Karlsruhe Institute of Technology

Fakultät für **Informatik**

Memory-Based Active Visual Search for Humanoid Robots

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

der Fakultät für Informatik

des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Kai Welke

aus Konstanz

Tag der mündlichen Prüfung: 03.06.2011
Erster Gutachter: Prof. Dr.-Ing. Rüdiger Dillmann
Zweiter Gutachter: Prof. Dr.-Ing. Aleš Ude

to Basia & Sebastian

Acknowledgement

This thesis was carried out in the course of my employment as research assistant at the Humanoids and Intelligence Systems Lab (HIS) of the Institute of Anthropomatics (IFA), Karlsruhe Institute of Technology (KIT).

First of all I want to thank my doctoral supervisor Prof. Dr.-Ing. Rüdiger Dillmann for giving me the opportunity to work on this fascinating topic. I want to thank Prof. Dillmann for his valuable advice and his support during the last years. The focus of his group, the facilities, and the environment made it possible to combine my scientific efforts in humanoid visual perception and their realization on real humanoid platforms. I also want to thank Prof. Dr.-Ing. Ales Ude for his interest in my work, for inspiring discussions, and for joining the committee as co-supervisor. I would particularly like to thank Dr.-Ing. Tamim Asfour, leader of the humanoids group, for his commitment to the group, his faith, and his support throughout the years.

Further, I want to express my gratitude to Dr. Mitsuo Kawato, head of the Computational Neuroscience Laboratories (CNS) of the Advanced Telecommunications Research Institute International (ATR), for the opportunity to work in his lab in 2005 / 2006. Especially, I want to thank Prof. Dr. Gordon Cheng for the supervision during my stay and the opportunity to investigate machine vision from the computational neuroscience perspective.

I always enjoyed the time with my colleagues in the humanoids group and want to thank them for the excellent teamwork and the great atmosphere. Especially, I want to thank my friend and officemate Dr. Nikolaus Vahrenkamp for his constant support and the great time. I owe many thanks to Dr. Pedram Azad for his way of supervising my diploma thesis, his faith in my skills, and the always fruitful exchange on computer vision methods. Furthermore, I am very grateful to Martin Do and Christian Böge for their support in profession and in private. My thank also goes to all the other colleagues of the humanoids group: Markus Przybylski, Julian Schill, Paul Holz, David Gonzalez, Stefan Ulbrich, Ömer Terlemez, Manfred Kröhnert, and Sebastian Schulz. I also want to thank my former colleagues Dr. Alexander Bierbaum, Steven Wieland, and Stefan Gärtner and my colleagues of the medicine group, the programming by demonstration group, and the cognitive cars group. Further, I want to thank our secretaries Christine Brand, Diane Krüger, and Isabelle Wappler for their help and the good teamwork. To all of my students I owe appreciation for their interest in the topic, their help, and their good work. Especially, I want to thank Jan Issac and David Schiebener, who have been backing me up for the last years.

Finally, I want to thank my parents for their endless support and most importantly my partner Barbara and my son Sebastian for their patience and their love.

Zusammenfassung

Die Vision der humanoiden Robotik besteht in der Bereitstellung von anthropomorphen autonomen Robotersystemen, die den Menschen in seinem täglichen Umfeld unterstützen. Um Serviceaufgaben in einer für den Menschen maßgeschneiderten Umgebung zu erfüllen, werden humanoide Roboter mit an den Menschen angelehnten Fähigkeiten zur Wahrnehmung und Aktionsausführung ausgestattet. Dabei wird die visuelle Wahrnehmung in der Regel mittels aktiver Kamerasysteme realisiert, welche die Erweiterung des Gesichtsfeldes durch Augenbewegungen ermöglichen. Des Weiteren wird die hochauflösende Fovea des menschlichen Auges mittels sogenannter fovealer Kamerasysteme nachempfunden.

Der Kopf des humanoiden Roboters ARMAR-III verfügt über ein solches aktives foveales Kamerasystem, bestehend aus einem peripheren Stereokamerapaar mit weitem Öffnungswinkel und einem fovealen Stereokamerapaar zur detaillierten Untersuchung von ausgewählten Bereichen der Szene. Die visuelle Wahrnehmung im erweiterten Gesichtsfeld des Roboters erfolgt dabei mittels der Ausführung von Blickrichtungswechselbewegungen der Augen, sogenannten Sakkaden, die jeweils in der Fixation von Bereichen der Umgebung im fovealen Kamerapaar resultieren.

Eine Grundvoraussetzung für viele Tätigkeiten im Umfeld des Menschen ist die visuelle Wahrnehmung von Objekten. Ausführungsrelevante Objekte müssen detektiert und lokalisiert werden. Der Vorgang der Detektion eines gesuchten Objektes in der Umgebung wird nach dem Vorbild des Menschen als visuelle Suche bezeichnet. Das aktive foveale Kamerasystem des humanoiden Roboters ARMAR-III erlaubt die Suche von Objektinstanzen im erweiterten Gesichtsfeld durch sakkadische Augenbewegungen. Die während dieser aktiven visuellen Suche wahrgenommene Umgebung setzt sich dabei aus unterschiedlichen Ausschnitten der Szene zusammen. Die Menge der erfassten Beobachtungen bildet die Basis für eine visuelle Repräsentation der Umgebung bezüglich des gesuchten Objektes.

Im Rahmen dieser Dissertation wurden Verfahren zur speicherbasierten aktiven visuellen Suche für humanoide Roboter untersucht, implementiert und evaluiert. Dazu wurde ein Ansatz verfolgt, der als Ziel die Bereitstellung einer konsistenten Repräsentation von gesuchten Objekten in der Umgebung des Roboters definiert. Um die Konsistenz der Repräsentation zu wahren, wurde eine Strategie zur Erzeugung von Sakkaden vorgeschlagen, die bekannte Verfahren um diese Anforderung erweitert. Die Realisierung der speicherbasierten aktiven visuellen Suche erfolgte auf einem humanoiden Roboterkopf. Dabei wurde die gesamte Kette von der Objektwahrnehmung bis zur Ausführung von Sakkaden untersucht und realisiert.

Zur konsistenten Speicherung von Instanzen gesuchter Objekte wurde ein transsakkadischer Speicher vorgeschlagen. Dieser transsakkadische Speicher dient zur Akkumulation von Eigenschaften beobachteter Objektinstanzen über mehrere Sakkaden. Sowohl die Position der Instanzen als auch ihre Ähnlichkeit zum gesuchten Objekt sind Bestandteil der gespeicherten Daten. Gemäß dem Wahrnehmungsprinzip des peripheren und fovealen Sehens wurde eine hierarchische Unterteilung des Speichers in präattentive Schicht und Objekt-Schicht vorgeschlagen. Die Detektion von Objektkandidaten erfolgt in den peripheren Kameras, welche mittels eines weiten Öffnungswinkels einen großen Ausschnitt der Szene abdecken. Die Speicherung der Objektkandidaten erfolgt in der präattentiven Schicht des transsakkadischen Speichers. Dabei werden Unsicherheiten in der Wahrnehmung und der Ausführung von Sakkaden berücksichtigt. Durch Fixation der Objektkandidaten kann die Erkennung von gesuchten Objektinstanzen im detaillierten fovealen Kamerabild durchgeführt werden. Die Speicherung des Ergebnisses der Objekterkennung erfolgt in der Objekt-Schicht des transsakkadischen Speichers.

Zur Erzeugung von Sakkaden wurde eine Strategie vorgeschlagen, die auf dem transsakkadischen Speicher basiert. Zusätzlich zur Detektion von Instanzen des gesuchten Objektes wurde dabei die Konsistenz des Speichers als Anforderung aufgenommen. Dazu wurde das Maß der aktiven Salienz formuliert, welches diese Anforderungen in einem probabilistischen Modell vereint. Dabei dient die präattentive Schicht des transsakkadischen Speichers zur Detektion von relevanten Veränderungen in der Umgebung. Die Verifikation einer Instanz in der Objekt-Schicht führt zur Validierung des Speicherinhalts. Basierend auf beobachteten Veränderungen und durchgeführten Validierungen wird ein Rückschluss auf die Konsistenz des Speichers gezogen. Hierbei erlaubt das aktive Salienzmaß den Grad der Konsistenz des Speichers in die Erzeugung von Sakkaden zu integrieren.

Zur Realisierung der speicherbasierten aktiven visuellen Suche auf dem Kopf des humanoiden Roboters ARMAR-III wurden sowohl Verfahren der Objekterkennung und -detektion als auch Verfahren zur Ausführung von Sakkaden entwickelt. Die Objektdetektion in den fovealen Kamerabildern erfolgt mittels Histogramm-basierter Methoden, welche für die aktive visuelle Suche erweitert und angepasst wurden. Die Objekterkennung in den fovealen Kamerabildern erfolgt basierend auf Texturmerkmalen. Zur Ausführung der Sakkaden wurde ein Verfahren zur kinematischen Kalibrierung der Augeneinheit entwickelt.

Die Evaluation der speicherbasierten aktiven visuellen Suche erfolgte am Beispiel einer Haushaltsumgebung. In dieser Umgebung wurden 200 Suchaufgaben absolviert. Die dabei gesuchten Objekte konnten innerhalb weniger Sakkaden detektiert werden. Die Validierung der Konsistenz des transsakkadischen Speichers erfolgte mittels mehrerer Instanzen eines gesuchten Objektes in veränderlichen Szenen.

Contents

1

Introduction

Technological innovations in our society are driven by two factors: demand and current technological know-how. The research field of humanoid robotics envisions such an innovation which constitutes a feasible technological challenge while making a big impact on the society.

The areas of possible applications for humanoid robot technology range from health care and security services to cooperative work in manufacturing. Among these areas, nursery care is one of the most important due to the current demographic trend in many industrialized countries. Introducing humanoid robots into nursery care would reduce the workload of staffs, in turn improving the quality of care. An even wider area of application consists in assistance and cooperation tasks in daily-life such as housekeeping. A humanoid robot equipped with the abilities to carry out a wide range of household tasks would substantially increase the quality of life. Overall, most areas of application are based on one common idea: the transfer of daily-life, monotonous and hard labor tasks to a humanoid robot in order to allow human to concentrate on higher cognitive tasks. Due to the attractive possibilities, the implementation of humanoid robot platforms with sophisticated capabilities in such real life environments is the focus of most ongoing research.

Through the last decades humanoid platforms have been designed and made available for an ever growing research community. Current platforms such as the humanoid robot ARMAR-III [Asfour et al., 2006] are highly integrated in terms of sensors, actuators, processing resources, and power supply (see Fig. 1.1). The realization of several assistance and manipulation tasks on such platforms has been successfully demonstrated in different human-centered environments. Yet, many challenges remain before humanoid platforms will be able to serve us in daily life. These challenges include the robustness of perception and task execution, adaptivity to changing environments and generalization and representation of acquired knowledge.

Fig. 1.1. ARMAR-III ([Asfour et al., 2006]) in the kitchen. The assistance in daily-life constitutes an important application area of humanoid robots. Equipped with the abilities to carry out a wide range of household tasks such a system could substantially increase the quality of life.

1.1 Motivation and Objective

Most areas of application envisioned for service robots are situated in the vicinity of human. As a consequence, typical operating environments are of human-centered nature. All elements of such environments exhibit ergonomics tailored for human, thus posing several requirements on the design of service robots. In order to fulfill a wide range of tasks the robot system has to offer similar abilities in terms of perceiving and acting as human. This requirement is attributed for by the anthropomorphic design of humanoid robots. The anthropomorphic appearance does not only provide the appropriate abilities to perceive and act, but also increases the acceptance by allowing for intuitive and natural interaction.

Considering the visual perception abilities of humanoid robots, the anthropomorphic design principle is implemented by mimicking the human visual system in several ways. Most humanoid robots allow controlling the cameras with a set of joints of the head-eye system and thus facilitate active gaze control. This ability to control the gaze enhances the field of view and on the other hand allows to produce natural head and eye postures. Further, foveated vision systems have been proposed that mimic the space-variant resolution of the human eye with a high resolution fovea at its center. For this purpose, either cameras with space-variant resolutions or multiple integrated cameras for each eye have been deployed. Humanoid platforms that combine active gaze control and foveated vision allow selective focusing on elements in a scene. Each gaze shift results in the fixation of a fraction of the scene within

Fig. 1.2. The humanoid robot ARMAR-III is equipped with a peripheral-foveal camera system which mimics the high resolution fovea of the human eye. Left: The peripheral views provide an overview of the scene. Right: Using foveal vision the detailed views of the scene are perceived within several glances by actively moving the eyes. The inner model of the scene is derived by considering a sequence of such foveal views at different gaze directions.

the high resolution fovea, thus allowing for increased acuity and robustness of the perceptual task (see Fig. 1.2).

Real world tasks usually involve the visual perception of several elements of the environment which are spatially distributed. The knowledge of the spatial layout of objects in such scenes is a prerequisite in order to decide the appropriate sequence of actions and to acquire necessary parameters for execution such as position and orientation. Reconsidering the selective nature of active foveated vision systems as introduced above, this knowledge is acquired by performing several gaze shifts, so-called saccades. Each saccade results in a fixation which provides a distinct detailled view of the scene. By integrating the information extracted from multiple views in a common representation, an inner model of the scene can be obtained. Since only a partial state of the world is visible on each fixation, the consistency of the inner model can be affected due to changes in the world through e.g. the interference of other agents. Thus, processes for the constant validation of the inner model are required in order to accurately reflect the scene with respect to relevant visual information.

The objective of this thesis is to endow a humanoid robot equipped with an active foveated vision system with the capabilities to actively guide its gaze in order to build such an inner model of the scene. The resulting representation constitutes a consistent perceptual basis for higher cognitive functions such as scene interpretation and manipulation.

1.2 Contribution

In order to enhance the perceptual capabilities of humanoid robots, an approach for retaining a stable representation of relevant objects in the scene is presented in this thesis. This representation is accumulated over different gazes performed by actively guiding the eyes. Deriving a solution for this problem prerequisites two capabilities: to search for objects utilizing an active camera system and the establishment of a consistent representation based on the search results. The key contributions of this thesis are summarized with respect to these two capabilities as follows:

- **Active Visual Search**
  The thesis presents a new approach for an active visual search procedure in the presence of a humanoid head with active foveated vision. The approach applies object detection in the peripheral images in order to identify object candidates and object recognition in the foveal cameras. For this purpose, state of the art methods for object detection are extended in terms of robustness and detection of multiple instances. For generating sequences of saccades based on detection results, an attentional mechanism is formulated in terms of a probabilistic inference problem, which maximizes the likelihood to detect the target object. Based on the generated target location, saccade execution by the active head is facilitated using a calibrated model of the head-eye system. For this purpose, a new approach for the kinematic calibration of the active head is proposed.

- **Transsaccadic Memory**
  In this work, the active visual search problem is extended by a spatial memory which is accumulated over multiple saccades. The proposed transsaccadic memory is organized in a hierarchical fashion which is derived from the peripheral-foveal design of the active camera system. A preattentive memory layer accumulates object candidates observed with the peripheral cameras while the object memory holds the positions of object instances as recognized in the foveal cameras. In order to support the accumulation of object candidates in the preattentive memory under varying gaze directions, the correspondence problem is approached by considering uncertainties from saccade execution and object detection in a probabilistic fashion. The location of an instance in the object memory is refined in a closed-loop manner. The proposed memory-based active visual search approach extends the outcome of active visual search from a single fixation point to the content of the transsaccadic memory. Considering this memory an integral part, its consistency becomes an essential drive for the generation of saccades. In order to assure consistency, the attentional mechanism is extended by the concept of active saliency, which takes into account the volatility of the physical world.

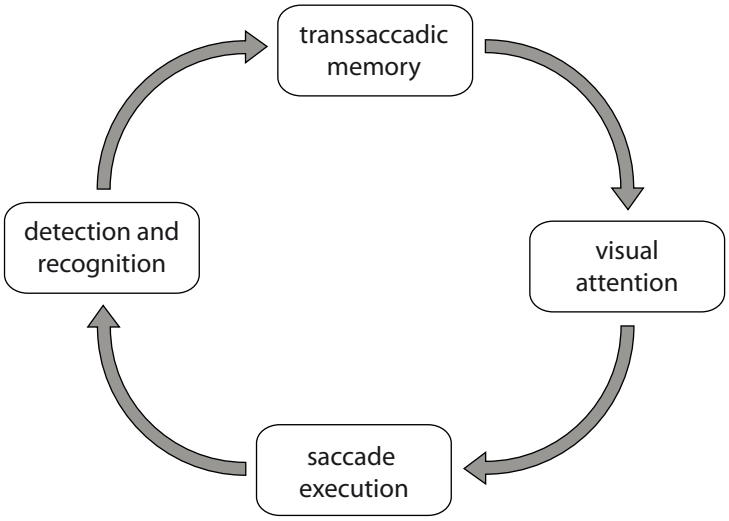

Fig. 1.3. The approach for memory-based active visual search addresses the four subproblems of object detection and recognition, saccade execution, memory, and visual attention.

The memory-based active visual search is implemented and evaluated on a humanoid head equipped with an active foveated camera system. As will be shown, the propoesed approach is able to build and retain a consistent inner model of the scene with respect to the target object even if the physical world is altered by the interference of other agents.

1.3 Outline

Human behavior and perception serves as a role model for many approaches that perform visual search in the robot literature. In order to provide an understanding of the basic terms and definitions, an overview of relevant principles of the human visual system is given in Chapter 2. The thesis proceeds with the review of recent research conducted in active visual search in Chapter 3, the focus being put on visual attention mechanisms and implementations of active visual search on anthropomorphic platforms. The overall approach for memory-based active visual search is outlined in Chapter 4. As depicted in Fig. 1.3, the subsequent four chapters address different subproblems of the memory-based active visual search problem, beginning with the developed approaches for peripheral object detection and foveal object recognition in Chapter 5. The thesis proceeds with the kinematic calibration of the head-eye

system and the execution of saccades in Chapter 6. The layout of the memory and an approach for solving the correspondence problem is discussed in Chapter 7, and taking into account the content and the consistency of transsaccadic memory, potential targets for saccadic eye movements are generated utilizing methods of visual attention which are formalized in Chapter 8. The approach for memory-based active visual search is evaluated on the humanoid head with respect to the search performance and consistency of memory retention in Chapter 9. Finally, the contributions of the thesis are summarized in Chapter 10 and future developments and applications are briefly discussed.

2

Visual Search in Humans

The term visual search has its origin in experimental psychology. A very broad but common sense explanation describes visual search tasks as those where humans look for something [Wolfe, 1996]. While this phenomenological explanation gives a good idea of what visual search stands for in human visual perception, more insight in the underlying principles is necessary in order to identify processes and derive implementations on technical systems. The following sections provide an overview of findings in the field of cognitive psychology on the topic of visual search. The overview is restricted to findings that directly relate to the state of the art on technical implementations of visual search approaches as discussed in the next chapter.

The visual search performance in humans is assessed with a class of experiments which share a common paradigm of experimental setup and procedure. According to [Tsotsos, 1990], a visual search experiment is defined as follows. Subjects in such an experiment visually perceive a test display, which shows a set of target items and a number of nontarget items (distractors). During the experiment, the time needed by the subject to detect a target among the distractors (response time) is measured. An example setup of a visual search experiment is illustrated in Fig. 2.1. The blue letter "T" constitutes the target object of the search task. Further, the display contains distractors (green letters "X" and brown letters "T"). The time required by a subject to detect the blue "T" is recorded. This response time varies in different setups of the experiment thus allowing to gain insights in the way humans process visual information. The variation in the complexity of distractors and targets allow to understand how human performance is affected by visual complexity. Increasing or decreasing the number of targets and distractors allows to correlate the response time with the set size. Further, the spatial location of targets can be varied in order to determine the dependency between viewing angle and search performance. Depending on the design of the experiment different aspects of the human visual search behavior can be probed in order to derive underlying principles.

The following sections provide an overview of fundamental principles found in human visual search. The focus is put on involved attentional mechanisms and the role of memory.

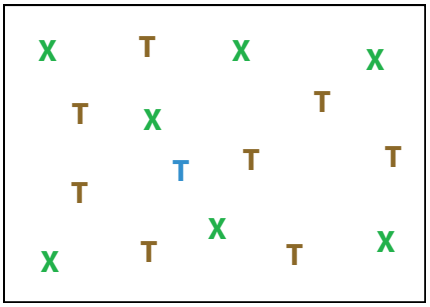

Fig. 2.1. A typical visual search experiment. The subject is supposed to detect the blue letter "T" among the distractors. The response time required for detection is recorded in order to assess human performance.

2.1 Visual Search and Attention

Natural environments offer a vast variety of visual information. Neisser et al. claimed that parallel processing of all visual information is implausible since the human brain does not possess the required resources [Neisser, 1967]. More likely, agents operating in such environments only process a small subset of information which is relevant for the current behavior. The restriction of processing to only a small relevant subset of possibilities allows to deploy the limited resources in an economic manner. In the context of visual perception this mechanism is usually referred to as visual attention.

The relation between visual attention and visual search has been demonstrated in an experiment according to the visual search experimental paradigm [Treisman et al., 1977]. The experiment is based on the variation of two parameters: the visual complexity of targets and distractors and the amount of distractors. By varying these parameters Treisman et al. could identify two different classes of visual search tasks, which show significant difference in the way the response time relates to the number of distractors. In the first class of tasks, the disjunctive tasks, the target is defined by one outstanding feature among the distractors. An example of a disjunctive search task is depicted in Fig. 2.1 where only the target is of blue color. In the second class of tasks, the conjunctive tasks, the target is outstanding only by considering a combination of features. Fig. 2.2, left illustrates a conjunctive task, where the target is defined by the features green and "T"-shaped. A series of such experiments revealed that the response time in disjunctive search tasks remains constant when varying the number of distractors while in conjunctive tasks the response time is about linear to the number of distractors (see Fig. 2.2, right).

