 91–97, 1999.

[71] **K. Kühnlenz**, M. Bachmayer, and M. Buss. A multi-focal high-performance vision system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2006)*, pages 150–155, Orlando, USA, 2006.

[72] **K. Kühnlenz** and M. Buss. Multi-focal switching visual servoing. *The International Journal of Robotics Research*. submitted.

[73] **K. Kühnlenz** and M. Buss. Multi-focal visual servoing strategies. In G. Obinata and A. Dutta, editors, *Vision Systems*. to appear.

[74] **K. Kühnlenz** and M. Buss. Stability issues in sensor switching visual servoing. In *Invited Session on Visual Servoing, Proceedings of the SPIE International Symposium on Optomechatronic Technologies*. to appear.

[75] **K. Kühnlenz** and M. Buss. Towards an emotion core based on a hidden Markov model. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2004)*, pages 119–124, 2004.

[76] **K. Kühnlenz** and M. Buss. Towards multi-focal visual servoing. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, pages 2336–2341, Edmonton, Canada, 2005.

[77] **K. Kühnlenz** and M. Buss. A multi-camera view stabilization strategy. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, pages 5303–5308, 2006.

[78] **K. Kühnlenz**, G. Lidoris, D. Wollherr, and M. Buss. On foveated gaze control and combined gaze and locomotion planning. In M. Hackel, editor, *Humanoid Robots*. to appear.

[79] **K. Kühnlenz**, S. Sosnowski, and M. Buss. Evaluating emotion expressing robots in affective space. In *Human-Robot Interaction*. to appear.

[80] Yasuo Kuniyoshi, Noboyuki Kita, Sebastien Rougeaux, and Takashi Suehiro. Active stereo vision system with foveated wide angle lenses. In *ACCV*, pages 191–200, 1995.

[81] V. Kyrki, D. Kragic, and H. Christensen. Measurement errors in visual servoing. *Robotics and Autonomous Systems*. in press.

[82] D. Liberzon. *Switching in Systems and Control*. Birkhauser, Boston, MA, USA, 2003.

[83] G. Lidoris, **K. Kühnlenz**, D. Wollherr, and M. Buss. Information-based gaze direction planning algorithm. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS 2006)*, pages 302–307, 2006.

[84] T. Lindeberg and J. Goarding. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing*, 15(6):415–434, 1997.

[85] V. Lipiello, B. Siciliano, and L. Villani. Eye-in-hand/eye-to-hand multi-camera visual servoing. In *Proceedings of the IEEE International Conference on Decision and Control (CDC 2005)*, pages 5354–5359, 2005.

[86] O. Lorch. *Beiträge zur visuellen Führung zweibeiniger Laufroboter in einem strukturierten Szenario*. PhD thesis, Institute of Automatic Control Engineering (LSR), TU-München, Munich, Germany, June 2003.

[87] E. Malis. Visual servoing invariant to changes in camera intrinsic parameters. In *Proceedings of the 8th International Conference on Computer Vision (ICCV 2001)*, pages 704–709, 2001.

[88] E. Malis and S. Benhimane. Vision-based control with respect to planar and non-planar objects using a zooming camera. In *Proceedings of the IEEE International Conference on Advanced Robotics*, 2003.

[89] E. Malis, F. Chaumette, and S. Boudet. 2 1/2 d visual servoing. *IEEE Transactions on Robotics and Automation*, 15(2):234–246, 1999.

[90] E. Malis, F. Chaumette, and S. Boudet. Multi-cameras visual servoing. In *IEEE International Conference on Robotics and Automation (ICRA 2000)*, pages 3183–3188, San Francisco, April 2000.

[91] D. Marr and S. Ullmann. Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London, Series B*, 211:151–180, 1981.

[92] N. Maru, H. Kase, S. Yamada, A. Nishikawa, and F. Miyazaki. Manipulator control by using servoing with the stereo vision. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots ans Systems (IROS 1993)*, pages 1866–1870, 1993.