For disjunctive search tasks, the constant response time indicates parallel processing of the display independent of the number of distractors. This only

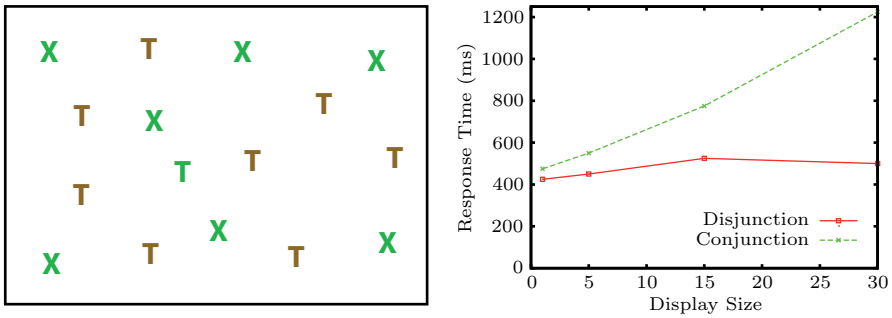

Fig. 2.2. Left: Finding the green letter "T" is a conjunctive search task. The target is defined by a combination of the two features "color" and "shape". Right: Disjunctive tasks are about constant in response time with respect to the display size, conjunctive tasks about linear (adapted from [Treisman and Gelade, 1980]).

occurs if target and distractors can be discriminated by a single feature. Based on the experiment it is possible to identify different so-called feature dimensions which can be processed in a parallel fashion. According to Treisman such feature dimensions include size [Treisman and Gelade, 1980], curvature, color, and intensity [Treisman and Gormican, 1988]. In subsequent research, the existence of more dimensions could be shown such as depth and motion [Nakayama and Silverman, 1986] and orientation [Wolfe et al., 1992]. The extraction of those feature dimensions is accomplished in parallel without the involvement of attention and is usually referred to as preattentive processing. In contrast, in conjunctive tasks where the target is defined by more then one feature the response time linear to the number of distractors indicates serial processing of the displayed items. A serial search through the display has to be performed and attention is drawn to one item at a time.

In accordance with physiological findings, the Feature Integration Theory (FIT) has been established in order to account for the human performance in conjunctive and disjunctive tasks [Treisman and Gelade, 1980]. The FIT attributes the parallel processing in disjunctive tasks to distinct specialized receptors and brain areas which coarsely correspond to different feature dimensions. In contrast, the processing of a combination of feature dimensions in conjunctive tasks involves higher areas and is performed in a serial fashion. According to the FIT, serial visual search is inseparably intertwined with the attentional mechanisms that allow to focus the processing resources to only a subset of the visual input. Within this subset different feature dimensions are bound together to a more complex representation, the so-called object file [Treisman, 1986].

2.1.1 Bottom-Up and Top-Down Attention

Focusing attention toward a location in the visual field requires a prior selection of this location. In order to understand this mechanism of selection a key question has been the type of information used for selecting one stimulus among others.

Two mechanisms have been identified which allow such a selection based on different sources of information: the bottom-up an the top-down guidance of attention. In bottom-up guidance, the selection of visual stimuli among others is performed based on the sensory information alone. Bottom-up attention is often referred to as data-driven since it only involves perceived information that is available from the current visual input. In top-down guidance, attention is focused toward a stimulus based on prior knowledge and inner models. This prior knowledge includes expectations about the scene or the objects within the scene. The top-down guidance is often referred to as task-driven since these expectations are tightly coupled to the current overall task. In contrast to bottom-up guidance, top-down mechanisms rely on information that does not reside in the stimulus such as memory representations of objects or tasks [Ullman, 1984].

In the context of visual search the question has been posed whether the search is guided mainly by bottom-up or top-down mechanisms. There is strong evidence that both, bottom-up and top-down guidance, affect visual search (see [Yantis, 2000] for a review). The interplay between bottom-up and top-down guidance in different search tasks is still subject to debate in current research. An important role of top-down guidance in visual search has been demonstrated in [Williams and Reingold, 2001] showing that distractors are fixated during search tasks which share features with the target objects. On the other hand, the effect of bottom-up guidance in visual search was demonstrated in [Theeuwes, 2004]. Strong color cues captured attention in their experiment even though they were irrelevant for the search task.

While many of the experiments that have been carried out in order to assess the influence of bottom-up and top-down guidance use simplified items as distractors and target, an experiment with photorealistic objects has been presented in [Chen and Zelinsky, 2006] (see Fig. 2.3, left). In the experiment, one outstanding object in terms of color serves as basis for the assessment of bottom-up influence. Prior to each search task a preview of the target object is shown in order to allow top-down guidance. The goal consisted in reporting a symbol embedded in the target object. The authors could demonstrate search performance independent of the color saturation used as bottom-up influence (see Fig. 2.3, right). These findings indicate that visual search in the case of complex realistic objects is dominated by top-down guidance.

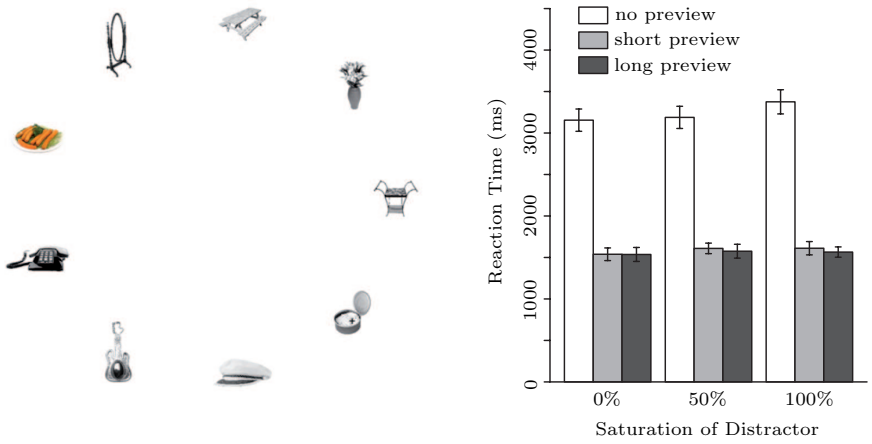

Fig. 2.3. Experiment on the influence of bottom-up and top-down guidance in photo realistic visual search tasks performed in [Chen and Zelinsky, 2006]. Left: The distractor object is displayed using colors. The target object is previewed before each trail. Right: In cases where a preview of the target object was presented, bottom-up saliency did not affect response time. Reprinted from [Chen and Zelinsky, 2006] with permission from Elsevier.

2.1.2 Covert and Overt Attention

In the previous sections, visual attention was introduced as a selective mechanism, which shifts processing resources from one location of the display to another in order to approach the complexity of search tasks. A common metaphor used to picture this kind of spatial attention is the spotlight [Posner, 1980]. According to this metaphor, the beam of the spotlight highlights an area of enhanced processing efficiency in the visual field. An extension to the spotlight metaphor has been proposed in terms of the zoom lens model, which also incorporates a variable width of the attended area [Eriksen and St. James, 1986]. The pure allocation of processing resources as considered in these models is usually referred to as covert attention. In contrast, overt attention denotes the application of eye movements in order to physically attend to elements of the environment by foveal fixation. In the context of active visual search, this guidance of the eye movements is a crutial element and consequently necessitates mechanisms of overt attention.

It seems apparent that a connection exists between actively changing focus in overt attention and covert allocation of processing resources to spatial locations. However, the strength of this dependency has been discussed controversially. The proposed theories range from complete independence between the attention system and the oculomotor system to theories, which state that both systems are identical. A series of experiments have been carried out that point in the direction of independent oculomotor and atten-

tion systems [Posner and Dehaene, 1994, Hunt and Kingstone, 2003]. Nevertheless, a growing body of literature follows the view of a strong dependency between covert and overt attention. Following the view of strong dependency, Rizzolatti et al. proposed the premotor theory of attention [Rizzolatti et al., 1987, Rizzolatti et al., 1994]. According to the premotor theory, covert attention is a by-product of the preparation for eye movements toward a target. Rizzolatti et al. suggest that attention does not result from a dedicated mechanism but is controlled by the same neural circuit that controls the motor behavior for overt focusing. Recent studies in neurophysiology lend support to the premotor theory of attention (see e.g. [Moore and Fallah, 2004, Ekstrom et al., 2008]). Furthermore, several recent studies in experimental psychology support the premotor theory (see e.g. [Van der Stigchel and Theeuwes, 2007, Awh et al., 2006]).

Overt attention involves the execution of saccades, which allow to fixate elements of the surrounding environment in the high resolution fovea. Saccadic eye movements have been well studied and several characteristic properties could be observed. Saccades are usually initiated with a latency of 100-200 ms after the onset of a stimulus [Carpenter, 1988] depending on the urgency of the perceptual task [Montagnini and Chelazzi, 2005]. The velocity of saccades can reach up to $700°/s$ [Rao et al., 1996]. Amplitude, duration and peak velocity of saccadic eye movements exhibit a strong relationship. A set of equations, the main sequence, has been established for these parameters describing their relationships in normal saccadic eye movement behavior [Carpenter, 1988]. The main sequence seems to trade-off accuracy versus speed in the presence of noise [Harris and Wolpert, 2006].

Saccadic eye movements are ballistic movements, which are controlled in an open-loop fashion. Thus, the target for the saccade has to be determined prior to saccade execution [Miles, 1983]. A model of the oculomotor system is required in order to map target locations to motor codes. In order to gain insights in the mapping model used by human, the accuracy of the eye landing position in relation to the target has been subject to extensive research. In general, saccades undershoot and a second corrective saccade is executed in order to allow fixation of the target [Becker, 1989]. Depending on the task, the accuracy can vary, leading to highly accurate landing with a standard deviation of less then 0.5° on the target in simple search displays [Findlay and Brown, 2006]. Inaccuracies of saccades occur in cases where another target interferes. If two targets are in proximity the saccade lands near the center of gravity of both targets [McGowan et al., 1998].

During the execution of a saccade, perception is suppressed. This saccadic suppression has first been reported in [Dodge, 1900] for the ability of flash detection. Later studies show, that also displacement is not perceivable during saccades [Bridgeman et al., 1975].

2.2 Visual Search and Memory

Overt and covert attention provide powerful mechanisms to select only relevant visual information and thus to reduce the computational load of the visual system. Based on those mechanisms, the visual world is perceived only through distinct fixations. Nevertheless, humans have the impression of a consistent and detailed world, in which everything is present at once. We are able to build an internal representation of the scene "across separate glances and over time" [Melcher, 2001]. The memories and representations involved in visual search that allow to build such an internal representation are discussed in the following sections.

Visual information perceived at a fixation point is first stored in the sensory memory [Coltheart, 1980]. Information in sensory memory resides in a very low level of abstraction, almost image-like, and is retained for about 80-100 ms. In the three level architecture of memory [Atkinson and Shiffrin, 1968] the sensory memory is connected to the visual short-term memory (VSTM), which has a higher level of abstraction, limited capacity, and is subject to much slower decay. The highest level of abstraction is provided by the long term memory (LTM) which accumulates the knowledge over long periods of time.

2.2.1 Models of Memory and Attention

Two prominent models for the representation and memorization of visual input during visual search have been proposed: the Object File Theory of Transsaccadic Memory [Irwin, 1992] and the Coherence Theory [Rensink, 2000]. In the following, both models are briefly introduced and their relation to visual search and attention is discussed.

According to the Feature Integration Theory (FIT, see Section 2.1), attention binds low-level features into a more complex representation, the object file. Based on this representation, a memory which retains object files for several fixations is proposed in the Object File Theory of Transsaccadic Memory [Irwin, 1992, Irwin and Andrews, 1996]. The transsaccadic memory is assumed to reside in the VSTM which is of limited capacity.

The Coherence Theory, on the other hand, models the interplay between attention, representation, and memorization [Rensink, 2000, Rensink, 2002]. According to the Coherence Theory, proto-objects are generated in the preattentive phase in a bottom-up fashion independent of attention. These proto-objects are compounds of structures generated in parallel over the visual field and are highly volatile. The state of volatileness of proto-objects is referred to as state of limited coherence. On allocation of attention, proto-objects are bound together in a spatial and temporal coherent representation called nexus (see Fig. 2.4). The nexus is linked to its corresponding proto-objects through

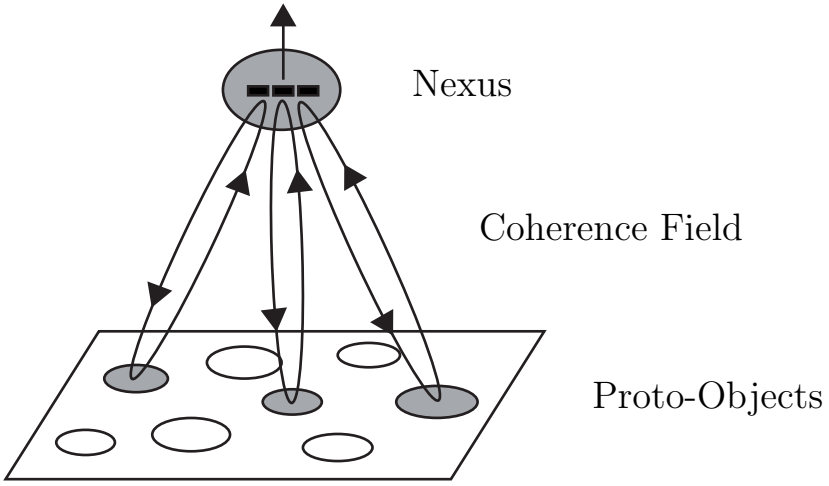

Fig. 2.4. In the Coherence Theory, proto-objects are generated in parallel and without attention over the visual field. If attention is focused on a set of proto-objects a temporary VSTM representation - the nexus - is generated and associated to the proto-objects via the coherence field. Illustration according to [Rensink, 2002].

the coherence field which forms a bidirectional communication channel. The up-link allows the nexus to obtain descriptions of visual properties from the attended proto-objects. The link from nexus to proto-objects guarantees temporal stability of the proto-objects e.g. in the case of occlusion. After removal of attention, the coherence field and the nexus are released. The proto-objects become unbound leaving them in the state of limited coherence again. In the view of Coherence Theory only one nexus can be formed at a time by focusing attention to a set of proto-objects. The nexus including the coherence field is proposed as VSTM representation. The Coherence Theory predicts that once attention is withdrawn and the nexus is released, no visual information is retained in the VSTM.

2.2.2 Persistence of Memory

The persistence of object representations in VSTM is discussed very controversially. The debate covers two main aspects: the duration of persistence and the level of detailedness of preserved information [Horowitz, 2006].

One line of research on the persistence of representations is based on the phenomenon of change blindness in human visual perception. This phenomenon was investigated in a series of experiments, where real world photographs were presented to subjects [Rensink et al., 1997]. In each trial, a part of the image was changed (e.g. engine removed from airplane). The experiment showed

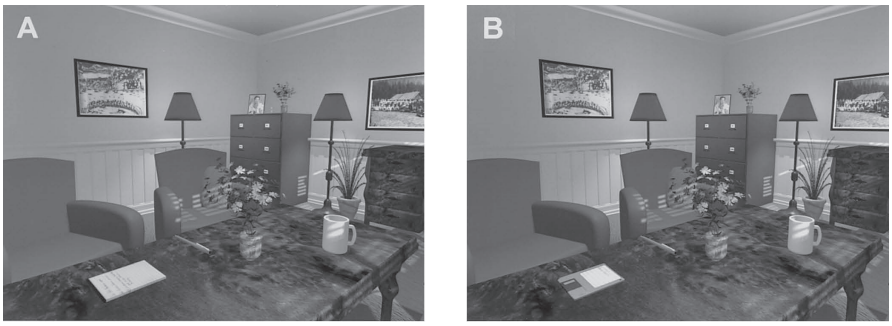

Fig. 2.5. Typical pair of images in the change detection experiment performed in [Hollingworth and Henderson, 2002]. In this item change condition, the previously fixated target object is changed in order to assess the persistence and detailedness of transsaccadic memory. Reprinted from [Hollingworth and Henderson, 2002] with permission from APA.

that subjects were unable to respond quickly to the changes. This observation lead to the conclusion that very little visual information about natural scenes is retained in memory. As consequence, the visual transience hypothesis was established stating that visual information acquired during one fixation is highly volatile. Both, the Coherence Theory and the Object File Theory of Transsaccadic Memory follow in principle the visual transience hypothesis.

In a more recent experiment, different levels of detailedness of representations were included in a change detection experiment [Hollingworth and Henderson, 2002]. During each trial one object of a computer rendered scene was changed (see Fig. 2.5). In contrast to the previous experiment it was guaranteed that subjects fixated the target object prior to change. Three types of changes were used in the experiment: type change, token change and rotation. In the type change setup, objects in the scene were replaced with objects of a different type (e.g. notepad with disk). In token change conditions, objects were replaced with other objects from the same family (notepad with different notepad). Rotation of the target object was performed in the rotation condition. Based on this experimental setup it could be shown that subjects were able to reliably detect the change for all three conditions. The authors conclude that visual representations at high detailedness are retained after withdrawal of attention and stored over several saccades. This result opposes the visual transience hypothesis.

The effects demonstrated by Hollingworth and Henderson lead to the model of long-term memory object files, which can be retained longer and are not bound to the limited capacity of VSTM. LTM object files are counterparts to the VSTM object files and code spatial locations and concepts associated with the VSTM object file. The LTM object files are accumulated in LTM and form

the basis for persistent scene representations. This model does not contradict the transience of VSTM as predicted in previous theories. Rather the model extends those theories by introducing LTM representations in terms of the LTM object files. Support was lend to the model of Hollingworth by other recent research which also reported the persistence of object files over several fixations and over long periods of time [Mitroff et al., 2005, Noles et al., 2005].

2.3 Discussion

In this chapter, basic principles and findings from the field of cognitive psychology in the context of active visual search were introduced. Many of these principles find their counterpart in implementations of active visual search on technical systems as discussed in Chapter 3. In the following, the most important principles are summarized.

Visual search for complex compounds of features involves focused attention. Thereby, attention serially allocates processing resources to distinct locations in the scene. The guidance of the attentional focus is performed using bottom-up data-driven mechanisms and top-down task-driven mechanisms. While both mechanisms are employed during visual search the search for complex objects seems to be dominated by top-down guidance. The shift of attention involving eye movements is referred to as overt attention. Different theories exist about the connection between pure allocation of processing resource in covert attention and overt attention. The premotor theory of attention predicts that overt and covert attention share the same mechanisms.

By overtly attending to different locations the surrounding visual world is perceived with distinct glances. In order to explain the impression of stability in human visual perception, the model of transsaccadic memory has been proposed. The observations of visual transience predicts that the content of transsaccadic memory can only be retained over a short period of time. Contradicting the visual transience hypothesis, a detailed representation retained over long periods was observed in a series of experiments. An attempt to bring both sides together uses the notion of LTM object files. The extension of the FIT by the introduction of LTM object files provides a way to explain persistence of even detailed representations without denying the transience of VSTM.

3

Active Visual Search on Technical Systems

In the following sections, research conducted in the context of active visual search on technical systems is reviewed. First, the related research fields of active vision and visual search are introduced and consequences for the proposed approach are discussed in Section 3.1. Thereby, the necessity of visual attention for approaches that perform active visual search is motivated for technical systems. Important classes of methods that have been applied to implement visual attention are reviewed and compared in Section 3.2. Subsequently, approaches for active visual search on technical systems that have been proposed in the literature are presented and discussed in Section 3.3. Therefore, the focus is put on implementations on platforms which provide the ability to actively look. Approaches which make use of foveated vision and consider a spatial memory during active visual search are highlighted. Finally, the discussed approaches are summarized in Section 3.4 and related to the problem addressed in this work.

3.1 Related Research Fields

In the following, the focus is put on research fields which are closely associated with the problem addressed in this work. Fundamental research will be highlighted that provides the basic principles drawn on in order to solve the problem of active visual search. As already reflected by the title of this thesis and as set out in the motivation and objective, the problem at hand consists in solving the visual search for objects using an active vision system. Consequently, the research field of visual search and its relation to machine vision and the problem of active vision are outlined in the following and the consequences for the proposed approach are discussed.

Machine Vision and Visual Search

The term visual search has been introduced in the previous chapter by the definition of the visual search experiment as used in experimental psychology. In its most general form, the problem of visual search consists in identifying target objects among a spatially distributed set of non-target objects and target objects, where the objects may vary in position, scaling, and orientation. Furthermore, the perception can be subject to perturbation, resulting in incomplete or noisy display.

While the term visual search is widely used in the context of computational approaches toward human perception and the implementation of similar skills on humanoid platforms, related problems have been addressed in the machine vision field during the last decades. In machine vision, the problem of visual object perception is usually subdivided into a set of more specific subproblems in order to cope with the diversity of possible perceptual tasks. A coarse and commonly used subdivision proposes the distinction of object detection and object recognition [Amit, 2002, Bishop, 2007]. In both cases, a model of the target object is assumed, which is usually geometric or appearance-based or a combination of both. In the case of object detection, the problem consists in identifying the location of the target object in the query image, or a registration between model parts and image parts. Typical object detection tasks involve query images, which include several objects, target as well as non-target objects. Object recognition in contrast refers to the classification of an object as target or non-target object. A typical query image for object recognition contains only a single object which is subject to classification. A third subclass of object perception problems is referred to as object classification. Object classification covers the case of assigning an object to one of multiple categories.

In order to solve the general visual search problem as defined above, a solution to at least two subproblems of object perception, namely object detection and recognition, has to be derived. Further, the term visual search implies the commitment to specific underlying processing principles deployed to address the problem. These principles have been derived in [Tsotsos, 1990] and are discussed in the following.

Tsotsos approaches the problem of visual perception from the complexity analysis perspective. Given the ability of the human visual system, he asks what might be the computational mechanisms that allow to implement these abilities on a system of limited resources such as the human brain or an artificial system. Therefore, the visual search tasks were divided into two classes: the bounded and the unbounded search tasks. In bounded visual search tasks, the target of the search is known a-priori and the goal is to detect the target among distractor objects. The unbounded task corresponds to the odd-man-out problem, where the goal consists in finding the one object, that does not

fit to the other objects in the display. Tsotsos showed, that the unbounded visual search problem can be mapped to the Knapsack Problem, which is known to be NP-complete [Tsotsos, 1989]. As a consequence, the complexity of the unbounded visual search is exponential in the number of pixels and as such computationally intractable. According to Tsotsos, an implementation of the ability to perform unbounded visual search has to satisfy these complexity constraints. Knowing the problem is NP-complete, principles have to be defined which allow the execution of visual search on systems with limited resources such as the human brain or artificial systems. From an analysis of the resources provided by the human brain, Tsotsos derived architectural principles including parallelism and hierarchical organization which allow the solution to unbounded visual search tasks under complexity constraints. By additionally applying the minimum cost principle to the problem, further characteristics are predicted such as columnar organization of processors, inverse magnitude of processing layers, and inhibitory attentional mechanisms. The insight provided by Tsotsos analysis of the visual search problem strongly supports the use of attentional mechanisms in visual search tasks. While these mechanisms have been proposed and deployed in different computational models, the complexity analysis shows the necessity and thus motivates the use of attentional mechanisms in models and implementations of visual search.

Active Vision

The research field of active vision has been initiated in the late eighties by the research of Bajcsy and Aloimonos et al. [Bajcsy, 1988, Aloimonos et al., 1988]. Machine vision research until that time was in its majority influenced by Marr's approach that formulated vision as a complex information processing task [Marr, 1982]. Marr's processing chain starts with the retinal projection of the scene, from which in subsequent steps the 3D model of the world is estimated. In artificial vision systems, the retina was usually implemented with a passive camera, which provided the input for the processing chain. Compared to how humans perceive their world these passive systems lack the ability of adapting parameters of the visual system in order to actively look.

In active vision approaches, several parameters of the camera system are considered and adapted in an active manner. Adaptation to illumination is achieved using a controlled aperture. Combined with an adaptable focus, the depth of sharpness can be controlled. For binocular systems the ability to converge or diverge and the realization of head movements provide parameters which allow to adapt the view point of the camera system. The major focus of early work in the active vision field was put on the question if and how the adaptation of these parameters can be beneficial in solving relevant problems of computer vision.

In [Aloimonos et al., 1988], the purposeful adaptation of the parameters for different computer vision problems is investigated. The motivation for the application of active vision stems from the fact that many computer vision problems are ill-posed given a passive observer and thus do not offer a unique solution. In order to restrict the solution space and to derive a unique solution additional constraints are usually defined. While many of these constraints are plausible, such as the smoothness of surfaces, they are not valid in general and lead to deficient results. A second problem of passive approaches consists in their instability in the presence of noise or errors in the input. In their work, Aloimonos et al. address the problems of shape from shading, shape from contour, shape from texture, and structure from motion which are ill-posed or unstable given a passive observer. For all four problems, a solution based on active vision is proposed where the position or orientation of an active observer is adapted in a way that the problem becomes well-posed and stable. For most problems the knowledge of the trajectory in parameter space allows to relax the constraints and to solve the ill-posed problem.

A different view of active vision has been proposed in [Bajcsy, 1988]. Again, active vision is motivated by abilities of the human visual system. The term "active" is derived from the active sensing domain and is used in order to describe a passive sensor which is used in an active fashion. In contrast to Aloimonos the notion of active vision emphasizes modeling and control strategies for perception. The processing chain of visual perception is decomposed into local models, comprising models for sensors and signal processing mechanisms including their parameters and resulting uncertainties. In order to solve perceptual tasks based on the local models, a global model is defined which describes the interconnection and external parameters of the system. In active vision approaches, this global model includes feedback which allows to adapt the parameters of the local models. Thereby, not only the parameters of the active camera system are considered but also the parameters of signal processing and data reduction mechanisms. The strength of this concept is demonstrated in focusing an unknown scene using feedback to adapt aperture and zoom, and zero disparity to adapt the vergence angle of the active camera system on the sensory level. Further, on higher processing levels, processes such as matching with object models or perceptual grouping allow to tune the system and provide feedback for prior processing. On the highest level, a search of information approach is proposed which involves decision making in order to choose parameters which are suitable with respect to the perceptual goal. Overall, this view on active vision sketches an intelligent data acquisition process with inherent feedback that adapts the parameters of all local models including the sensor.

Based on this pioneering work, the definition of active vision evolved and was further extended in subsequent research. Most notably Ballard introduced the concept of sensori-motor behaviors, which are task dependent and compete for the visuo-motor system [Ballard, 1991]. The task dependency and the defini-

tion of such behaviors led to the notion of purposive vision, where the problem of vision is considered with respect to the embodiment [Aloimonos, 1993]. The human visual system and implementations on anthropomorphic systems share the same purpose and should have similarities in computational methods, representations, and architecture. The proposed principles in active vision define such similarities. Further, in the view of purposive vision, the visual system is not isolated and composed of different modules. Rather, the visual system is embedded in a larger system with an overall task, which provides different embodied behaviors. These embodied behaviors together constitute the ability of visual perception.

Consequences for the approach

The approach for active visual search proposed in the remainder of this thesis builds on insights gained in the field of visual search and active vision. The problem of finding an object in a complex scene ultimatively involves classical machine vision problems such as object detection and recognition. The complexity analysis performed by Tsotsos reveals, that those problems have to be addressed in a manner, which takes into account the complexity constraints of artificial system. The application of visual attention as well as the use of hierarchical representation has been put forward in order to cope with the computational complexity of the visual search problem and builds the basis for solving the active visual object search problem in this work.

The anthropomorphic head of a humanoid robot allows to adapt several parameters, which influence gaze direction and head pose. For this work, the ability of the active camera system is exploited in order to extend the visual field and to provide more natural ways of human machine interaction. In the sense of Bajscy, the approach generates feedback to the head controllers in order to accomplish the overall task in an active vision fashion. The perceptual task is defined by the target object; this information is used within an action-perception cycle in order to provide a consistent answer to the active visual search problem.

3.2 Visual Attention

The importance of attention for visual search has already been discussed in the context of the complexity analysis. The selective nature of attention allows to focus on relevant information in order to cope with the complexity constraints. As discussed in Section 2.1, research on attention has been performed in the psychophysical domain in order to understand how human perception copes with the diversity of the visual world. In the psychophysical domain, the goal usually consists in deriving models of attention which allow to predict human performance.

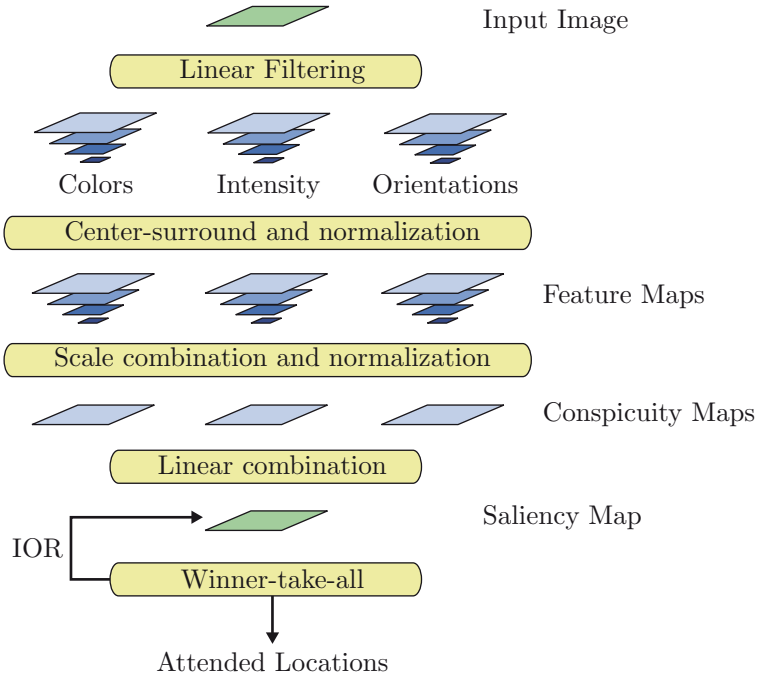

Fig. 3.1. Model of visual attention according to [Itti et al., 1998]. The saliency map is calculated from the input image in several stages including calculation of scale pyramids, feature calculation and combination. A winner-take-all network determined the location of maximum saliency and inhibition of return (IOR) ensures the sequencing.

In the following the focus is put on technical implementations of visual attention. The common goal of all discussed approaches consists in the calculation of saliency data from the visual input. The saliency data is usually hold in a two dimensional saliency map, where each element corresponds to a location in the visual input as introduced in [Koch and Ullman, 1985]. Thus, the saliency map is a topological map which encodes a measure of conspicuity. According to the different theoretical frameworks used for the calculation of saliency, approaches of visual attention for technical systems can be roughly categorized in the three classes of psychophysical approaches, information theoretic approaches, and Bayesian approaches. Relevant approaches from these three categories are reviewed and discussed in the following paragraphs.

3.2.1 Psychophysical Approaches

The class of psychophysical approaches are largely influenced by biological findings and provide models for the human ability in terms of visual attention. Here approaches are of interest, which are motivated by principles derived from the human visual system and which at the same time serve a technical purpose. The basic mechanisms, which are at the core of many psychophysical approaches toward visual attention have been proposed in [Koch and Ullman, 1985]. Their computational model lead to a first implementation of visual attention on a technical system in [Itti et al., 1998], which is outlined in Fig. 3.1. The saliency map is calculated successively from the input image in four stages. In the first stage, invariance to scaling is achieved by computing a pyramid of different scales. Each of the scales is subject to feature extraction, where all features are processed in parallel. Usually features are used, which correspond to one feature dimension as defined by Treisman et al. (see Section 2.1). The original implementation proposed to use the feature channels orientation, intensity, and color. In the successive stage, the center-surround difference as proposed by Marr is calculated at different scales from the features in order to mimic similar processing principles in the human visual system [Marr, 1982]. The resulting feature maps are sensitive e.g. to transitions between different colors or to edges in the case of intensities. In order to derive one conspicuity map per feature, the different scales are combined in a single map in the third stage. The final saliency map is calculated by a linear combination of the different conspicuity maps. In addition to the model for saliency map calculation, Koch and Ullman also proposed a mechanism which allows to generate shifts of attention based on the encoded saliency data. The target for the attentional shift is defined by the maximum activation of the saliency map and is determined based on the winner-take-all principles. Therefore, the application of a dynamic neural network, the winner-take-all network (WTA, [Feldman, 1982]), is proposed. In order to approximate the calculation of the maximum over the complete saliency map with biological plausible processing principles, the WTA network includes local inhibition, which assures that the activation of a region around the winning neuron is inhibited. This mechanism is usually referred to as inhibition of return (IOR).

Many variations and extensions to the original implementation have been proposed. The largest variety exists in the different feature channels used for the saliency map calculation. The proposed features include corners [Heidemann et al., 2003, Ouerhani and Hugli, 2005], symmetry and eccentricity [Backer et al., 2001, Heidemann et al., 2003], optical flow [Maki et al., 2000, Vijayakumar et al., 2001, Ude et al., 2005], and disparity [Maki et al., 2000, Ude et al., 2005]. Also specialized feature maps have been proposed such as skin color and facial features [Lee et al., 2005] and feature maps dedicated to auditory input [Trifa et al., 2007].

In most approaches discussed so far, no models of targets, the scene, or the current task are considered during saliency generation. Such approaches, where the saliency is calculated only based on the visual input fall into the class of bottom-up attention (see Section 2.1.1). In order to integrate top-down guidance, several advances of the Itti, Koch and Ullman model have been proposed. The majority of the research deals with top-down guidance based on target information, which is available directly or derived from a task definition. In the simplest approach, the guidance is implemented on top of the bottom-up process. In [Lee et al., 2005], the saliency map is combined with another top-down map, which is calculated using color cues in order to determine a modulated saliency map. By training a neural network with the features of distractors, corresponding locations in the bottom-up saliency map are inhibited in [Choi et al., 2004]. Another option to bring top-down guidance into the model consists in changing the way the different feature and conspicuity maps are combined to a single saliency map. In [Frintrop et al., 2005] and [Rasolzadeh and Eklundh, 2006], weights for feature and conspicuity maps are introduced in order to consider the current task. A similar principle is applied in [Navalpakkam and Itti, 2005], but the focus is put on internal representations and how they contribute in deriving the correct weights given a verbal task description. Another approach shows how the knowledge of distractors can be exploited in order to adapt the weights during saliency computation [Navalpakkam and Itti, 2006]. A different approach is proposed in [Moren et al., 2008]. In order to enable top-down modulation in the bottom-up model, the authors propose to combine the Feature Gate model with the Itti, Koch and Ullman model. The Feature Gate successively filters the input given a feature vector of the target from fine to coarse until only areas survive, which share similarities with the search target [Cave, 1999]. Moren et al. propose to consider the similarity of the target's features at different levels of detail in the calculation of the conspicuity maps resulting in a top-down modulated saliency map.

Until this point only approaches which have been derived from the initial work of Koch and Ullman and the implementation of Itti et al. have been discussed. In [Frintrop et al., 2010], these approaches are subsumed in the class of filter models according to their foundation in image filtering techniques. Additionally, the class of connectionist approaches is identified which comprises approaches that rely on neural processing at their cores. Prominent examples of such approaches are the Dynamic Routing Circuit [Olshausen et al., 1993] or the Selective Tuning Model [Tsotsos et al., 1995]. While those approaches are build on psychophysical evidence, the implementation of large scale systems based on neural networks is still challenging given current hardware.

3.2.2 Information Theoretic Approaches

The common basis for information theoretic approaches is the saliency calculation by means of an information theoretic measure. In the first attempts of applying information theory in saliency calculation, salient regions were defined by high information in terms of Shannon entropy in a local region. In [Gilles, 1998] gray scale histograms are calculated on local image patches and the entropy is determined from the histogram in order to define a saliency measure for the local region. The work in [Heidemann et al., 2003] includes an entropy map based on the entropy of local intensities in a psychophysical model. The approach proposed by Gilles was extended in order to include different scales by varying the patch size in [Kadir and Brady, 2001] and evaluated on different image sequences. This technique has been applied e.g. in image matching using a compact representation of images based on the salient regions [Hare and Lewis, 2003].

Another measure that has been proposed for the generation of saliency is the self information of local image patches. In [Bruce and Tsotsos, 2006], the self information is calculated by taking into account a local patch and adjacent patches in order to derive a measure for local contrast considering such an ensemble of patches. In contrast to the entropy, which encodes how a local patch differs from uniformity, the application of self information takes into account the surrounding regions in order to determine how outstanding a particular patch is. A complete framework for saliency based on self information is presented and evaluated in [Bruce and Tsotsos, 2009].

Several frameworks have been proposed using information theoretic measures, which formulate the problem of directing the gaze as information maximization process. In this context, an information theoretic framework for saccadic eye movements is presented in [Lee and Yu, 2000]. Based on the entropy of an ensemble of neurons and the mutual information with adjacent ensembles and considering a memory of attended locations, a saccadic behavior is modeled which maximizes the gathered information. This idea was further developed in [Renninger et al., 2005] where saccadic eye movements result from a sequential information maximization process in silhouette scanning tasks.

3.2.3 Bayesian Approaches

The common background of Bayesian approaches consists in the definition of saliency as a problem of Bayesian inference. Using the concepts of probability theory as basis, these approaches allow to consider uncertainties which result from noisy or incomplete visual measurements in a natural way.

The most prominent approach in this class is the Bayesian Strategy which uses a formulation of saliency which is explicitly derived from the goal of

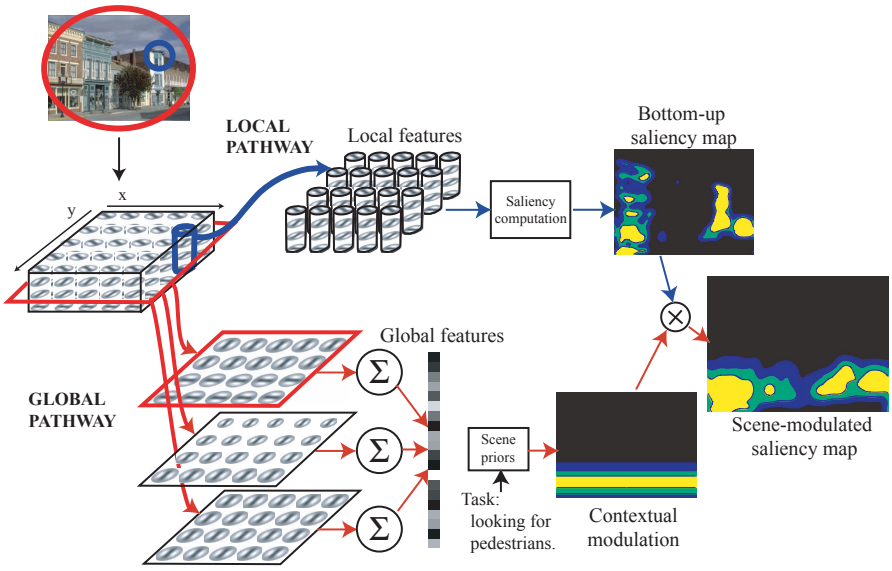

Fig. 3.2. Model for contextual guidance in detection of pedestrians using visual attention from [Torralba et al., 2006]. The approach follows the Bayesian Strategy, both bottom-up and context saliency result from the factorization of the model. The local pathway calculates bottom-up saliency, while the global pathway introduces priors in terms of spatial locations. Reprinted from [Torralba et al., 2006] with permission from APA.

object detection [Torralba, 2003]. Formally, saliency is defined as the posterior probability of detecting an object at a spatial location given local features and global features. Through the application of Bayes rule, this conditional probability is split into several factors to model different aspects of visual attention. The factors include a model of bottom-up attention as inverse to the conditional probability of the local feature given the global features, a model for target-driven control in terms of the conditional probability of local features given the scene in terms of global features and the target object, and a model of contextual priors. The contextual priors provide means to include the knowledge about the object identity and location given the global features of the scenes. The major focus in the work of Torralba and similar work in [Oliva et al., 2003] is put on these contextual priors which are again split into the factors of object-class priming, contextual selection of location, and contextual selection of object features. In [Torralba et al., 2006] the approach for contextual guidance is evaluated together with a bottom-up model for the detection of pedestrians (see Fig. 3.2).

A different Bayesian formulation is proposed in [Zhang et al., 2008]. There saliency is defined as the conditional probability of detecting an object in the

image given local features and the location. In contrast to the formulation proposed by Torralba, this saliency measure directly models the conditional dependency between the location and the presence of an object. After application of Bayes rule and expressing the resulting term with log likelihoods, the authors define the pointwise mutual information between the presence of the target and the features. The formulated saliency measure is evaluated in a bottom-up task.

In [Itti and Baldi, 2009] saliency is formulated as Bayesian surprise. Therefore, the prior believe of the observer is modeled using a random variable. New sensed data lead to an update of the prior using Bayes rule resulting in the posterior believe. Surprise and thus saliency is then described as the difference between prior and posterior believe in terms of the Kullback-Leiber divergence. The authors evaluate their model in human experiments using image sequences as data. It could be shown that Bayesian surprise predicts human performance well in comparison to models based e.g. on entropy.

3.2.4 Summary

As already highlighted in the introduction and discussed in the context of machine vision, the visual search for complex objects involves object models, which are used during the detection and recognition procedure. The ability of approaches for saliency calculation to include top-down information in terms of target knowledge is a necessary prerequisite for visual search applications in the real world. Further, a formalization of saliency should not only include information on the basis of a single factor but should allow the inclusion of different sources of saliency such as bottom-up information, top-down target specific knowledge and top-down scene knowledge in order to enable seamless integration of the different source. In the following, the different approaches for saliency calculation are discussed regarding their ability to exploit top-down target knowledge and to integrate different sources of information. An overview of the most relevant approaches is provided in Table 3.1.

Psychophysical approaches allow to integrate top-down target knowledge in different ways. While the initial approach proposed in [Itti et al., 1998] only allows to calculate saliency based on bottom-up information, further developments take into account the appearance of the target. This knowledge is either considered in terms of weights between the different feature and conspicuity maps [Frintrop et al., 2005, Itti and Baldi, 2006] or by means of the Feature Gate approach [Moren et al., 2008]. Only using weights for the maps provides little control over the saliency generation process. Only if the target is identified by a combination of the low level features, the results are promising. The Feature Gate approach goes one step further and allows to consider guidance at different levels of detail. However, in order to apply this technique, a

approach	bottom-up	target knowledge	scene knowledge
psychophysical			
[Itti et al., 1998]	X	-	-
[Frintrop et al., 2005]	X	X	-
[Moren et al., 2008]	X	X	-
information theoretic			
[Lee and Yu, 2000]	X	-	-
[Renninger et al., 2005]	X	-	-
[Bruce and Tsotsos, 2009]	X	-	-
Bayesian			
[Torralba, 2003]	X	X	X
[Zhang et al., 2008]	X	-	-
[Itti and Baldi, 2009]	X	(X)	(X)

Table 3.1. Classification of different approaches for the generation of saliency in visual attention systems. The ability of including bottom-up and top-down knowledge in the saliency generation based on the approaches is highlighted. For top-down knowledge, a distinction is made between knowledge about the target in terms of its features and knowledge about the scene such as location priors and target priors.

hierarchical representation of the target is required. Such representations are subject to ongoing research activities.

The dominant saliency measures in information theoretic approaches are entropy and self-information. These measures are calculated based on the information which is available in the current scene. In order to derive the necessary statistics, a collection of scenes is required [Bruce and Tsotsos, 2009] but not applied in terms of top-down guidance. Consequently, non of the approaches explicitly models top-down influence neither for target knowledge nor for scene knowledge.

The class of Bayesian approaches comprises three different strategies for exploiting Bayesian inference in order to calculate saliency. In [Zhang et al., 2008] the pointwise mutual information is proposed as saliency measure, which is derived from a model taking into account local features and positions. The saliency measure thus only includes bottom-up information. The more general concept of Bayesian surprise as proposed in [Itti and Baldi, 2009] allows to model bottom-up as well as top-down influence. However, the implementation presented only considers information available in the input and further development and specification is required in order to apply the model using top-down guidance. In contrast, the Bayesian Strategy as proposed in [Torralba, 2003] allows to seamlessly integrate different sources of information and especially the inclusion of target knowledge.

Until now, only the contributions in terms of saliency generation of the different approaches were compared. For the application in active visual search another important requirement are methods which allow to generate gaze sequences from the saliency data. In psychophysical approaches, the WTA network is commonly used in order to generate scan patterns using the inhibition of return mechanism [Itti et al., 1998]. Taking into account a memory representation of attended locations, the sequential information maximization allows to generate scan patterns in the class of information theoretic approaches [Lee and Yu, 2000]. In the class of the Bayesian approaches no such mechanisms have been proposed until now.

3.3 Active Visual Search

The literature offers a vast amount of approaches towards active visual search due to the ambiguity of the term. In the context of this work, we focus on relevant research which has a similar goal in terms of visual capabilities: the detection of one or multiple known target objects in the scene using an active camera system. Such approaches typically implement strategies for the generation of visual attention and then use the active system to guide the gaze in order to recognize the target object. In particular, approaches are not covered that implement pure bottom-up attention on an active system such as e.g. [Ude et al., 2005] and [Ruesch et al., 2008]. While such approaches are able to generate very natural gaze patterns, they do not share the goal of detecting and recognizing the target object.

The focus in the following is put on development in active visual search which deals with implementations for hardware platforms that offer active capabilities. Such platforms range from active stereo sensors, active anthropomorphic heads and mobile platforms to integrated humanoids. In the following, research that deals with egomotion of such platforms is not covered. Rather, active visual search refers to generating a gaze sequence in terms of head and eye movements resulting in ego-centric perception. In contrast, when considering egomotion, the problem of finding an object is often formulated as a sensor planning problem involving locomotion (e.g. [Shubina and Tsotsos, 2010, Andreopoulos et al., 2011]) or embedded in a simultaneous localization and mapping task (e.g. [Frintrop et al., 2007, Frintrop and Jensfelt, 2008]) which is not the scope of this work.

Beginning with approaches of active visual search using actuated eyes, the following review of research proceeds with systems that apply special techniques for foveated vision. Approaches which integrate memory are treated in the subsequent section before all relevant approaches are compared and discussed.