[93] M. Maurer, R. Behringer, S. Furst, F. Thomanek, and E.D. Dickmanns. A compact vision system for road vehicle guidance. In *13th International Conference on Pattern Recognition (ICPR 1996)*, 1996.

[94] G. Metta. An Attentional System for a Hunamoid Robot Exploiting Space Variant Vision. In *Proceedings of IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS2001)*, Tokyo, Japan, November 2001.

[95] J. Miura, T. Kanda, S. Nakatani, and Y. Shirai. An active vision system for on-line traffic sign recognition. *IEICE Transactions on Information and Systems*, E85-D(11):1784–1792, 2002.

[96] K. Morooka and H. Nagahashi. A method for integrating range images with different resolutions for 3-d model construction. pages 3070–3075, 2006.

[97] A. Muis and K. Ohnishi. Eye-to-hand approach on eye-in-hand configuration within real-time visual servoing. In *Proceedings of the IEEE International Workshop on Advanced Motion Control*, pages 647–652, 2004.

[98] H.J. Müller. *The effect of selective spatial attention on peripheral discrimination thresholds*. PhD thesis, University of Durham, Durham, UK, 1986.

[99] B. Nabbe and M. Hebert. Toward practical cooperative stereo for robotic colonies. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2002)*, pages 3328–3335, May 2002.

[100] S. Nayar, S. Nene, and H. Murase. Subspace methods for robot vision. CUCS-06-95, Technical Report, Department of Computer Science, Columbia University, New York, 1995.

[101] B. Nelson and P. Khosla. An extendable framework for expectation-based visual servoing using environment models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 1995)*, pages 184–189, 1995.

[102] B. Nelson and P.K. Khosla. The resolvability ellipsoid for visually guided manipulation. Technical Report CMU-RI-TR-93-28, The Robotics Institute, Carnegie Mellon University, 1993.

[103] K. Nickels and S. Hutchinson. Weighting observations: The use of kinematic models in object tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 1998)*, pages 1677–1682, 1998.

[104] M.I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25, 1980.

[105] J. Pretlove and G. Parker. The development of a real-time stereo-vision system to aid robot guidance in carrying out a typical manufactoring task. In *Proceedings of the International Symposium of Robotics Research (ISRR 1991)*, pages 1–23, 1991.

[106] A. Rieder. Trinocular divergent stereo vision. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 1, pages 859–863, 1996.

[107] D.A. Robinson. The oculomotor control system: A review. *Proceedings of the IEEE*, 56:1032–1049, 1968.

[108] H. P. Rotstein and E. Rivlin. Optimal servoing for active foveated vision. *Computer Vision and Pattern Recognition*, pages 177–182, 1996.

[109] A. Ruf and R. Horaud. Vision-based guidance and control of robots in projective space. In *Proceedings of the 6th European Conference on Computer Vision (ECCV 2000)*, pages 50–66, 2000.

[110] D. Sagi and B. Julesz. Enhanced detection in the aperture of focal attention during simple discrimination tasks. *Nature*, 321(12):693–695, 1986.

[111] B. Scasselati. A Binocular, Foveated Active Vision System. MIT AI Memo 1628, March 1998.

[112] B. Scasselati. Eye finding via face detection for a foveated, active vision system. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI 1998)*, pages 969–976, Madison, WI, USA, 1998.

[113] C. Scheering and B. Kersting. Uncalibrated hand-eye coordination with a redundant camera system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 1998)*, pages 2953–2958, 1998.

[114] J. Schiehlen and E.D. Dickmanns. Design and control of a camera platform for machine vision. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 1994)*, pages 2058–2063, Neubiberg, Germany, 1994.

[115] C. Schmid and R. Mohr. Local gray-value invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(5), 1997.