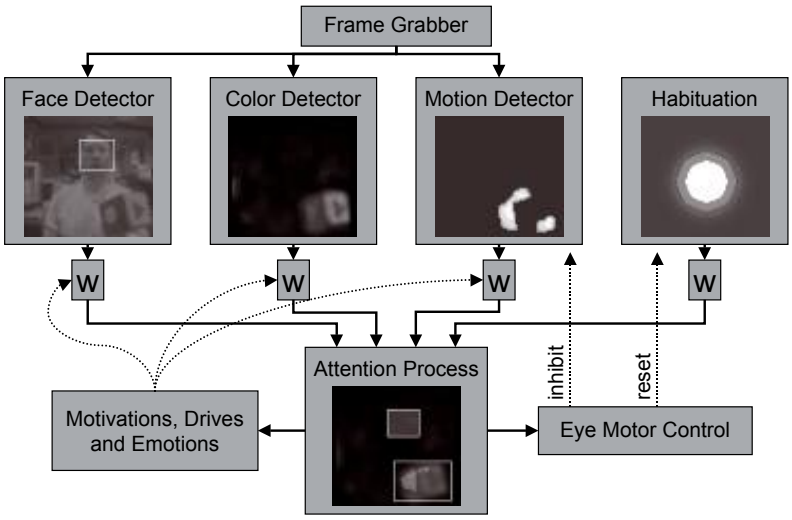

Fig. 3.3. Attention system implemented for the social robot Kismet focusing on natural human machine interaction. The active visual search for colored objects and faces is supported. The system uses conspicuity maps for faces, colors, and motion combined to a saliency map using weights which define the top-down bias. A habituation map ensures gaze sequencing. Reprinted from [Breazeal and Scassellati, 1999] with permission from Elsevier.

3.3.1 Visual Search with Actuated Eyes

From a machine vision point of view, the benefit of the ability to focus on objects using actuated eyes in active stereo systems is twofold: First, the part of the scene which is visible to both cameras can be adapted to the needs of the application. Second, the focused element projects to the same coordinate in the left and right rectified images. Both advantages ease the common task of depth processing in stereo vision. A common technique introduced early to the field which makes use of these advantages in order to segment focused elements is zero disparity filtering [Coombs, 1992]. This approach stands at the core of many approaches for active visual search since it enables figure ground segmentation of focused elements. An early system for object detection and recognition on an active stereo platform that makes use of zero disparity filtering is presented in [Rao and Ballard, 1995]. At the same time, attentive mechanisms have been introduced in order to implement active visual search (see [Grimson et al., 1994]) again exploiting zero disparity filtering.

An approach which highlights the importance of eye movements for humanoid robots that act in a social manner has been presented in [Breazeal et al., 2001]. The Kismet platform [Brooks et al., 1999] as depicted in Fig. 3.3, left has been

developed in order to investigate social aspects in humanoid robotics. Based on a psychophysical attention system emotions, motivations, and drives are employed as top-down guidance [Breazeal and Scassellati, 1999]. As depicted in Fig. 3.3, right the top-down guidance is implemented in terms of weights between the feature maps which include faces, colors and motion. A sequence of gaze directions is generated using a habituation map which acts as a feature map and amplifies the currently focused point and is subject to decay in order to allow gaze shifts. Using a static mapping between saliency map and eye position, the gaze of the cameras is directed to the region with maximum saliency.

3.3.2 Visual Search using Foveated Vision

The active stereo systems covered in the last paragraph lack one essential property of the human visual system. The human eye exhibits space-variant resolution over the visual field with a high resolution area, the fovea, close to the optical axis and with decreasing resolution towards the periphery. This property seamlessly integrates with attentional mechanisms since already at the physical level a reduction of information is provided. Further, techniques such as overt visual attention are motivated, which allow to focus on scene elements whenever acuity is needed. There are essentially two techniques how this principle is realized in state of the art platforms. In log polar vision systems, the goal is to accurately mimic the decreasing resolution towards the periphery using a structure derived from a log polar coordinate system [Schwartz, 1980]. In contrast, quadrifocal systems use two cameras per eye. One camera is equipped with wide angle lenses and mimics the periphery. The fovea is implemented using cameras with a small field of view. Additionally, some systems offer a motorized zoom for one camera [Sharkey et al., 1993]. Since the quadrifocal concept allows to integrate out of the shelf cameras most platforms make use of this implementation of foveated vision.

A system which is based on log polar vision has been presented in [Orabona et al., 2005] and [Orabona et al., 2007]. From the log polar images color opponencies are calculated in different channels using a center surround mechanism in order to mimic human processing. The resulting maps are subject to edge detection and combination to a single map. Regions from the map of combined edges are calculated and grouping is performed based on color. The resulting blobs are termed proto-objects within this approach. Based on the proto-objects, a top-down cue is integrated, which determines the similarity between blobs and a given color signature and attention is drawn to the most salient location. In order to generate a sequence of gaze directions, object based inhibition of return is proposed which is associated to each instance of a proto-object. The bottom-up part of the attention system has been demonstrated using the BabyBot platform in the context of sensori-motor learning [Natale et al., 2005].

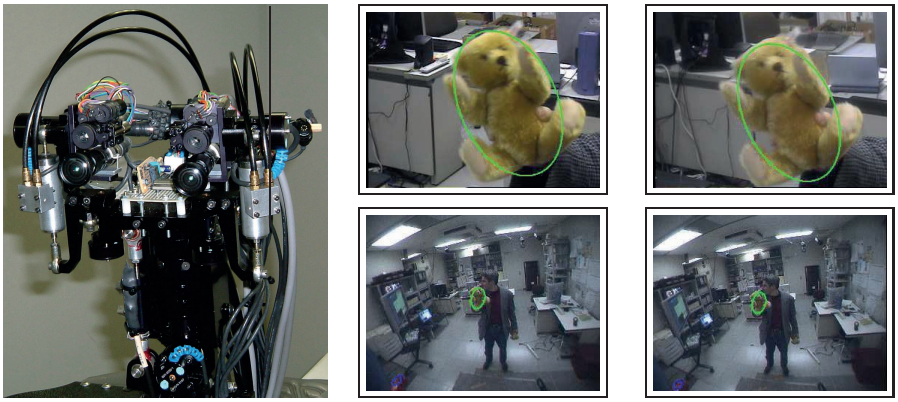

Fig. 3.4. Left: The humanoid platform DB offers an active head with foveated vision implemented using two cameras per eye. Right: The active visual search approach proposed in [Ude et al., 2003] actively tracks the target object in the peripheral image to perform recognition in the detailed foveal view. Reprinted from [Ude et al., 2003] ©2003 IEEE.

The application of a quadrifocal system using peripheral and foveal cameras in order to detect and recognize objects has been presented in [Ude et al., 2003]. In constrast to active visual search in a static scene, the focus is put on recognizing objects in the detailed view of the foveal cameras which are subject to movement. As experimental platform, the humanoid robot DB [Atkeson et al., 2000] is used, which offers an active head with peripheral and foveal cameras, 2 DoF per eye and 3 DoF in the neck (see Fig. 3.4, left). In addition to the head DoF, three DoF of the torso are also controlled by the active system. Using wide angle lenses in the peripheral cameras, the captured images are used in order to detect and track target objects. As cues for object detection the system relies on color and shape processed in a probabilistic fashion. The object color provides hypotheses for the initial detection, whereas the space of elliptical shapes is randomly sampled. Once the detection was successful, the system closes the loop in order to bring the target object into the foveal camera. While still tracking the object in the periphery, the same detector is executed on the foveal images in order to detect the corresponding region. Once the target object is detected in the foveal cameras, recognition is performed. While the orientation of the object is known from the ellipse of the detection process, invariance to scaling is achieved by normalizing the detected region. Further, the resulting view is filtered using LoG in order to achieve robustness toward illumination. Based on the resulting image vector, principle component analysis (PCA) is performed in order to project the image vector to a low dimensional feature space, where classification is performed by means of the minimal Euclidean distance to previously acquired prototypes. In [Ude and Cheng, 2004], a different approach for the

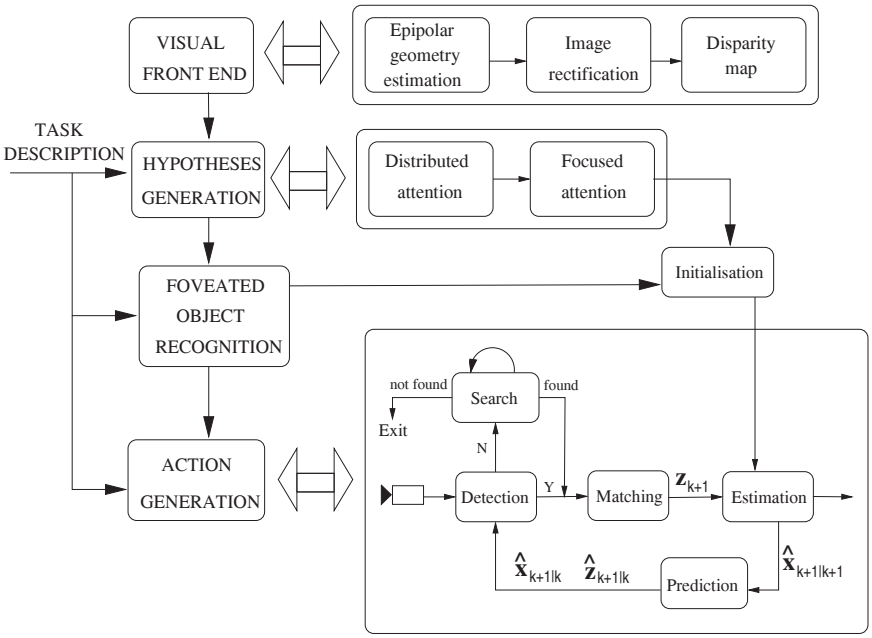

Fig. 3.5. The active visual search approach in [Bjorkman and Kragic, 2004] includes peripheral vision. Hypotheses of object locations are determined in the peripheral image pair based on disparity using object size and color. Attention is focused on the maximum salient region taking into account random noise. Object recognition based on SIFT and color is performed in the detailed view of the foveal cameras. On top of the active visual search, several actions are implemented such as smooth pursuit and pose estimation. Reprinted from [Bjorkman and Kragic, 2004] ©2004 IEEE.

recognition in the foveal cameras is proposed. The normalized images are processed using Gabor filters. The filter response is then classified using Support Vector Machines. Further development toward a general closed loop controller in the context of active vision is integrated and discussed in [Ude et al., 2004].

Another system that combines active and foveated vision has been initially proposed in [Bjorkman and Kragic, 2004] and [Bjorkman and Eklundh, 2004]. As target platform for the evaluation the Yorick Head [Sharkey et al., 1993] is used which consists of a peripheral and a foveal camera pair. The 4 DoF of the head include seperate pan movements of the left and right camera pairs. The proposed system not only covers active visual search but integrates tracking and pose estimation based on the estimated object positions. The following discussion of the system is restricted to the active visual search for known objects.

In Fig. 3.5, left the building blocks of the overall system as proposed in [Bjorkman and Kragic, 2004] are illustrated. The visual frontend is responsible for the calculation of disparity maps based on the peripheral camera images. Therefore, online estimation of the epipolar geometry of the stereo camera pair is performed using the optical flow constraint based on 2D points extracted using the Harris corner detector. Based on the estimated model, a disparity map is calculated. The disparity map is used in the hypotheses generation block together with a search task specification in order to guide attention toward promising locations in the scene. In the initial implementation, the object size and a hue signature are used as top-down cue. The generation of object hypotheses is illustrated in Fig. 3.5, right. The list of candidates resulting from the hypotheses generation step is weighted according to the similarity with the hue signature of the search target. In order to cope with cases, where the target object does not exhibit the largest peak, random noise is added to the weights and the gaze is directed toward the highest peak. Once the saccade toward this position has been executed, object recognition based on the foveal camera pair is performed. After fixating the hypothesis location, figure ground segmentation is performed in the foveal camera pair using a mean-shift approach by taking into account the known object size. The resulting segmented region is subject to recognition using a combination of SIFT matching and color cooccurrence histograms (see Fig. 3.5, right). If the recognition fails, the procedure is repeated usually resulting in fixating another hypothesis due to the random noise added to the weight of the peaks. This procedure continues until the object has been found. A more detailed discussion of the overall system is provided in [Bjorkman and Eklundh, 2006], an application to manipulation of familiar and unseen objects is presented in [Kragic et al., 2005].

The active vision system described above is sophisticated in terms of the integration of different abilities such as tracking and pose estimation which are referred to as actions. The methods for coping with depth maps using active cameras, relating peripheral to foveal cameras, and figure ground segmentation using zero disparity in the foveal cameras are emphasized. In contrast, the generation of gaze sequences is straight forward, using the maximum salient peak with additional random noise. A more sophisticated approach for the generation of attention has been integrated into the system in [Rasolzadeh et al., 2010]. Using the 7 DoF Karlsruhe Humanoid Head [Asfour et al., 2008] as target platform, the authors have integrated the attention mechanism from [Rasolzadeh and Eklundh, 2006]. As already discussed in Section 3.2.1 attention is thereby generated using a combination of bottom-up saliency according to the Itti and Koch psychophysical model and a top-down saliency map which is derived by adapting the weight of different feature and conspicuity maps. The combination of both saliency maps is performed in a way that prevents one map of dominating the other one.

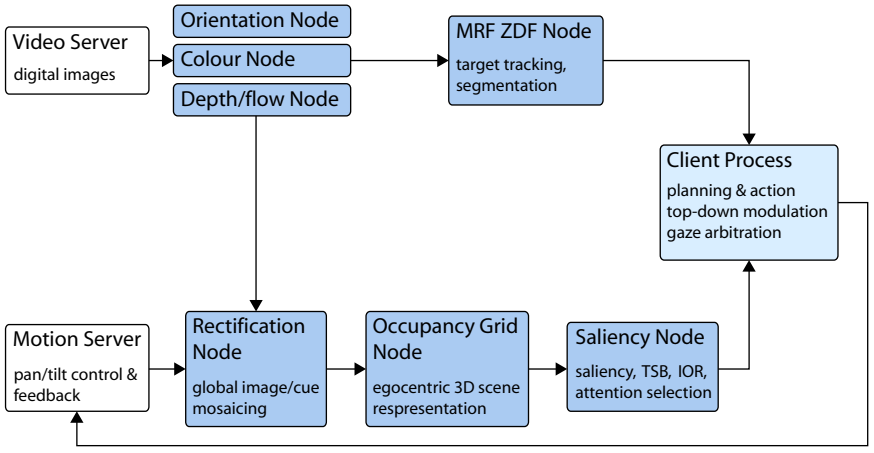

Fig. 3.6. The approach proposed in [Dankers et al., 2009] integrates memory into active visual search. An egocentric spatial representation of depth, color, and orientation is accumulated over different gaze directions. This memory is considered during saliency generation and allows to focus on invisible scene elements which have been previously attended. On fixation, zero disparity filtering is performed in order to allow classification of the object.

3.3.3 Active Visual Search and Memory

All systems that have been discussed so far exhibit the one common property that they report the target object in terms of the last fixation of the system. Once the target object is recognized, the perceptual task is finished. This property poses a restriction on the applicability in the real world in two ways. First, such systems only allow to report one instance of a searched object which usually corresponds to the highest saliency. In embodied tasks such as manipulation on humanoid platforms the object with highest saliency is not necessarily the optimal choice. Second, such approaches do not account for the dynamic nature of real world scenes. An object which has been recognized at a previous point in time might have been moved or removed by e.g. interfering agents. In order to cope with such situations, a memory of attended regions that is build during different gaze shifts is necessary. It provides the means to handle multiple hypotheses. Further, by considering memory and visual search as a continuous process, events such as relevant changes in the world can be detected and appropriate actions can be taken. In the following, memory representations and their application in active visual search are reviewed.

The application of a sensory ego-sphere [Peters II et al., 2001] as short term memory of spatial locations in active visual search has been proposed in [Figueira et al., 2009b] and [Figueira et al., 2009a]. The approach is imple-

mented for the active head of the iCub platform which has 3 DoF in the neck and 3 DoF in the eyes [Metta et al., 2008]. A top-down saliency map is calculated by matching SIFT features of the target object within the currently perceived scene. Once an object is recognized, its position is mapped to a unit sphere and the associated angular part of the spherical polar coordinates is stored as 2D position. From the memory, a spherical saliency map is calculated, where each memory entity maps to a 2D Gaussian distribution with covariance according to the errors as encountered during 3D localization. In order to generate a gaze sequence between different memory entities, a fixed decay time is implemented which modulates the saliency within the top-down map. A similar representation of spatial memory is proposed in [Aryananda, 2006]. Based on face, color, and auditory input, saliency is calculated on the 2D projection of the world coordinate frame using Gaussian distributions. For each of the projections, a saliency function is defined which includes decay in order to generate gaze sequences.

A different kind of spatial memory is proposed in [Dankers et al., 2007] and [Dankers et al., 2009]. The authors propose an attention system which includes the execution of gaze shifts on an active head. As experimental platforms the 3 DoF CeDAR head [Truong et al., 1999] and the iCub head [Metta et al., 2008] are used. The system structure is outlined in Fig. 3.6. Attention is generated based on orientation, color, motion, and depth which are processed in a bottom-up fashion. Using the disparity estimates from stereo processing, a 3D occupancy grid is used as memory for spatial locations which allows the fusion of depth measurements from each gaze. Additionally, the extracted bottom-up cues are associated with the corresponding grid cells in order to enrich the memory representation. Gaze shifts are executed based on saliency calculated from local contrast of the different cues. Upon fixation, zero disparity filtering is performed in order to segment the currently fixated entity. The generation of gaze sequences is realized by implementing decay using an inhibition of return mechanism.

3.3.4 Summary

In the following, the discussed approaches are compared in terms of the methods used in order to direct the gaze and the integration of spatial memory. An overview of the most important approaches is provided in Table 3.2.

The most frequently deployed methods of visual attention in active visual search are modifications of the Itti-Koch psychophysical model. Thereby a saliency map is calculated based on conspicuity from different feature channels. While [Rasolzadeh et al., 2010] use the channels proposed in the model of Itti and Koch, other authors propose the application of different features which are tuned to the task envisioned by the approaches. Faces, colored objects, and motion are used in order to detect and pursue human faces and simple

approach	target platform	foveated vision	attentional mechanism	spatial memory
[Breazeal and Scassellati, 1999]	Kismet	-	**bu:** face, color, motion **td:** weighted maps **seq:** habituation map	-
[Ude et al., 2003]	Dynamic Brain (DB)	quadrifocal	**bu:** - **td:** color, shape **seq:** maximum saliency	-
[Bjorkman and Kragic, 2004]	Yorick Head	quadrifocal	**bu:** depth map **td:** size and hue **seq:** additive noise	-
[Rasolzadeh et al., 2010]	Karlsruhe Head	quadrifocal	**bu:** Itti-Koch **td:** weighted maps **seq:** additive noise	-
[Orabona et al., 2005]	BabyBot	log polar	**bu:** color opponency **td:** - **seq:** FIFO	-
[Figueira et al., 2009b]	iCub	-	**bu:** depth **td:** SIFT **seq:** memory decay	sensory egosphere (2D)
[Dankers et al., 2009]	iCub	-	**bu:** orient, color, depth **td:** - **seq:** Gaussian inhibition	occupancy grid (3D)

Table 3.2. Comparison of relevant approaches for active visual search. The target platform and the type of foveated vision is given in columns two and three. The mechanisms for bottom-up attention (**bu**), top-down attention (**td**) and gaze sequencing (**seq**) are listed in column four. For systems that implement spatial memory, the type of memory is provided in column five.

colored objects in [Breazeal and Scassellati, 1999]. Color opponency is used to provide bottom-up saliency in [Orabona et al., 2005]. The approach proposed in [Dankers et al., 2009] combines orientation, color, and depth in the bottom-up processing. The top-down guidance in these approaches is either implemented using weighted saliency maps [Breazeal and Scassellati, 1999, Rasolzadeh et al., 2010] or happens in a later stage after saliency calculation [Orabona et al., 2005, Dankers et al., 2009]. A different approach for bottom-up processing is pursued in [Bjorkman and Kragic, 2004] and [Figueira et al., 2009b], where depth is used as exclusive bottom-up cue. Using the depth cue, image fragments are grouped and are subject to matching with object models in a top-down fashion. The recognition is performed using SIFT and color features. The approach proposed in [Ude et al., 2003] does not make use of bottom-up cues but rather performs a top-down search of the object signature using shape and color cues.

In order to generate gaze shifts from the saliency map, all approaches identify the maximum saliency and use the corresponding 3D position as target point. In order to generate sequences of gaze shifts, different approaches are proposed. One class of approaches uses decay of the attended locations over time. In [Breazeal and Scassellati, 1999], the decay is realized using habituation maps which are treated as additional feature channels while in [Dankers et al., 2009] a Gaussian kernel serves as bias for the saliency gen-

eration. Another approach is proposed in [Orabona et al., 2005], where attended locations are stored in a FIFO buffer of fixed size. Locations are not revisited if they are present in this buffer and if their color signature did not change. The approaches proposed in [Bjorkman and Kragic, 2004] and [Rasolzadeh et al., 2010] randomly add noise to the peaks in saliency in order to allow focusing to areas which differ from the maximum saliency. Since the objective in [Ude et al., 2003] is foveation and subsequent pursuit of an object, only one saccade is generated toward the target object before smooth pursuit is initiated. An elaborated approach for the generation of gaze sequences is proposed in [Figueira et al., 2009b] where a spatial memory is taken into account during saliency generation. For each memory entity which resides in the spatial memory a constant decay factor is implemented which allows refocusing of objects that are temporarily out of sight of the active vision system.

Two recent approaches integrate spatial memory into the system [Figueira et al., 2009b, Dankers et al., 2009]. In both approaches, the spatial layout of the memory and the information stored differs. In [Dankers et al., 2009] a 3D occupancy grid map is used in order to store low level information such as depth, orientation, and color. The approach in [Figueira et al., 2009b] proposes a 2D representation on the unit sphere which encodes information about spatial locations of objects. Both approaches retain information of previously attended locations and apply their principle of gaze sequencing to the memory entities. As such they are able to explicitly handle multiple object candidates and to direct attention toward previously seen objects which are not in the current visual field of the system. The application of memory requires to solve the correspondence problem of perceived entities during changing gaze directions. In [Figueira et al., 2009b], this problem is solved by representing objects with a 2D Gaussian spatial uncertainty equivalent to the maximum inaccuracy in object localization as determined for the system. The correspondence problem in [Dankers et al., 2009] is solved based on the occupancy grid cells.

3.4 Discussion

In the review of related work, two essential aspects of active visual search were covered. First, attentional mechanisms which are an integral part of active systems that search were discussed in Section 3.2. Second, state of the art system that perform active visual search were reviewed and discussed with focus on integrated systems in Section 3.3.

The complexity of the visual search problem motivates the necessity to approach the problem using techniques that provide a feasible way to deal with restricted resource [Tsotsos, 1989]. Attentional mechanisms are such techniques which allow to focus only on relevant input data and neglect data

which is irrelevant for the task. Methods proposed in order to implement visual attention on technical systems fall into one of three classes: the psychophysical approaches, the information theoretic approaches, and the Bayesian approaches. In the context of active visual search the approaches were compared according to their ability to integrate different sources of information in order to determine regions of interest in the image. The focus was put on the ability to integrate top-down knowledge in terms of a target object model which is a prerequisite for bounded active visual search. Furthermore, the strength of the approaches in integrating different sources of information in a general formalization has been discussed. The psychophysical approaches offer restricted abilities to integrate top-down guidance into the model. Top-down knowledge is either applied after saliency has been calculated or the knowledge serves as weights for the different feature maps [Frintrop et al., 2005]. The weighting of feature maps provides weak control for the guidance of attention and strongly depends on the type of features deployed in order to calculate the conspicuity maps. The information theoretic approaches use local statistics of the image in order to determine a measure of information (e.g. [Bruce and Tsotsos, 2009]). Thereby, top-down knowledge in terms of object models is not considered. In constrast, the Bayesian Strategy allows to seamlessly integrate different sources of information in a general formalization [Torralba, 2003]. In particular, it allows to integrate top-down information in a very general way providing strong control for guiding attention.

The application of an active stereo system with moving eyes in order to implement active visual search allows to focus on relevant elements of the scene as determined by the attentional mechanism. In foveated vision systems the principle of attention to focus processing resources on relevant data is mimicked on a physical level by providing an area of increased acuity in terms of resolution. In order to guide the gaze based on attention on such platforms, most approaches rely on the original or a modified Itti-Koch model (e.g. [Dankers et al., 2009, Rasolzadeh et al., 2010]). While using this psychophysical approach is biologically plausible it is accompanied with the disadvantage of weak control over attention guidances as discussed above. This problem is approached by defining appropriate feature channels in a way that weighting of different maps is sufficient for the perceptual task [Breazeal and Scassellati, 1999]. Nevertheless, no general solution to the problem of active visual search is achieved. Most control over the guidance of attention is provided when performing pure top-down processing [Ude et al., 2003].

While the integration of a spatial memory in approaches for visual search has been proposed in previous work (e.g. [Backer et al., 2001]) actual systems that make use of memory have only been demonstrated recently. Memory in these systems has either a 2D [Figueira et al., 2009b] or a 3D [Dankers et al., 2009] spatial layout and accumulates information about attended locations over several gaze shifts. The application of a spatial memory in active visual search is beneficial in several ways. The persistence of previously attended locations

allows to relax the restriction to target locations which are visible to the cameras during the complete search procedure. Further, a memory which holds information about attended entities allows to handle multiple hypotheses. Answering the problem of visual search with the memory content enables the robot to choose the most feasible hypotheses given the current task. Most importantly, including memory allows to consider active visual search as a continuous process which provides a consistent and persistent answer to the location of target objects.

In this work, an approach is proposed which brings together the Bayesian strategy to visual attention and spatial memory in an integrated foveated active visual search system. By considering active visual search as a process, the requirement of a consistent memory provides the drive which guides the gaze toward salient locations in the scene in a top-down manner. In contrast to [Dankers et al., 2009] where information is accumulated on a feature level, the accumulation of information about objects provides a consistent answer to the search problem. A similar approach is proposed in [Figueira et al., 2009b] which does not make use of foveated vision. Further, the authors do not propose a general way how to integrate memory with existing techniques for the generation of saliency but rather rely on disparity information in order to segment the scene and recognize objects. Further, the restriction to a 2D spatial layout introduces inaccuracies if the cameras are moved in an active fashion.

4

Memory-Based Active Visual Search

In the following, the approach for memory-based active visual search is outlined. Before the objective is further refined in the light of previously discussed mechanisms of active vision and visual search in Section 4.2, the target platform is specified in the following section.

4.1 The Target Platform

The goal of this thesis consists in enhancing the visual perception capabilities of the humanoid robot platform ARMAR-III [Asfour et al., 2006] which has been designed for the application in human-centered environments. As depicted in Fig. 4.1, left the robot is equipped with two 7 DoF arms, five-fingered hands, and an active head. For locomotion, the upper body is mounted on a mobile platform which allows holonomic movements. The system has been built with a strong focus on autonomy in terms of power supply as well as the on-board integration of the necessary processing resources. Given its ability of dexterous two arm manipluation and its high degree of autonomy, ARMAR-III is well suited for the realization of typical tasks in a household scenario.

The active head of the robot is also available as a stand-alone version under the name Karlsruhe Humanoid Head [Asfour et al., 2008] as depicted in Fig. 4.1, right. The head features an active camera system with two cameras per eye, six microphones, and a gyroscope-based orientation sensor. The neck is actuated with four joints where the first three joints realize roll, pitch, and yaw and the fourth joint implements an upper neck tilt. The eye system comprises three DoF, a common tilt for both eyes and one pan joint per side in order to allow vergence of the eyes. The camera system is quadrifocal, each eye is equipped with two cameras mimicking foveal and peripheral vision of the human visual system. All four PointGrey Dragonfly2 cameras provide a resolution of 640 $\times$ 480 pixels at a frame rate of 60 frames per second. The camera images are captured via IEEE 1394 either with the on-board embedded PC or externally.

Fig. 4.1. Left: The target platform for this thesis is the humanoid robot platform ARMAR-III. Right: The active head of ARMAR-III is available as stand-alone version under the name Karlsruhe Humanoid Head. The head is equipped with a active foveated vision system actuated by 7 DoF.

4.2 Outline of the Approach

The problem to be solved in this thesis has been formulated as the detection and recognition of objects using moving eyes. Further, a memory of relevant objects in the scene is to be established in order to provide consistent and persistent spatial information. While the first requirement corresponds to the active visual search problem, the second requirement extends the answer of active visual search from the fixation point of a single object to the content of a spatial memory.

As described above, the Karlsruhe Humanoid Head offers an active foveated camera system which allows to implement the eye movements required to realize active visual search. Making use of the active eyes and the quadrifocal foveated camera system the field of view of the robot is augmented in terms of the covered scene fraction. From the complexity analysis of the visual search problem as discussed in Section 3.1, principles have been derived to solve the problem on technical systems with limited resources. Among those principles, the application of attentional mechanisms and a hierarchical organization of visual processing have been proposed. Thereby, visual attention allows to focus the limited processing resources to only relevant input data. The hierarchical organization enables a coarse to fine approach thus decomposing the search task. For both mechanisms, a solution is presented in this thesis.

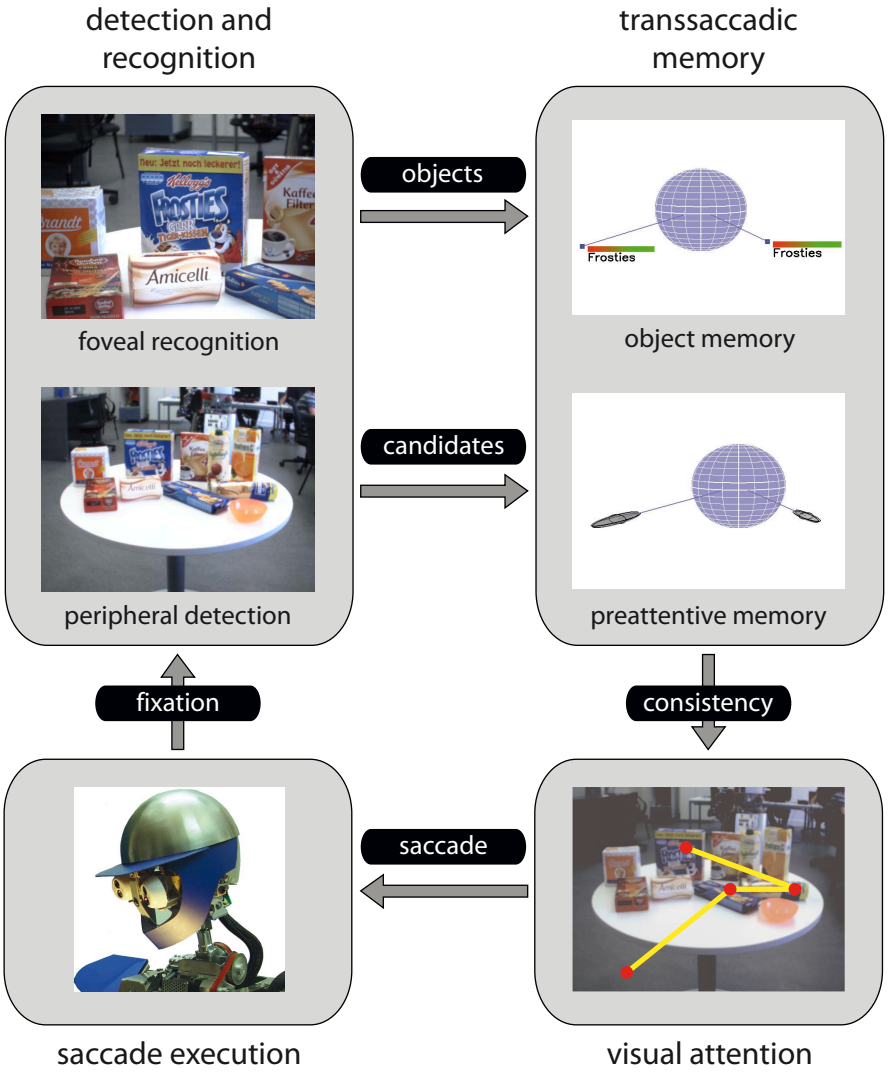

Fig. 4.2. Decomposition of the active visual search problem into subproblems addressed in this work. Object candidate detection is performed based on the peripheral images. The correspondence problem for object candidates from different gaze directions is solved based on the transsaccadic memory content resulting in preattentive memory entities. From the memory content, the attentional mechanism generates a sequence of saccadic eye movements. The fixation of an object candidate allows to perform object recognition. The result is stored in the object memory which constitutes the output of the approach.

Hierarchical organization

A hierarchical organization is proposed which is tightly coupled to the foveated nature of the vision system. The task of visual search is decomposed into a coarse peripheral analysis and foveal verification. This decomposition results in two processing chains (see Fig. 4.2). One processing chain is associated to the peripheral cameras, another to the foveal cameras. In the peripheral chain the input images are first subject to object candidate detection. Thereby, regions of interest are determined that likely contain the target object. Based on the regions of interest, entities are generated in the preattentive layer of the transsaccadic memory. In the course of the active visual search, object candidates are extracted from different gaze directions. The fusion of these candidates in the preattentive memory involves solving the correspondence problem between past observations and the current measurement. The resulting content of preattentive memory serves as prior for locations in the scene which are relevant for the search task and thus restricts the search space.

In the foveal processing chain, object recognition is performed based on the detailed views of the foveal cameras. These detailed views are taken according to the restricted search space which results from the preattentive memory content. The recognition procedure generates entities in the object memory layer of transsaccadic memory and associates them with corresponding object candidates in the preattentive layer. Each time an entity in object memory is brought into the foveal cameras, the accompanied position is corrected in a closed-loop fashion in order to derive consistent spatial information. The object memory constitutes the output of the memory-based active visual search approach.

Integrating attention and memory

In order to guarantee the consistency of transsaccadic memory, continuous verification of its content is performed by directing the gaze of the active system. For this purpose, an attentional mechanism is proposed which takes into account the probability of detecting an object and the consistency of the object memory layer at the target location. The attentional mechanism is derived from the Bayesian Strategy to visual attention which offers seamless integration of top-down guidance in a general formalization (see Section 3.2). Thereby, gaze shifts are generated based on the content of transsaccadic memory. Each saccade brings one object from transsaccadic memory to the view of the foveal cameras. The occurrence of change in the scene as well as the validation using foveal vision affect the consistency of object memory and are considered by the attentional mechanism. In the context of the overall approach, the memory and the requirement of its consistency play a central role in the generation of the saccadic behavior.

Overview of the approach

In Fig. 4.2 the layout of the memory-based active visual search approach is illustrated by subdividing the problem into four smaller tasks. The left part of the figure comprises the subproblems of object perception and of saccade execution which directly interface with the hardware in terms of cameras and motor controllers. The right part of the illustration contains the transsaccadic memory and the visual attention mechanism which close the perception-action cycle.

The realization of the memory-based active visual search approach is discussed in the subsequent four chapters. Each chapter corresponds to a subproblem as illustrated in Fig. 4.2. First, the approaches for object candidate detection and object recognition are introduced and evaluated in Chapter 5. In order to execute saccades to given target positions, a kinematic calibration of the head eye system is necessary which is outlined and evaluated in the context of saccade execution in Chapter 6. The organization of the transsaccadic memory is discussed in Chapter 7 accompanied with a solution to the correspondence problem. Finally, Chapter 8 provides a formalization of the attentional mechanism which allows to generate gaze sequences that maximize the detection probability and the memory consistency.

5

Object Detection and Recognition

The active visual search involves a two stage approach for recognizing instances of the target object in the scene as introduced in Section 4.2. First, object detection is performed in the peripheral images. Once attention is focused to a potential target which has been identified using peripheral object detection, the foveal camera images allow robust recognition of the searched object. In the following, the approach for object detection in the peripheral images is detailed in Section 5.1 and the approach for object recognition in the foveal images is described in Section 5.2.

5.1 Object Detection in the Peripheral Images

The images of the peripheral cameras provide a comprehensive view of the scene as illustrated in Fig. 5.1. Each object in the image only covers a small region making robust object recognition difficult. Instead, the proposed two stage approach performs object recognition in the detailed foveal views. In order to avoid the execution of costly processing steps in the foveal images, the peripheral image processing acts as filter for potential target object locations, so-called object candidates, which are then subject to foveal object recognition. Object candidates are determined based on their similarity to the target object using a descriptor suitable for small object regions.

In the following sections the approach for object candidate detection in the peripheral images is introduced. Based on the image representation described in Section 5.1.1 an appearance based image descriptor is proposed in Section 5.1.2 accompanied with a suitable similarity measure. Subsequently, the detection of object candidates is outlined in Section 5.1.3 and extended to the generation of 3D locations of object candidates in Section 5.1.4.

Fig. 5.1. Table setup captured in the peripheral camera using lenses with 4 mm focal length.

5.1.1 Image Representation

The representation of the acquired peripheral images has a large impact on the performance in object detection tasks. The decision for an adequate representation of the image involves choosing a color space as well as suitable image properties which are then provided for the appearance based descriptor.

In the following, different channels c are defined which are represented by functions $f_c(u, v)$. Each channel extracts a specific property of the image at the spatial location (u, v). Images are either represented in RGB color space with $\mathbf{I_{rgb}}(u, v) = (R, G, B)$ or in HSV color space with $\mathbf{I_{hsv}}(u, v) = (H, S, V)$.

RGB color space

In the RGB color space, each pixel is represented with three values corresponding to the primary colors red, green and blue. The resulting color is an additive mixture of these primary colors. In order to achieve invariance to constant light intensity changes, the RGB values can be represented using the normalized $R'G'$ representation [Finlayson et al., 1998]:

$$R' = \frac{R}{R+G+B} \ , \ G' = \frac{G}{R+G+B}.$$

The following channels are calculated from the $R'G'$ representation:

- **Chromaticity:** The RG-chromaticity of the image is represented using the functions $f_r = R'$ and $f_g = G'$ defined on the normalized $R'G'$ representation. The blue channel is omitted in the RG-chromaticity.
- **Gradient:** The image gradient is calculated from the partial derivatives in the image plane. The gradient reflects the color change with respect to the spatial domain. As channels, the magnitude of the gradient is represented using the functions

$$f_{\nabla_R} = \sqrt{R'^2_x + R'^2_y}\ ,\ f_{\nabla_G} = \sqrt{G'^2_x + G'^2_y}.$$

- **Laplacian:** The second spatial derivative of the image is termed Laplacian. The response of a Laplace filter on the image produces zero crossings at edge locations and is typically applied in edge detection tasks. For the application in RGB color histograms, the Laplacian is calculated based on the chromaticities using

$$f_{\nabla^2_R} = R'_{xx} + R'_{yy}\ ,\ f_{\nabla^2_G} = G'_{xx} + G'_{yy}.$$

HSV color space

The HSV color space is a re-organization of the RGB color space, which is closer to human color perception. Colors in the HSV space are represented with the three components hue, saturation and value. The hue component H defines a perceptual color and the value component V describes the intensity of the color. The second component S encodes the saturation of the color.

The following channels are calculated from the HSV representation:

- **Hue:** For the hue component H the corresponding channel is calculated with the function f_H. Since hue is not defined for colors with zero saturation or value, only colors with $S > S_{min}$ and $V > V_{min}$ are considered in the calculation of f_H.
- **Gradient and Laplacian:** The channels for gradient and Laplacian of the hue component are calculated analog to the RGB color space and are denoted by f_{∇_H} and $f_{\nabla^2_H}$.

The proposed channels allow for several alternative image representations. The descriptors introduced in the following make use of several combinations of the presented channels. In Section 5.1.3 an evaluation with different combinations of channels is provided in order to identify the best suitable set for the detection of object candidates.

5.1.2 Descriptors for Object Candidate Detection

Descriptors for the visual appearance of objects either fall into the class of geometrical methods or into the class of statistical methods. Geometrical descriptors use properties of geometrical primitives in order to derive a representation of the image while statistical methods analyze frequencies of pixel properties. The decision for a suitable descriptor of object appearances depends on the desired application. The application in the detection of object candidates in wide angle views motivates a descriptor which can deal with objects that only cover a small region in the query image. The extraction of geometrical primitives in these regions is difficult since only few pixels are available. Thus, a statistical approach is proposed for the detection of object candidates in the peripheral images. The statistical analysis in the proposed approach is based on the image channels discussed in Section 5.1.1 which comprise color components and their spatial derivatives. In the following, statistical methods for the representation of images are discussed and the LCCH and NLCCH descriptors are proposed and evaluated in object detection tasks.

Statistical Descriptors

The first application of statistical methods as descriptors for color images, the color histogram, was proposed in [Swain and Ballard, 1991]. The color histogram is a first order statistical measure which records the frequency of colors within the image in a one-dimensional array, thus representing an image by the statistical distribution of colors.

One major drawback of color histograms is their low specificity which results from the omission of spatial information. In order to improve the specificity, the color histogram descriptor has been augmented by properties of the image texture derived from the spatial arrangement of colors [Haralick et al., 1973]. Therefore, the distribution of pairs of colors and their spatial relation is encoded in so-called color co-occurrence descriptors. These descriptors are a second order statistical measure, augmented with the spatial domain. The feasibility of co-occurrence descriptors for the representation of textures has first been demonstrated in [Haralick et al., 1973]. In their work, the frequency of co-occurring intensities is recorded and represented with the Spatial-Dependence Matrix M. The spatial relationship between pairs of co-occurring intensities is defined by their distance in pixels d and their angle a, which is quantized to 45° intervals as depicted in Fig. 5.2, left. For each distance d up to a maximum distance d_{max} and for each angle a, a Spatial-Dependence Matrix $M(d, a)$ is calculated. The entries $m_{ij}(d, a)$ of the matrix $M(d, a)$ represent frequencies of quantized co-occurring intensities i and j with the spatial relation encoded by d and a. The resulting set of matrices $\mathbb{M} = \{M(d_0, a_0), \ldots, M(d_n, a_m)\}$ is processed by a number of feature extractors in order to identify similar textures.

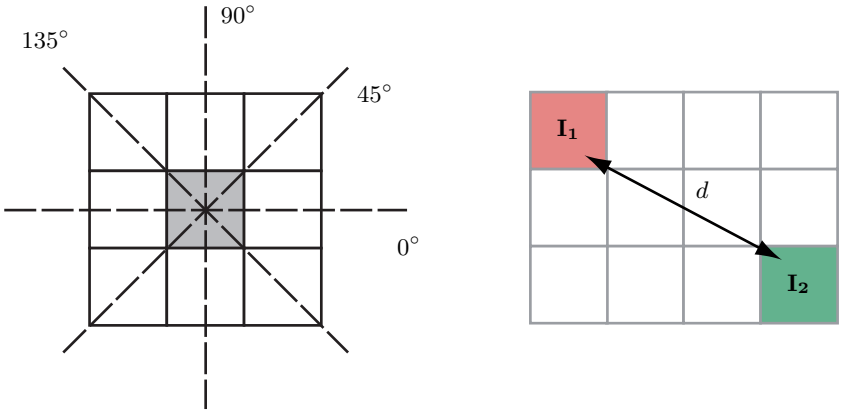

Fig. 5.2. Calculation of co-occurrences. Left: Neighborhood for the calculation of the Spatial-Dependence Matrix. The direction is quantized to 45° intervals. Right: Co-occurrences of colors $\mathbf{I_1}$ and $\mathbf{I_2}$ are considered if the Euclidean distance d is smaller then a maximum distance d_{max}.

The application of color co-occurrences for object recognition has been proposed in [Chang and Krumm, 1999]. Instead of using pixel intensities as underlying representations, co-occurring colors are considered. Let $\mathbf{I} = \mathbf{I_{rgb}}$ be the representation of a pixel in RGB color space. The frequency of a pair of colors $\mathbf{I_1}, \mathbf{I_2}$ is measured as a function of their distance d using

$$ch_c(\mathbf{I_1}, \mathbf{I_2}, d). \tag{5.1}$$

Chang et al. omit the directional component of the spatial relation between co-occurring colors in order to allow invariance toward rotations in the image plane. In order to build compact histograms, the colors are quantized using k-means clustering with s clusters. Further the distances are quantized using equally sized intervals resulting in t bins. Thus, the resulting co-occurrence histogram $H_c \in \mathbb{N}_0^{s \times s \times t}$ consists of entries

$$H_c = (h_{ijk}) \text{ with } h_{ijk} = ch_c(i, j, k), \tag{5.2}$$

where $i, j \in [1, s]$ are indices of co-occurring color clusters and $k \in [1, t]$ is the index of the distance interval.