[116] A. Shademan and F. Janabi-Sharifi. Using scale-invariant feature points in visual servoing. In S. Kaneko, H. Cho, G. K. Knopf, and R. Tutsch, editors, *Optomechatronic Sensors, Actuators, and Control. Edited by Moon, Kee S. Proceedings of the SPIE, Vol. 5603*, pages 63–70, 2004.

[117] O. Shakernia, R. Vidal, C. Sharp, Y. Ma, and S. Sastry. Multiple view motion estimation and control for landing an unmanned aerial vehicle. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA 2002)*, pages 2793–2798, 2002.

[118] R. Sharma and S. Hutchinson. Motion perceptibility and its application to active vision-based servo control. *IEEE Transactions on Robotics and Automation*, 13(4), 1997.

[119] T. Shibata, S. Vijayakumar, J. Conradt, and S. Schaal. Biomimetic oculomotor control. *Adaptive Behavior*, 9(3–4):189–207, 2001.

[120] T. Shibata, S. Vijayakumar, J. Conradt, and S. Schaal. Humanoid Oculomotor Control Based on Concepts of Computational Neuroscience. In *Proceedings of IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS2001)*, Tokyo, Japan, November 2001.

[121] S. Sosnowski, A. Bittermann, **K. Kühnlenz**, and M. Buss. Design and evaluation of emotion-display EDDIE. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, pages 3113–3118, 2006.

[122] S. Sosnowski, **K. Kühnlenz**, and M. Buss. EDDIE – an emotion display with dynamic intuitive expressions. In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2006)*, pages 569–574, 2006.

[123] N. Sprague and D. Ballard. Eye movements for reward maximization. *Advances in Neural Information Processing Systems (NIPS)*, 16, 2003.

[124] S. Thrun, Y. Liu, D. Koller, A.Y. Ng, Z. Ghahramani, and H. Durrant-Whyte. Simultaneous localization and mapping with sparse extended information filters. *International Journal of Robotics Research*, 23(7–8):693–716, 2004.

[125] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, RA-3(4):323–344, August 1987.

[126] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, and F. Nuflo. Modeling visual attention. *Artificial Intelligence*, 78(1–2), 1995.

[127] A. Ude, C. Gaskett, and G. Cheng. Foveated vision systems with two cameras per eye. In *Proc. of the 2006 IEEE Int. Conf. on Robotics and Automation (ICRA 2006)*, Orlando, USA, 2006.

[128] S. Ullmann. Visual routines. *Cognition*, 18:97–159, 1984.

[129] J. van der Spiegel, G. Kreider, C. Claeys, I. Debusschere, G. Sandini, P. Dario, F. Fantini, P. Bellutti, and G. Soncini. *A foveated retina-like sensor using CCD technology*. DeKluwer, Boston, MA, USA, 1989.

[130] T. Vidal-Calleja, A.J. Davison, J. Andrade-Cetto, and D.W. Murray. Active control for single camera SLAM. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2006)*, 2006.

[131] S. Vijayakumar, M. Inoue, and A.D. Souza. Maveric – oculomotor experimental vision head. http://homepages.inf.ed.ac.uk/svijayak/projects/maveric/index.html, 2004.

[132] R. Wallace, P. Ong, B. Bederson, and E. Schwartz. Space variant image processing. Technical Report TR1991-589-R256, 1991.

[133] L. Weiss, A. Sanderson, and C. Neumann. Dynamic visual servo control of robots: An adaptive image-based approach. *IEEE Journal on Robotics and Automation*, 3(5):404–417, 1987.

[134] G. Westheimer. Mechanism of saccadic eye movements. *A.M.A. Arch. Ophthalmol.*, 52:710–724, 1954.

[135] W. Wilson, C.W. Hulls, and G. Bell. Relative end-effector control using cartesian position-based visual servoing. *IEEE Transactions on Robotics and Automation*, 12(5):684–696, 1996.

[136] J. Wolfe. *Visual search: A review, Attention*. University College London Press, London, UK, 1996.

[137] A.L. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967.