In addition to the invariance toward image plane rotation, the invariance to object scaling is essential in real world recognition tasks. In [Chang and Krumm, 1999] the distance is encoded in the third dimension of the histogram H_c. Thus, the histograms of an object and the histogram of its scaled equivalent will differ significantly. In order to achieve robustness toward scaling, the distance d can be omitted as dimension of the histogram [Ekvall and Kragic, 2005]. Co-occurrences of pixels are recorded up to a distance d_{max} as depicted in Fig. 5.2, right. The authors could demonstrate an

increased robustness toward scaling using the proposed descriptor. Again, let i and j be indices of the clusters in the utilized color space. Then the frequency of co-occurring colors is calculated with

$$ch_e(i,j) \tag{5.3}$$

for pixels with spatial distances smaller then d_{max}. Using a quantization with overall s clusters, the frequencies form a two-dimensional histogram $H \in \mathbb{N}_0^{s \times s}$ with

$$H_e = (h_{ij}) \text{ with } h_{ij} = ch_e(i,j). \tag{5.4}$$

As underlying channels different combinations in RGB color space including first and second derivatives were evaluated in [Ekvall and Kragic, 2005]. Using the set of channels $C = \{c_1, \ldots, c_m\}$, k-means clustering is performed in the m-dimensional space of all considered channels.

The LCCH and NLCCH Descriptors

In this work the two-dimensional histogram defined in Eq. (5.4) is used in order to benefit from the image plane rotational invariance and the robustness toward scaling. The frequency of co-occurring colors with indices i, j can be formalized using

$$ch(i,j) = \left|\{\mathbf{p_1}, \mathbf{p_2} \in P | L_2(\mathbf{p_1}, \mathbf{p_2}) \leq d_{max}, q(\mathbf{p_1}) = i, q(\mathbf{p_2}) = j\}\right|, \tag{5.5}$$

where $\mathbf{p_1}$ and $\mathbf{p_2}$ are two points within the spatial domain P of the image, L_2 is the Euclidean distance and d_{max} is the maximum distance for the consideration of co-occurrences. The quantization function $q(\mathbf{p_n}) = i_n$ applies a quantization method to the color of the pixel at location $\mathbf{p_n}$ and returns the corresponding histogram bin index i_n.

The quantization method used for co-occurrence histogram calculations has an essential impact on the resulting representation. The method proposed in [Ekvall and Kragic, 2005] performs quantization in the space of all considered channels

$$q(\mathbf{p_n}) = q\begin{pmatrix} f_{c_1}(\mathbf{p_n}) \\ \vdots \\ f_{c_m}(\mathbf{p_n}) \end{pmatrix}. \tag{5.6}$$

The resulting receptive field co-occurrence descriptors (RFCH) not only covers co-occurrences within each channel but also co-occurrences between different channels resulting in an increased specificity. Nevertheless, this type of quantization implicitly assumes that a correlation between the different channels, e.g. normalized $R'G'$ values and their derivatives, exists. If the channels are largely uncorrelated the k-means clustering will either fail to produce a descriptive representation or will require a larger set of clusters.

In order to overcome this assumption, the descriptors for object candidate detection proposed in this work make use of quantization performed on single channels only. In order to encode co-occurrences between different channels based on such single channel quantization a $2m$-dimensional histogram would be required. For memory consumption and efficiency reasons co-occurrences between different channels are not encoded. Rather, a separate histogram is calculated for each channel. By concatenating these histograms from all channels, the descriptor is formed. In the following, two descriptors based on this principle are defined using different quantization methods. Both descriptors exhibit different properties in terms of specificity and computational complexity. An evaluation of the proposed descriptors and a comparison with the RFCH descriptor proposed in [Ekvall and Kragic, 2005] is provided in the context of object candidate detection in Section 5.1.3.

Linear Color Co-occurrence Histogram (LCCH)

A linear quantization of the image representation is performed in order to build the linear color co-occurrence histogram (LCCH). The quantization function subdivides the space into equally sized intervals

$$i_n = q_L(\mathbf{p_n}) = \left\lfloor \frac{f_c(\mathbf{p_n})}{b} \right\rfloor, \tag{5.7}$$

where b is the size of the intervals and i_n is the index of the histogram bin. The image is represented with the function $f_c(\mathbf{p_n})$ which extracts the property corresponding to channel c at position $\mathbf{p_n}$. For each channel the co-occurrence histogram has a size of $s \times s$ where s is the number of bins used for quantization.

The LCCH descriptor consists of a set of histograms $\mathbb{H}_{L,C}$ where each histogram H_{L,c_t} is defined over a channel from $C = (c_1, \ldots, c_m)$. Thus, the size of the complete LCCH descriptor for m channels is $s^2 m$.

Non-Linear Color Co-occurrence Histogram (NLCCH)

The non-linear color co-occurrence histogram (NLCCH) involves non-linear quantization. For each considered channel c k-means clustering is performed resulting in cluster centroids $E = (e_1, \ldots, e_s)$. The bin index is then determined by the minimum Euclidean distance to the cluster centroids

$$i_n = q_N(\mathbf{p_n}) = \underset{w}{\operatorname{argmin}}(|L_2(e_w, f_c(\mathbf{p_n}))|). \tag{5.8}$$

As in the linear case, a set of histograms is calculated from all considered channels C resulting in a set of histograms $\mathbb{H}_{NL,C}$ with elements H_{NL,c_t}. Further, the cluster centroids E_c are stored for each considered channel. For each centroid $e_{c,t}$ the variance $\sigma^2_{c,t}$ is calculated from the training data and stored together with the descriptor.

Fig. 5.3. Object candidate detection in the peripheral cameras. Left: Views of the target objects are trained and matched with the scene. Right: Typical scene for detection queries.

Descriptor Matching

In order to determine the similarity between two images, the corresponding LCCH or NLCCH descriptors have to be matched. The descriptor matching for object detection involves two representations: the representation of the object to be detected in the scene and the representation of a subregion of the scene for comparison with the object. The goal is to determine the similarity of objects and subregions of the scene. While the representations of searched objects are kept in a database and are constant over several scenes and search tasks, the scene is changing constantly with each eye movement. In the following, the feature extracted from the subregion of the scene is denoted with F while the feature of the object used as query is denoted with F_q. Examples of object views used for query feature extraction are illustrated in Fig. 5.3, left. An exemplary scene is shown in Fig. 5.3, right.

LCCH Matching

Using linear quantization, the histogram of the scene subregion and the object can be calculated independently from each other. The linear quantization function from Eq. (5.7) is used for the calculation of frequencies as defined in Eq. (5.5). The feature vectors F and F_q are directly derived from the LCCH descriptor $\mathbb{H}_{L,C}$ using the following definition

$$F = (H_{L,c_1} \ \ldots \ H_{L,c_m}), \tag{5.9}$$

where each co-occurrence histogram in $H_{L,c}$ is represented as a row vector. The resulting feature vectors are then matched using a histogram metric.

NLCCH Matching

If non-linear quantization is deployed, the scene and the object histograms cannot be calculated independently. As described above, a representation of the query object is derived by performing k-means clustering on the considered channels. The matching procedure between query object and a subregion of the scene involves applying the same quantization to the image representation of the scene.

Given the cluster centroids E_c from the NLCCH descriptor of a query object, all pixels within the subregion of the scene are assigned to the corresponding cluster for each channel c. Pixels are only considered if the distance to the cluster centroid $e_{c,t}$ is less then $\beta\sigma_{c,t}^2$. Formally, for a pixel n and the channel c the bin index i_n is calculated using the rule

$$i_n = \begin{cases} \underset{t=1,\ldots,s}{\operatorname{argmin}}(L_2(e_{c,t}, f_c(\mathbf{p_n})), & \text{if } L_2(e_{c,t}, f_c(\mathbf{p_n})) < \beta\sigma_{c,t}^2 \\ -1, & \text{otherwise} \end{cases}. \tag{5.10}$$

Using the indices resulting from the above rule, the histograms $H_{NL,c}$ for each channel are again concatenated as row vectors to form the feature

$$F = (H_{NL,c_1} \ \ldots \ H_{NL,c_m}). \tag{5.11}$$

Similarity Measure

The literature offers a variety of similarity measures for histograms including geometric, information theoretic, and statistical measures [Liu et al., 2008]. One of the most commonly used statistical measures is the histogram intersection method proposed in [Swain and Ballard, 1991]. The histogram intersection calculates the normalized sum of the minimum bin entries for all components of the features F and F_q using

$$d(F, F_q) = 1 - \frac{\sum_{i=1}^{n}(min(F(i), F_q(i))}{min(|F|, |F_q|)}, \tag{5.12}$$

where $F(i)$ is the number of elements in the bin i of the feature and $|F|$ the overall number of entries in all bins. The normalization by the number of overall entries assures that the histogram intersection lies within the interval $[0, 1]$. A desirable property of the histogram intersection is the ability to also match partial histograms. If a region is considered which contains a significant part of the query histogram the measure will still report a significant similarity. The histogram intersection is used for LCCH as well as for NLCCH feature matching.

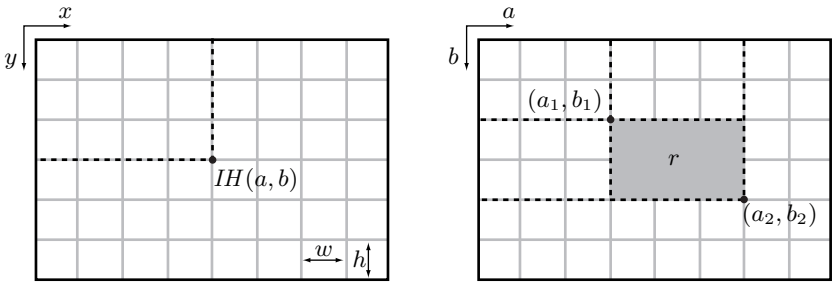

Fig. 5.4. CCH matching procedure. Left: The integral histogram is calculated on a regular grid for regions starting at the upper left corner. Right: For a rectangular subregion of the grid the histogram can be efficiently determined using the integral histogram representation.

5.1.3 Object Candidate Detection

Object detection is performed in the images of the peripheral cameras based on the proposed LCCH and NLCCH color co-occurrence descriptors. The goal is to identify regions within an image which have significant similarity with the representation of the query object. Since the spatial extent and the location of object candidates within the image are not known, the image has to be scanned at different scales and different locations. Chang and Krumm propose to use a search window of fixed size at discrete spatial locations in order to determine the match between the region represented by the window and the object histogram [Chang and Krumm, 1999]. In order to allow invariance for different scalings of the object, Ekvall and Kragic proposed a vote matrix for overlapping search windows which collects evidence for object candidates [Ekvall and Kragic, 2005]. In the following, the work by Chang and Krumm and the further development of the detection approach by Ekvall and Kragic is outlined and extended in terms of the extraction of multiple object candidates. Subsequently, the detection performance is evaluated in typical search tasks using different variants of the LCCH, the NLCCH, and the RFCH descriptor.

Object Candidate Detection Approach

In order to detect object candidates at different locations, the scene image is subdivided into windows with a spatial extent of $w \times h$ in a regular grid as depicted in Fig. 5.4, left. The search for object candidates is performed on rectangular subregions of the image defined by one or multiple cells of the grid. For fast computation of the CCH representation for subregions, the integral histogram is calculated. The integral histogram method is based on the integral image representation as proposed in [Crow, 1984].

Let (a, b) be the index of a cell in the grid with the spatial location $x = aw$ and $y = bh$. The integral histogram $IH(a, b)$ represents the histogram of a region starting from the upper left corner with index $(0, 0)$ to the cell with index (a, b) as depicted in Fig. 5.4, left. Further, let $H(x, y)$ be the color co-occurrence histogram of the pixel at location (x, y). Then the integral histogram can be formalized using

$$IH(a,b) = \sum_{\substack{(x' \leq aw) \\ \wedge\, (y' \leq bh)}} H(x', y'). \tag{5.13}$$

The integral histogram can be efficiently calculated for the complete image in one pass. Based on the integral histogram, an arbitrary rectangular region within the grid can be calculated with only four operations as described in the following.

Let r be a region defined by the indices of the upper left corner (a_1, b_1) and the lower right corner (a_2, b_2) as illustrated in Fig. 5.4, right. Then the CCH descriptor of the region r can be calculated from the integral histogram representation using

$$CCH(r) = IH(a_2, b_2) + IH(a_1, b_1) - (IH(a_1, b_2) + IH(a_2, b_1)). \tag{5.14}$$

In order to search for a query object, histograms over regions calculated by the above equation are compared to the object histogram using histogram intersection as described in the previous section. Thereby, only regions are considered if their width and height lies within a specified range in order to restrict the search space to reasonable scalings of the object. The result of the comparison is a list of regions R, where each region is associated with its spatial extend defined by $x_{min}, y_{min}, x_{max}$ and y_{max}. The spatial extent in pixels is recovered from the cell indices. Additionally, for each region r_s in the list R the result of the histogram intersection ψ is stored:

$$R = \{r_1, \ldots, r_n\} \text{ with } r_s = (x_{min}, y_{min}, x_{max}, y_{max}, \psi). \tag{5.15}$$

The list R not only contains regions corresponding to the query object but also regions which only contain parts of the object and major parts of other objects or of the background. Hence, the object candidate detection algorithm processes the region list R in order to generate a minimum set of regions X where in the optimal case each region describes the spatial extent of one object candidate in the scene.

Algorithm 1: Calculation of object candidate regions

input : A list of regions R
output: Regions of detected object candidates X

```
while Size(R) > 0 do
    V ←CalculateVoteMatrix(R);
    r ←FindMaximumEntry(V);
    r ←GrowRegion(V,r);
    Append(X,r);
    R ←RemoveOverlapping(R,r);
return X;
```

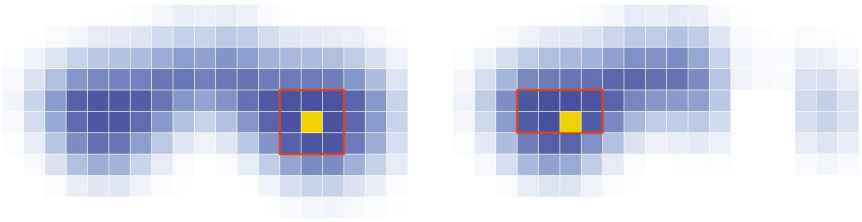

Fig. 5.5. Illustration of the object candidate detection algorithm. Left: In each iteration, the vote matrix (blue) is calculated based on the list of regions determined by the matching procedure. The maximum of the vote matrix (yellow) is calculated and the spatial extent is determined using region growing on the vote matrix. Right: In the next iteration all regions that overlap with the detected region are removed. This procedure is repeated until no regions are left.

Algorithm 1 outlines the approach for calculation of object candidate regions. Input is the set of processing regions R, calculated as described above. Each region is accompanied with the histogram intersection match ψ. In each iteration, a vote matrix is calculated from processing regions in the set R by accumulating the matches ψ of all regions that overlap with an entry of the vote matrix as proposed in [Ekvall and Kragic, 2005]. A typical vote matrix is depicted in Fig. 5.5, left. In order to identify object candidates in the image, the entry r with maximum match in the vote matrix is determined. Using this entry as seed point, a region growing method is deployed which extends r by adjacent regions if the overall match of the new region does not drop below p_{grow} percentage of the maximum entry's match. The new region resulting from growing based on the vote matrix is appended to the list of candidate regions X.

In order to detect all candidates in the scene, regions overlapping with the previously calculated candidate region r are removed from the list of processing regions R. With the updated list of processing regions, a new iteration is started until the processing list is empty. The first two iterations of the algorithm are depicted in Fig. 5.5. In iteration one, the maximum of the vote matrix and the region r illustrated in the left image are calculated. Removing all regions from the processing list that overlap with r results in the updated vote matrix depicted in the right image. Based on the updated vote matrix, a new maximum is determined. Using region growing, a new candidate region is calculated.

While Algorithm 1 iterates until no regions are left in the processing list, for the application for object candidate detection only regions are considered which have an activation higher the p_{region} percentage of the best matching candidate region.

Evaluation of Detection Performance

Having introduced the complete object candidate detection procedure - from low-level image representation over descriptors to the detection algorithm itself - allows to evaluate all proposed components in the context of object candidate detection. First, sets of channels and their applicability using the LCCH and NLCCH descriptors are evaluated. Subsequently, the invariance of different descriptors under illumination change and scaling of the target object are investigated. Based on the best performing combination, the detection performance using the proposed algorithm is evaluated.

Descriptor Evaluation

For the evaluation of the proposed descriptors three scenes containing a number of different objects were considered. For each scene, ten target objects were selected and cut out of the scene manually. Further, the scene was annotated with rectangular regions corresponding to the objects as ground truth. Each scene was scaled to 320×240 pixels. In the matching procedure, a grid with a cell size of 8×8 pixels was used for the integral histogram calculation. The minimum and maximum size of extracted regions were set to 2×2 and 16×16 grid cells, respectively. For all experiments, co-occurrences up to a distance of $d_{max} = 1.6$ were considered. Thus, all eight neighbors of a pixel were taken into account during the calculation of co-occurrences. The number of clusters used for the descriptors largely influences the quality of matches and are subject to evaluation.

In order to judge the performance of the descriptors their ability to identify a target object within a scene has to be quantified. Therefore, the query

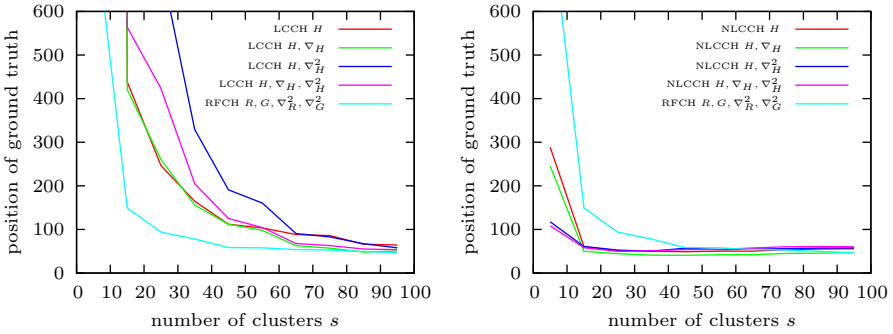

Fig. 5.6. Evaluation of the CCH descriptors in a search task comprising 30 objects in 3 different scenes. The evaluation is performed for different numbers of clusters s. From the ordered list of reported regions, the position of the region which best fits the ground truth is used as quality measure. The results for the RFCH descriptor are given for comparison.

histogram extracted from the target object is matched with all regions R generated using the integral histogram approach. The list of regions R is sorted by descending match quality. The region within this list which best matches the ground truth is determined using the Hausdorff distance

$$d_H(r_1, r_2) = \max\{ \sup_{\mathbf{p_1} \in r_1} \inf_{\mathbf{p_2} \in r_2} L_2(\mathbf{p_1}, \mathbf{p_2}), \sup_{\mathbf{p_2} \in r_2} \inf_{\mathbf{p_1} \in r_1} L_2(\mathbf{p_1}, \mathbf{p_2}) \}, \quad (5.16)$$

where r_1 and r_2 are the detected region and the ground truth region, respectively. The points $\mathbf{p}$ are positions on the contour of the regions. In order to derive a quality measure, the position of the region in the region list R with minimal Hausdorff distance d_H is drawn on.

In Fig. 5.6 the position in the result list of the ground truth region for different descriptors is illustrated. Only channels defined on the hue components were considered for the LCCH and NLCCH descriptors, since the HSV color space allows for a more compact representation of color. The combinations of channels include the hue component H, the hue component accompanied with the hue gradient H, ∇_H, the hue component accompanied with the Laplacian H, ∇_H^2, and the combination of the three channels H, ∇_H, ∇_H^2. For both the LCCH and the NLCCH plot the results achieved using the RFCH approach are provided for comparison. We chose a combination of the normalized R and G components accompanied with their Laplacian ∇_R^2 and ∇_G^2. This combination proved to be the best trade-off between matching performance versus number of channels as demonstrated in [Ekvall and Kragic, 2005]. For all descriptors, different numbers of clusters s were evaluated in order to determine the optimal size of the descriptor.

The LCCH results as illustrated in Fig. 5.6, left show that by using the RFCH approach the position of the ground truth rapidly improves with increasing number of clusters. Using 25 clusters the ground truth region lies within the 100 best regions calculated by the approach. The linear quantization of the LCCH descriptor results in an increase of clusters required to achieve the same result. Furthermore, the plot shows that including the Laplacian as channel more clusters are required in order to achieve a good result.

Replacing the linear quantization with a non-linear quantization significantly improves the performance of the descriptors. For the evaluation of the NLCCH a threshold distance for the assignment to clusters according to Eq. (5.10) of $\beta = 2$ was used. As depicted in Fig. 5.6, right the NLCCH descriptor reports the ground truth at a position of approx. 50 using only 15 clusters. Any combination of channels evaluated for the NLCCH outperforms the RFCH descriptor. The performance slightly degrades with an increasing number of clusters due to over fitting to the training data.

While the LCCH descriptor is constantly outperformed by the RFCH and NLCCH descriptors, the LCCH descriptor is still appealing due to its computational simplicity. In contrast to non-linear quantization, the linear quantization defined in Eq. (5.7) does not require to match values from each channel to cluster centers allowing a rapid calculation of LCCH descriptors. The LCCH descriptor offers advantages when multiple views of a target object or multiple target objects are considered during search. While the NLCCH and RFCH approaches require the expensive generation of integral histograms for each query feature, the LCCH descriptor requires only the calculation of one integral histogram. Comparing an NLCCH with 25 clusters and an LCCH descriptor with 40 clusters which both provide satisfactory performance, the LCCH allows faster matching when searching for more then three objects or views.

Evaluation of Invariances

In order to evaluate the behavior of the different descriptors under changing illumination, the previous experiment was repeated for 10 objects and one scene with changing lighting conditions. The lighting conditions were not controlled. Rather three different lighting conditions were captured resulting from changing illuminations in the lab during the day.

In Fig. 5.7 the results of the experiment based on changed illumination are illustrated. While the performance does not vary significantly between different combinations of channels for the LCCH and NLCCH descriptor, the performance of both descriptors increases for changed lighting conditions compared to the RFCH descriptor. This is the expected behavior when using the HSV color space, since the hue component is lesser affected by changes in light intensities than the components of the $R'G'$ color space.

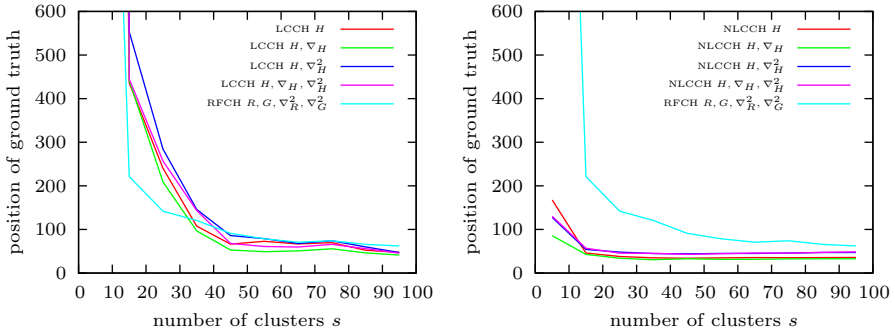

Fig. 5.7. Evaluation of the invariance to illumination changes. 10 Objects in one scene captured with 3 different lighting conditions were used. Results are shown for the LCCH and NLCCH descriptors for different number of clusters s. From the ordered list of reported regions, the position of the region which best fits the ground truth is used as quality measure. The results for the RFCH descriptor are provided for comparison.

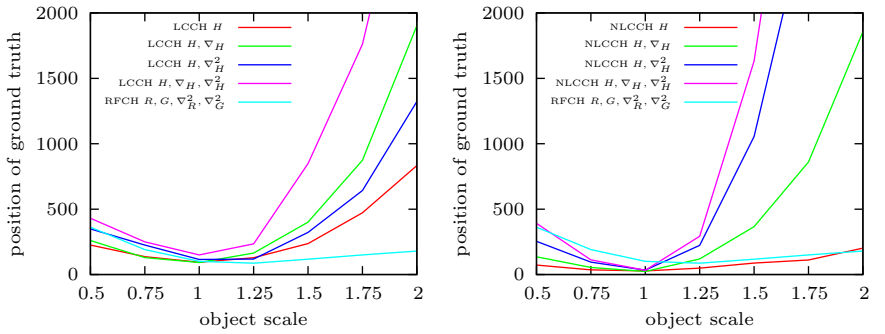

Fig. 5.8. Evaluation of the proposed CCH descriptors using different scalings of the target object. The position of the region which best fits the ground truth from the ordered list of matched regions is used as quality measure. The results for the RFCH descriptor are provided for comparison.

The invariance to rotation and scaling of color co-occurrence descriptors has already been evaluated in [Ekvall and Kragic, 2005]. While the invariance to rotation results from the definition of the CCH which makes no use of orientation, the invariance to scaling depends on the channels used for the calculation of the descriptor. In order to verify the suitability of the proposed channels for the LCCH and NLCCH descriptors, the invariance to scaling was evaluated by resizing the target object in several steps. All ten objects of one scene were used as test data. In Fig. 5.8 the results for target object scalings from 0.5 times to 2 times the size of the original object are illustrated. The behavior of LCCH and NLCCH descriptor is very similar, since the scaling mainly affects the underlying channels. The results indicate that gradient and

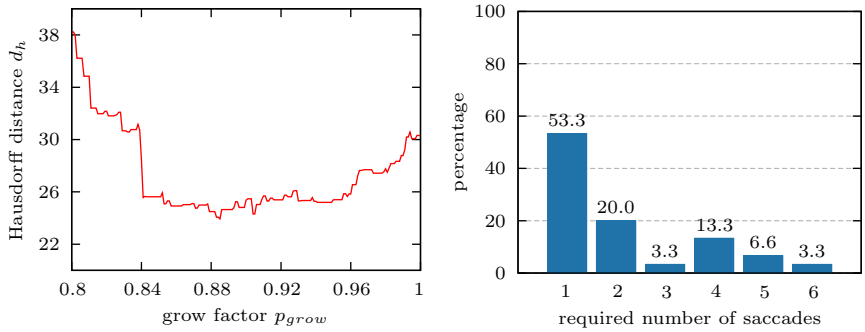

Fig. 5.9. Results of the detection performance evaluation. Left: The optimal region grow factor was determined using the Hausdorff distance between detected region and ground truth. Right: Using this factor about 53% of the object are reported as first match. All objects are reported within the first 6 regions.

Laplacian channels are largely affected by scaling which results in degrading performance of the matching procedure. The quantization method based on all dimensions used in the RFCH descriptor results in a better behavior in the scaling experiment.

Reconsidering all performed experiments, the hue channel proves to be the best choice for representing the image for color co-occurrence calculation. The experiments on changed lighting conditions show the advantage of using the HSV color space for the histogram calculation. Further, by only using the hue component invariance to scaling can be preserved. The NLCCH descriptor outperforms both the LCCH and the RFCH descriptor in all our experiments. The computational complexity is of the same order as the RFCH descriptor. Only for cases in which fast response is preferred over matching quality, e.g. when matching multiple query features, the LCCH descriptor provides a reasonable alternative due to its computational simplicity. For the object candidate detection experiments in the next section, only the NLCCH descriptor based on the hue channel with 30 dimensions was considered.

Evaluation of Object Candidate Detection

In order to evaluate the object candidate detection approach, again all 30 objects and 3 scenes were considered. The accuracy of the region detection depends strongly on the threshold p_{grow} since it directly influences the size of the extracted region. In order to evaluate the optimal threshold, the Hausdorff distance d_h between the reported regions and the ground truth was measured for different thresholds. In Fig. 5.9, left the results of the experiment are illustrated. The results show, that within a range of $p_{grow} = [0.84, 0.96]$ the mean Hausdorff distance for all query objects stays about constant. For all subsequent experiments, a threshold of $p_{grow} = 0.85$ was used.

Fig. 5.10. Calculation of stereo correspondences for object candidates. In each image of the stereo camera pair candidate regions are extracted using the CCH descriptor. Using the center of regions in the left image, the epipolar line in the right image is calculated (white). For regions in the right image that intersect the epipolar line, correspondences are calculated using correlation techniques (red).

The overall detection performance was evaluated by determining whether the reported object candidate regions overlap with the manually annotated ground truth regions. With respect to the active visual search approach the position of the first overlapping reported region in the list of object candidates corresponds to the number of saccades required before the target object is fixated. In Fig. 5.9, right a histogram of the number of required saccades is illustrated for all 30 objects. 53% of the objects are reported as first region in the list of object candidates. All objects from the test set were covered by the first 6 extracted regions in the list.

5.1.4 Calculation of Stereo Correspondences

In the previous sections the detection of object candidates in the peripheral camera image was discussed. The resulting regions describe the spatial extent of object candidates in the image plane. In order to determine the corresponding 3D location of the object candidate, both images of the peripheral stereo camera pair are considered. Thereby, the epipolar constraint is exploited in order to restrict the search space of possible correspondences within the stereo image pair.

The process of calculating correspondences is illustrated in Fig. 5.10. The object candidate detection approach processes both stereo images resulting in the regions marked by red frames. The calculation of 3D locations is performed using a two-step approach in order to reduce the computational complexity of stereo correspondence calculation. In the first step, corresponding regions in the left and right image are determined using the epipolar constraint. For

Fig. 5.11. The detailed view of the foveal camera allows for robust object recognition.

this purpose, the epipolar line for the center of each region in the left image is calculated in the right image. For each intersection of the epipolar line with an extracted region in the right image, a potential correspondence is reported. In a second step, these correspondences are verified using image correlation based on the zero mean normalized cross-correlation (ZNCC) (see [Aschwanden and Guggenbühl, 1992]). The ZNCC is calculated for different offsets between the center of left and right region. With the offset, inaccuracies of the stereo calibration and possible shifts along the epipolar line are accounted for. The correspondence is defined by the center of the left region and the pixel in the right image where the correlation has its maximum. The 3D location is determined from stereo calibration which is calculated based on the calibrated kinematic model of the head as discussed in Chapter 6.

5.2 Object Recognition in the Foveal Images

As illustrated in Fig. 5.11, the foveal cameras allow a more thorough inspection of scene fragments compared to the peripheral cameras. The extent that an object of interest covers within the image plane can be dramatically increased by fixation using the foveal cameras. While the object detection in the peripheral images produces candidates of objects using a coarse analysis of the scene, the foveal view allows for robust recognition of objects. Consequently, the goal of foveal object recognition consists in ascertaining whether an object candidate corresponds to the target object or not.

5.2.1 Features for Foveal Object Recognition

The small extent of the objects in the peripheral images motivated a global approach to object representation. The complete view of the object was encoded using a single statistical feature. For the purpose of object recognition, the information encoded in such a single global feature does not yield robust results. Hence, local features are used for object recognition in the foveal images. Local features have been shown to be well suited for recognition as they are robust to occlusion and background clutter. The process of local feature extraction can be divided into two separate phases: the detection of interest points and the representation of their properties in a feature descriptor. The literature offers a variety of solutions for both the detection of interest points and the feature descriptor. Typical requirements are invariance to rotation, scaling, and illumination and computational efficiency.

A minimum requirement for interest points is their invariance to shifting in the image plane. Two widely applied approaches in this class are the Shi-Tomasi interest points [Shi and Tomasi, 1994] and the Harris interest points [Harris and Stephens, 1988]. While the invariance to shifting in the Shi-Tomasi interest points is achieved by analyzing the tracking behavior, the Harris interest point detector analyzes the local neighborhood in order to identify salient pixels. In order to achieve invariance toward scaling, most common approaches make use of the scale space analysis. For this purpose, the image is processed at different scales and the interest points are detected in the scale space. Typical approaches calculate the difference of Gaussians filtered images at different scales (DoG,[Lowe, 1999]) or use the second derivative of scale space representations in order to identify interest points (Laplacian of Gaussian, LoG, [Lindeberg, 1998]). The Harris-Laplacian, a scale invariant extension to the Harris interest point detector has been proposed based on LoG [Mikolajczyk and Schmid, 2001]. In order to identify objects which are subject to rotation out of the image plane, the invariance to affine transformation can be considered during interest point detection. An example of an affine invariant interest point detector is the Harris affine detector proposed in [Mikolajczyk and Schmid, 2004]. A comprehensive review of different interest point detectors is provided in [Schmid et al., 2000].

In order to match the query against previously observed interest points, an ideally unique descriptor has to be calculated for each interest point. Such descriptors can be roughly divided into statistical methods and geometric methods. Statistical methods build statistics over the local neighborhood of the interest point. The statistics capture intensities (e.g. SIFT [Lowe, 1999], SURF [Bay et al., 2006]) and color properties as it is the case for the CCH descriptor introduced in Section 5.1.2. In geometrical approaches, the descriptor comprises geometric properties of the region, e.g. the contour (Local Affine Frames, LAF, [Obdrzálek and Matas, 2002]).

For the recognition of objects in the foveal cameras invariance and computational efficiency are required. The Harris-SIFT approach [Azad et al., 2009] combines the robustness of the SIFT descriptor with the computational efficiency of the Harris interest point detector. Instead of analyzing the scale space for each observed image in order to achieve scale invariance, Azad et al. propose to extract Harris interest points at different scales accompanied with a SIFT descriptor for each corner point. Combined with the recognition framework presented in [Azad, 2008] the approach allows for efficient and robust recognition of textured objects. In the following sections, the fundamental components of the Harris-SIFT approach are introduced. The detection of interest points using the Harris corner detector as well as the SIFT descriptor of local image patches are described. Subsequently, in Section 5.2.2 object recognition based on the Harris-SIFT features is discussed.

Harris Corner Detector

In order to determine salient pixels in the image, Moravec proposed to use the sum of squared differences (SSD) of an image patch and patches shifted by an offset $(\Delta u, \Delta v)$ using

$$\epsilon(\Delta u, \Delta v) = \sum_{(u,v)} w(u,v)(I(u+\Delta u, v+\Delta v) - I(u,v))^2, \tag{5.17}$$

where $w(u,v)$ defines the image patch [Moravec, 1980]. Thereby, offsets corresponding to the horizontal, vertical and diagonal neighbors were used. The basic idea consists in determining how shifting affects the image patch. Shifting along a line or within a uniform area will produce no significant error in ϵ. On the other hand, shifts will produce significant errors, if the patch is cluttered and differs from surrounding patches. Consequently, the cornerness is defined as the minimum SSD $\epsilon(\Delta u, \Delta v)$ produced by shifting the patch in all considered directions. If this value is maximal in a local neighborhood, the point (u,v) is used as interest point.

Further evaluation of the cornerness measure revealed its anisotropicity. Only shifts to the three main directions are considered, similarities with neighbored patches which result from shifts in other directions are not covered. Consequently, an isotropic cornerness measure which considered all local shifts of the patch hass been proposed in [Harris and Stephens, 1988]. The cornerness measure is again derived from the sum of squared differences of patches, as defined in Eq. (5.17). The intensity of a pixel shifted by $(\Delta u, \Delta v)$ can be locally approximated using the Taylor expansion

$$I(u+\Delta u, v+\Delta v) \approx I_u \Delta u + I_v \Delta v + I(u,v), \tag{5.18}$$

where I_u and I_v are the partial derivatives of I at the position (u,v).

Fig. 5.12. Harris corner points extracted in the foveal camera image with a minimum cornerness of $C_{min} = 0.001$.

Substituting the Taylor expansion from Eq. (5.18) in Eq. (5.17) yields

$$\epsilon(\Delta u, \Delta v) \approx \sum_{(u,v)} w(u,v)(I_u \Delta u + I_v \Delta v)^2, \tag{5.19}$$

which can be rewritten as

$$\begin{aligned}\epsilon(\Delta u, \Delta v) \approx & \quad \Delta u \ \Delta v \begin{pmatrix} \sum\limits_{(u,v)} w(u,v) I_u^2 & \sum\limits_{(u,v)} w(u,v) I_u I_v \\ \sum\limits_{(u,v)} w(u,v) I_u I_v & \sum\limits_{(u,v)} w(u,v) I_v^2 \end{pmatrix} \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} \\ \approx & \quad \Delta u \ \Delta v \, S \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix},\end{aligned}$$

where S is called the structure tensor. In order to derive a measure for the cornerness on the basis of the structure tensor, Harris et al. examined the eigenvalues λ_1, λ_2 of S. Based on these eigenvalues, the cornerness measure is defined using

$$C = det(S) \quad k \, trace(S)^2. \tag{5.20}$$

The measure C describes the curvature of the local image patch [Noble, 1988]. Corners are reported if the measure is above a threshold $C > C_{min}$. In Fig. 5.12 an example of the Harris interest points extracted on a typical foveal camera view with $C_{min} = 0.001$ is illustrated.

SIFT Descriptor

The SIFT descriptor has been proposed in the context of the scale invariant feature transform (SIFT) approach for textured object recognition [Lowe, 1999, Lowe, 2004]. For the recognition of objects in the foveal cameras only the SIFT descriptor is used in conjunction with the Harris corner detector, as proposed in [Azad et al., 2009].

The SIFT descriptor is calculated based on the gradient magnitude $m(u, v)$ and orientation $\theta(u, v)$ for each pixel within a region around the corner point using

$$m(u,v) = \sqrt{I_u^2 + I_v^2}$$
$$\theta(u,v) = arctan\frac{I_v}{I_u}.$$

In order to achieve invariance to rotation in the image plane, dominant orientations within a region centered at an extracted corner point are determined. For this purpose, a histogram of orientations $\theta(u, v)$ for all pixels within the region is calculated. The histogram is built in a linear manner, orientations are quantized to bins of 10 degree size. The vote of each pixel is weighted with the gradient magnitude $m(u, v)$ and with an isotropic Gaussian kernel centered at the corner point. Dominant orientations are determined by calculating the maximum over the 36 bins. In addition, bins with an activation of 80% of the maximum are considered as dominant orientations. For each of the dominant orientations, a polynomial of second order is fit to the corresponding entry and its two neighbors in order to refine the position of the peak.

The descriptor is calculated from the region which is re-sampled according to the dominant orientations, for each corner point and each dominant orientation. For a region of width w, the gradient magnitudes are again weighted with an isotropic Gaussian kernel with standard deviation $\sigma = \frac{w}{2}$. A region of width $w = 8$ is illustrated in Fig. 5.13, left. The region is again subdivided into subregions. For each subregion a histogram of gradient orientations and weighted magnitudes is calculated with 8 bins for the orientation as depicted in Fig. 5.13, right. For illustration, the region was subdivided into 2×2 subregions for the calculation of histograms. An optimal size of the histogram was determined experimentally with 4. Using 8 bins for orientations, the SIFT descriptor is formed by the gradient magnitudes from the $4 \times 4 \times 8 = 128$ histogram bins.

In order to achieve invariance to changes in image contrast, the resulting feature vector is normalized to unit length. Non-linear illumination changes are accounted for by thresholding all entries above 0.2.

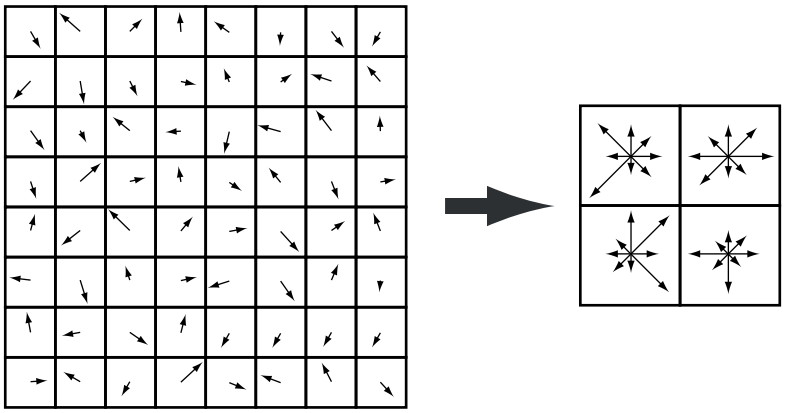

Fig. 5.13. In order to calculate the SIFT descriptor a histogram of gradient orientations weighted by their magnitudes is calculated over an image patch. The patch is divided into subregions. For each subregion the directions are quantized. The magnitude of the resulting 8 directions is used to build the SIFT descriptor.

Harris-SIFT Feature Extraction

The SIFT descriptor has been proposed in conjunction with a scale space analysis, based on maximum detection in the DoG representation. Since the scale space analysis is computational expensive, the goal in the Harris-SIFT approach consists in replacing the interest point detector based on the scale-space with the Harris corner detector [Azad et al., 2009]. In order to re-incorporate invariance toward scaling, the robustness of the SIFT descriptor toward scaling was evaluated. From the evaluation a robust response for images scaled by about 10-15% could be shown. Consequently, this robustness is exploited in order to extract SIFT detectors at different scales without generating gaps in the scale coverage. A relative scale factor of $\Delta s = 0.75$ is proposed in order to provide sufficient robustness toward scaling. SIFT descriptors are extracted at five different scales i which are calculated by scaling the input image with $(\Delta s)^i$.

5.2.2 Object Recognition

For the object recognition in the foveal camera, SIFT descriptors resulting from Harris corner points of trained objects are matched with descriptors extracted at Harris corner points in the foveal image. For this purpose, the correspondence between database and scene feature vectors is determined using an Euclidean classifier. If the match quality is above the threshold e_{match}, the corner points associated to the considered descriptors are considered as match. In

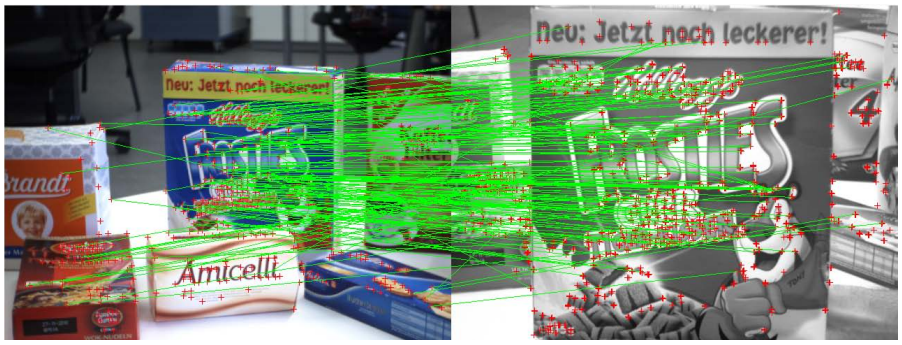

Fig. 5.14. Correspondences between object and scene resulting from matches of Harris-SIFT features using the Euclidean distance. The matching procedure generates several outliers.

Fig. 5.14 correspondences calculated by Euclidean classification between scene and database features are illustrated with green lines. Apparently, among a large amount of correct correspondences the matching generates many outliers which are not consistent in terms of related spatial locations in database and scene image. In order to sort out invalid correspondences the spatial relation between detected corners is exploited using the Hough transform as described in the following.

The Hough transform was introduced first in [Hough, 1962] for the detection of 2D shapes in images. The extension to the generalized Hough transform has been proposed in [Ballard, 1981]. In general, the Hough transform involves two essential steps: Hough space voting and maximum calculation. While in the voting step, properties of extracted primitives such as pixels or edges are used to activate entries in the Hough space by voting, the maximum calculation step determines the bin with the maximum vote which represents the maximum accordance between all primitives. In the following, the Hough transform for the recognition of the object position from oriented corner features as proposed in [Azad et al., 2009] is reviewed briefly.

The Hough space is a two dimensional voting matrix where the coordinates of the entries correspond to spatial locations (u, v) in the image plane. The quantization is performed in a linear manner with a quantization factor $r = 1/p$ where p is the number of pixels per bin. For each matched pair of corner points, the mapping of the origin of the database object within the scene is voted for in the Hough space. For matching pairs of corner points (u, v) and (u', v') with orientations ϕ and ϕ' the corresponding entry in the two-dimensional Hough space (u_k, v_k) is calculated using

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix} = r \left[\begin{pmatrix} u \\ v \end{pmatrix} \quad s_k \begin{pmatrix} \cos \Delta\phi & \sin \Delta\phi \\ \sin \Delta\phi & \cos \Delta\phi \end{pmatrix} \begin{pmatrix} u' \\ v' \end{pmatrix} \right], \tag{5.21}$$

Fig. 5.15. Purged correspondences after applying the Hough transform to the matches. The illustrated matches correspond to the bin in Hough space with maximum vote.

where s_k represent the considered scalings of the object as introduced in the previous section and $\Delta\phi = \phi \quad \phi'$ is the relative orientation difference derived from the dominant orientation. In order to achieve robustness toward scaling five different scales s_k are considered.

After voting in Hough space with all corresponding pairs, the maximum vote over all bins is calculated. Correspondences which fall into the maximum bin are considered valid, while all other correspondences are considered outliers. In Fig. 5.15 the correspondences that survive the Hough space voting are illustrated. The number of inconsistent matches can be significantly reduced using the Hough transform resulting in reliable recognition.

In order to filter out cases where only few matches have been determined, a minimum percentage of object features p_m has to be associated with the maximum bin. Furthermore, in order to specify the quality of the match, a score is calculated based on the features that are associated with the maximum bin. Let n_m be the number of features that fall into the maximum bin and n_o be the number of spatially distinct object features, then the score for the object match is defined with

$$s_f = \frac{n_m}{n_o}. \tag{5.22}$$

5.3 Summary

In this chapter, the approaches for the detection of object candidates in the peripheral camera pair and the recognition of objects in the foveal camera pair were introduced. In order to restrict the search space and thus reduce

the number of saccades required for object recognition, a coarse analysis is performed based on the peripheral images. Thereby, object candidates are detected based on a statistical descriptor of neighbored colors. Two descriptors, the LCCH and the NLCCH descriptor, which offer different properties in terms of specificity and computational expense, have been proposed and evaluated. The matches between target object and scene are established based on the integral histogram which determines the corresponding region at different scales. The approach for object candidate detection makes use of a vote matrix in order to determine multiple hypotheses of regions.

The detailed view of the foveal images allows for robust recognition of the target objects. For this purpose, an approach based on local features has been drawn on. The combination of Harris-SIFT features and Hough space voting allows to reliably recognize textured objects.

The calculation of 3D locations from the object candidates in left and right image is based on epipolar geometry. A necessary prerequisite for the calculation of stereo correspondences is the knowledge of the extrinsic parameters of the camera system. These parameters are derived from the kinematic calibration as discussed in the next chapter.

6

Calibration and Saccade Execution

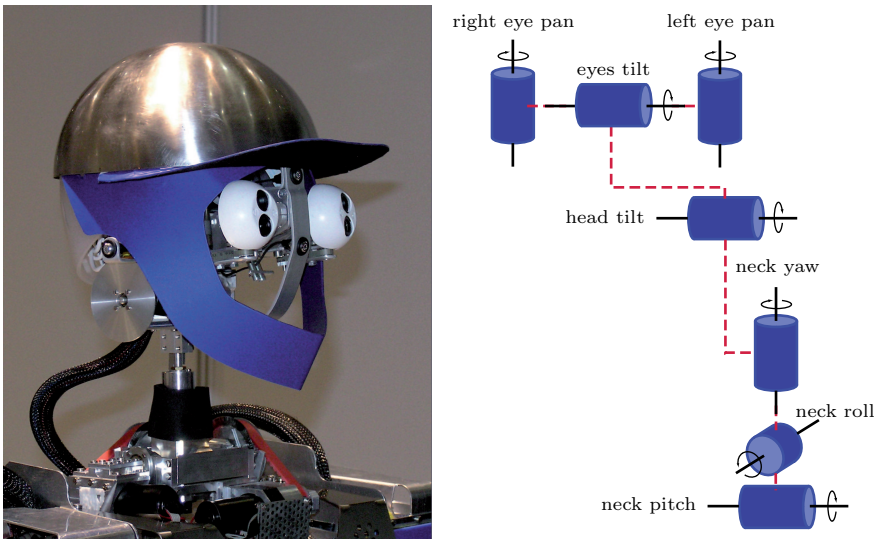

Fig. 6.1. The Karlsruhe Humanoid Head allows to adapt the gaze of the peripheral and foveal camera systems using seven degrees of freedom: three neck joints, one head tilt joint, one common eye tilt joint, and separate pan joints for each eye.

The Karlsruhe Humanoid Head [Asfour et al., 2008] offers seven degrees of freedom (DoF) which allow to actively control the gaze direction of the system. As depicted in Fig. 6.1, the head is actuated with 3 DoF in the neck, a head tilt joint, a common tilt joint for both eyes, and individual pan joints for each eye. The independent eye pan joints allow to separately control left and right cameras in order to direct the gaze of both eyes to relevant parts of the scene.

In the course of an active visual search task, different parts of the scene are brought into the view of the foveal cameras in order to perform robust object recognition based on the detailed views. This is accomplished by executing saccadic eye movements. As already discussed in Section 2.1.2, saccadic eye movements are ballistic movements controlled in an open-loop fashion. Given a Cartesian position $\mathbf{x}$ of an object candidate, the joint angles $\boldsymbol{\theta}$ which point the optical axis of the foveal cameras toward the position $\mathbf{x}$ have to be determined. The joint angles $\boldsymbol{\theta}$ are the solution to the inverse kinematics problem $\boldsymbol{\theta} = IK(\mathbf{x})$. A prerequisite for solving the inverse kinematics problem is the knowledge of the kinematic model of the system.

The kinematic structure of the head is available as CAD model from the construction process. For the neck, head tilt, and eye tilt joints these models match the head except for inaccuracies during manufacturing. In contrast, the position of the cameras relative to the pan joints cannot be derived from this model. The camera reference system depends on properties of the sensor and varies over different sensors of the same type. In order to determine the kinematic model of the complete head, a calibration procedure is required which allows to estimate the unknown or inaccurate parameters.

In the following, the approach for the kinematic calibration of the head-eye system as proposed in [Welke et al., 2008b] is introduced and evaluated. Based on the calibrated model, Section 6.2 discusses the execution of saccades by solving the inverse kinematics problem for the eye system of the Karlsruhe Humanoid Head.

6.1 Kinematic Calibration

The kinematic calibration of an active camera system falls into the wider class of the head-eye calibration problem. The general head-eye calibration problem is outlined in Fig. 6.2. The goal consists in determining the transformation X from the activated joint to the camera. The problem can be solved if the transformation B resulting from the actuation of the joint and the camera movement A corresponding to the actuation are known. The unknown X is calculated using the analogy $AX = XB$ [Shiu and Ahmad, 1989].

The literature offers a variety of approaches for head-eye calibration, where the main differences are representations of the transformation X and the method used to solve the $AX = XB$ problem. The solution can be derived in closed form [Young et al., 1992] using linear least-squares [Neubert and Ferrier, 2002] or using non-linear least squares [Li et al., 1994] minimization techniques. In comparative studies of the different methods, it has been shown that the non-linear least squares techniques provide the most accurate results in the presence of noise [Dornaika and Horaud, 1998, Li, 1998]. Consequently, the proposed method makes use of non-linear least squares techniques.

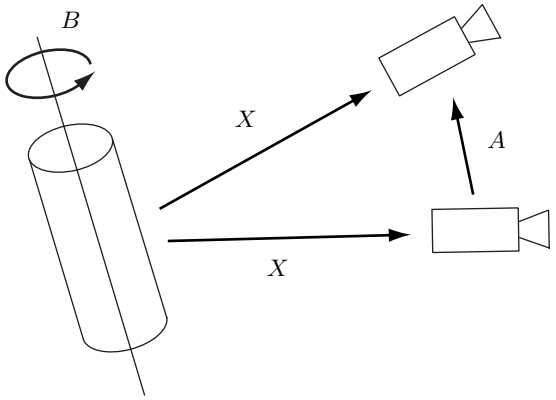

Fig. 6.2. The problem of head-eye calibration consists of estimating the unknown transformation X from the camera coordinate system to the local joint coordinate system. Given the actuation of the joint B and the corresponding movement of the camera A the problem can be solved using $AX = XB$.

Different representations used to describe the unknown transformation X have been proposed in the literature. These representations include DH parameters [Young et al., 1992], axis-angle representations [Tsai and Lenz, 1989, Shiu and Ahmad, 1989], quaternion representations [Dornaika and Horaud, 1998], Lie theory [Neubert and Ferrier, 2002], and logarithmic maps in $SE(3)$ [Ude and Oztop, 2009]. While the representation itself does not play a significant role for the accuracy of the calibration, there are fundamental differences in how the rotational part R and the translational part t of X are treated. In general, there are two classes of approaches for handling R and t. The first class estimates rotation R and translation t in separate steps while in the second class of approaches, R and t are estimated simultanously. The decomposition of the estimation in separate steps eases the optimization. Usually first the rotational part R is calculated using optimization techniques. In a second step, the translational part t is directly calculated without optimization. Since the separate estimation minimizes only the rotational error, the remaining translation error can still be significant due to noise or unmodelled factors. In contrast, the simultanous estimation allows to control the accuracies of translational and rotational part and to adapt them to the needs of the application.

According to [Tsai and Lenz, 1989], [Shiu and Ahmad, 1989], and [Park and Martin, 1994] two independent axes of rotation are necessary to determine all parameters of the transformation X. When only one rotation axis is considered, the problem is under-determined. The common assumption of two intersecting rotation axis which are calibrated at the same time restricts the approach to specific types of kinematic chains.

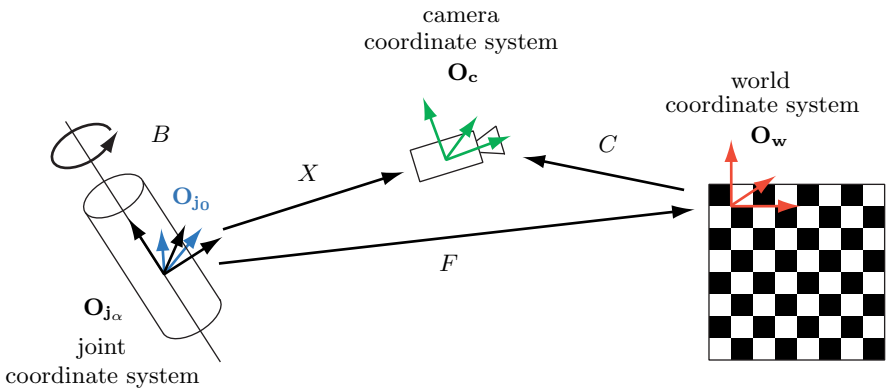

Fig. 6.3. Coordinate systems and transformations involved in the kinematic calibration procedure.

For the kinematic calibration of the Karlsruhe Humanoid Head a method is introduced in the following which allows to estimate the kinematic model of the head by calibrating one joint at a time. The approach is based on a representation with four parameters which models the translational and rotational parts of X and provides a solution also for the under-determined case. By estimating the rotational and the translational part at the same time, the accuracy of either part can be controlled. The solution to the $AX = XB$ analogy is estimated using non-linear optimization techniques in order to allow stability in the presence of noisy measurements. Before the approach is discussed in detail in the subsequent sections, the conventions used for coordinate systems and transformations are defined. Further, the acquisition of the data required for the optimization process in terms of the camera movement A is explained.

Coordinate Systems and Transformations

During calibration the camera which is rigidly attached to a moving joint observes a chessboard calibration rig. In Fig. 6.3 the coordinate systems and transformations involved in the calibration procedure of a single joint are illustrated. The following coordinate systems are defined:

- $\mathbf{O_w}$ denotes the fixed world coordinate system.
- $\mathbf{O_c}$ denotes the camera coordinate system defined by the optical center and the optical axis of the camera.
- $\mathbf{O_{j_0}}$ is a fixed coordinate system on the rotation axis of joint j at zero position.
- $\mathbf{O_{j_\alpha}}$ denotes the rotated coordinate system of the joint j after actuation with α.

Based on the coordinate systems, the following transformations are defined:

- $B(\alpha)$ describes the coordinate transformation from the fixed joint coordinate frame $\mathbf{O_{j_0}}$ to the rotated joint coordinate frame $\mathbf{O_{j_\alpha}}$. As only revolute joints are considered, $B(\alpha)$ describes a rotation around the rotation axis of the respective joint by an angle α, which is obtained from the encoder readings.
- F denotes the transformation from the fixed joint coordinate frame $\mathbf{O_{j_0}}$ to the world coordinate system $\mathbf{O_w}$. As both systems are fixed F is constant.
- C is the transformation from the world coordinate frame $\mathbf{O_w}$ to the camera coordinate frame $\mathbf{O_c}$. It depends on the camera position and therefore on the joint angle α.
- X denotes the transformation from the rotated joint coordinate system $\mathbf{O_{j_\alpha}}$ to the camera coordinate system $\mathbf{O_c}$. As the joint movement is already described by the transformation B, X remains constant independent of the actual joint actuation.

The goal of the calibration process is to determine the transformation X.

Estimation of the Camera Movement

In order to solve the head-eye calibration problem it is necessary to determine the movement of the camera. Therefore, the kinematic calibration is supported by the observation of a fixed calibration pattern in the peripheral cameras which defines the world coordinate system $\mathbf{O_w}$. The 6D pose of the camera C relative to the calibration pattern is calculated using the model-based method provided by OpenCV [Bradski, 2000]. As prerequisite, the intrinsic parameters of all cameras are calibrated.

The calibration pattern is observed at different joint angles α_k with $k \in [1; N]$. For each observation, the camera pose C_k is recorded and stored in the set of poses $\mathbb{C} = \{(C_1, \alpha_1), \ldots, (C_N, \alpha_N)\}$ along with the corresponding joint angles. This set of poses is used as input for the calibration process.

The following sections provide a detailed formal description of the proposed kinematic calibration method based on the defined coordinate systems and transformations. The model which is subject to optimization is derived in Section 6.1.1. Subsequently, the objective function for optimization is derived based on the model in Section 6.1.2 and representational aspects are discussed. While the formalization focuses on the calibration of one joint in the kinematic chain, the extension to multiple joints is detailed in Section 6.1.2. Further, the stereo calibration for moving eyes is derived from the calibrated model in Section 6.1.3. Finally, the proposed approach is evaluated in Section 6.1.4.

6.1.1 Derivation of the Model

In the following, the model for the optimization problem is derived. For a single joint of the head, given the set of camera poses $\mathbb{C}$, the sequence of transformations is derived which corresponds to the $AX = XB$ formulation of the head-eye calibration problem. In the formalization a pose defined in one of the coordinate systems $\mathbf{O_{j_0}}, \mathbf{O_{j_\alpha}}, \mathbf{O_w}$ and $\mathbf{O_c}$ is denoted with T using the same subscript as the coordinate system.

The transformation F between world coordinate system and joint coordinate system at zero position is constant for all values of the joint angle. By exploiting this fact a formalization similar to the head-eye calibration problem can be derived. First, F is expressed by combining the joint rotation $B(\alpha)$, the unknown transformation X, and the camera pose C. The coordinate transformation from the fixed joint frame to the rotated joint frame is

$$T_{j_\alpha} = B(\alpha)T_{j_0}. \tag{6.1}$$

A transformation from the rotated joint frame to the camera coordinate frame can be written as

$$T_c = XT_{j_\alpha}. \tag{6.2}$$

The transformation from the world coordinate system to the camera coordinate system depends on the position of the camera. This can be formulated as

$$T_c = CT_w. \tag{6.3}$$

The transformation from the fixed joint coordinate system to the world reference system is given by

$$T_w = FT_{j_\alpha}. \tag{6.4}$$

By combining Eqs. (6.1), (6.2), (6.3) and (6.4), F can be expressed as

$$F = C^{-1}XB(\alpha). \tag{6.5}$$

For two different values of the joint angle α_s and α_t, the corresponding transformation F_s and F_t with

$$F_s = C_s^{-1}XB(\alpha_s) \tag{6.6}$$

and

$$F_t = C_t^{-1}XB(\alpha_t) \tag{6.7}$$

can be calculated. Since the transformation F is constant for all joint angles the condition

$$F_s = F_t \tag{6.8}$$

always holds. The above equation corresponds to the $AX = XB$ formulation of the head-eye calibration problem. The $AX = XB$ form can be derived by setting $B(\alpha_s)$ in Eq. (6.6) and C_t in Eq. (6.7) to identity.

6.1.2 Solving the Calibration Problem

In theory, the model introduced in Eq. (6.8) could be solved without involvement of optimization. In practice however, the recorded calibration pattern pose and the joint encoder readings are not entirely accurate. Due to these errors it is impossible to find a matrix X which satisfies Eq. (6.8) for all possible combinations F_s and F_t with $s, t \in [1; N]$. Hence, we seek to estimate a transformation $\hat{X}$ which best fits for all combinations. Therefore, non-linear optimization is performed using the model defined by Eq. (6.8) using the estimated camera movements $\mathbb{C}$ as data.

In order to establish an objective function which is subject to minimization during the optimization process, a metric $d(F_s, F_t)$ has to be defined which expresses the similarity between two matrices F_s and F_t. Since both matrices lie within $SE(3)$, F can be decomposed in a rotation R and a translation $\mathbf{v}$. For our approach, two different metrics d_R and d_T are defined for rotational and translational part and are later combined to build the objective function. As metric for the translational part, the Euclidean distance is employed. The translation error is calculated using

$$e_T(s,t) = d_T(F_s, F_t) = ||\mathbf{v}_s - \mathbf{v}_t||,$$

where $\mathbf{v}_s$ and $\mathbf{v}_t$ are the translational components of F_s and F_t, respectively. The rotational part of F_s and F_t is represented using the unit quaternions $\mathbf{q_s}$ and $\mathbf{q_t}$. The rotational error is calculated using

$$e_R(s,t) = d_R(F_s, F_t) = f(\mathbf{q_s}^{-1}\mathbf{q_t}),$$

where f calculates the rotation angle from the resulting unit quaternion. Both errors, the translational error $e_T(s,t)$ and the rotational error $e_R(s,t)$ are combined using

$$e(s,t) = we_R(s,t) + (1-w)e_T(s,t), \tag{6.9}$$

where w is a weight that allows to balance the accuracy of translational versus rotational errors. The objective function evaluates the error for all pairs of recorded data s, t

$$\epsilon = \frac{1}{N(N-1)} \sum_{s \in \{1,\ldots,N\}} \sum_{t \in \{1,\ldots,N\} \setminus s} e(s,t). \tag{6.10}$$

Given the objective function, the optimization process estimates the solution which best fits the underlying model in terms of the model parameters. The problem of estimating a matrix in $SE(3)$ has six degrees of freedom, three rotations α, β and γ and three translations x, y and z. Thus, the complete set of parameters is given by $\mathbf{p} = (\alpha, \beta, \gamma, x, y, z)$.

As discussed in the introduction, when only one joint is calibrated, the problem is under-determined and not all six parameters can be estimated. In order to derive a kinematic model suitable for the solution of the inverse kinematics problem, the estimation of four parameters is sufficient. The two parameters that are not estimated, are the translation along the rotation axis and the initial rotation around the rotation axis. These parameters can be chosen freely and are set to zero in our approach.

Using the reduced set of four parameters $\mathbf{p'}$ for the optimization process, the optimal solution for $\hat{X}$ is estimated by minimization of the objective function

$$\min_{\mathbf{p'}} \left(\frac{1}{N(N-1)} \sum_{s \in \{1,\ldots,N\}} \sum_{t \in \{1,\ldots,N\} \setminus s} e(s,t) \right). \tag{6.11}$$

The estimate of the unknown transformation $\hat{X}$ is then recovered from the parameters $\hat{\mathbf{p}}'$ corresponding to the optimal solution. The optimization problem formalized in Eq. (6.11) is solved using the Levenberg-Marquardt method for non-linear optimization [Levenberg, 1944, Marquardt, 1963].

Extension to Multiple Joints

While the formal description in the previous paragraph only covered the calibration of a single joint, the approach is now extended to a kinematic chain comprising multiple joints. Starting the calibration process with the joint where the camera is attached, the preceeding joints can be calibrated one by one in a recursive manner.

Let X_j with $j \in [i+1, \cdots, n]$ be the result of the calibration of the joints which succeed joint i up to the last joint n in the kinematic chain. Let further be the actuation of all joints with index $j > i$ be zero. Reconsidering Eq. (6.5) the constant transformation F for a joint angle α can be rewritten for joint i with

$$F = C_i^{-1} \prod_{m=0}^{n-i-1} (X_{n-m}) X_i B_i(\alpha).$$

By replacing the chain of transformations with the matrix A_i, the above equation becomes

$$F = C_i^{-1} A_i B_i(\alpha).$$

and can be solved analog to the method proposed above. The seeked transformation X_i can be recovered using

$$X_i = \left[\prod_{m=0}^{n-i-1} (X_{n-m}) \right]^{-1} A_i. \tag{6.12}$$

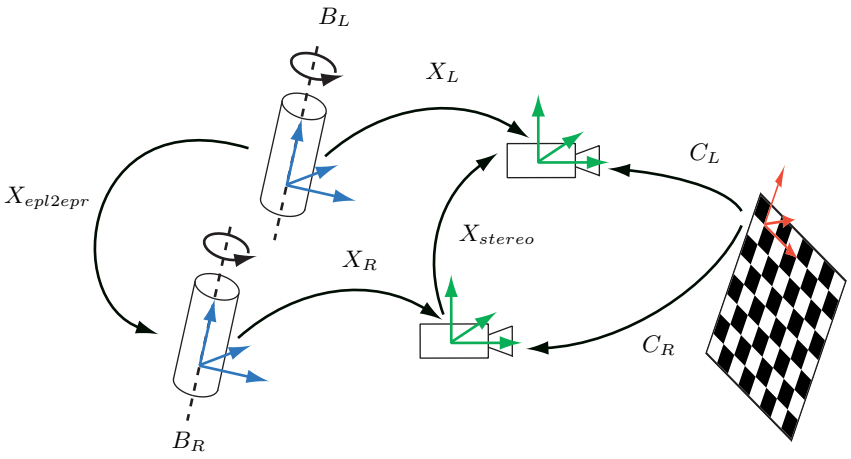

Fig. 6.4. Calculation of the stereo calibration X_{stereo} from the calibrated model for moving eyes.

6.1.3 Stereo Calibration

As discussed in Section 5.3 the calculation of the 3D location from 2D correspondences requires the knowledge of the extrinsic camera parameters. The extrinsic parameters can be derived from the transformation between the coordinate systems of left and right camera. For active camera systems, as the case on the Karlsruhe Humanoid Head, this calibration is not static but changes with the movement of the eyes. In the following, the stereo calibration will be derived from the kinematic model which has been calibrated with the method proposed in the previous sections.

The goal consists in determining the transformation X_{stereo} between left and right camera coordinate system as depicted in Fig. 6.4. Once both eye pan axes are calibrated with the kinematic calibration procedure, X_{stereo} can be determined from the calibrated model for arbitrary camera actuations. This enables methods of stereo vision with active eyes.

In order to calculate X_{stereo}, the transformation $X_{epl2epr}$ has to be made available in addition to the calibrated model of the kinematic chains corresponding to left and right camera. For this purpose, a calibration rig is positioned in front of both cameras of the stereo pair and the transformation for the fixed system is determined.

For the transformation $X_{epl2epr}$, the following holds

$$X_{epl2epr} = B_{epr}^{-1}(\alpha_R) \cdot X_R^{-1} \cdot C_R(\alpha_R) \cdot C_L^{-1}(\alpha_L) \cdot X_L \cdot B_{epl}(\alpha_L).$$

The matrices X_L and X_R represent the kinematic calibrations of both eye pan joints. $B_R(\alpha_L)$ and $B_L(\alpha_R)$ represent the rotations of the respective joints by the angles α_L and α_R. The transformations C_L and C_R depend on the same angles. In theory the transformation $X_{epl2epr}$ could be determined at any position of the two joints. In practice however, modeling the joint movements using $B_R(\alpha_R)$ and $B_L(\alpha_L)$ introduces errors. Therefore, the most accurate result is obtained with both joints at zero positions, i.e. $\alpha_L = \alpha_R = 0$ where $B_R(\alpha_R)$ and $B_L(\alpha_L)$ become identity. In order to calculate $X_{epl2epr}$, the calibration pattern is positioned in the visual field of both stereo cameras of the considered pair and C_L and C_R are determined using the 6D pose of the calibration pattern.

The stereo calibration X_{stereo} for arbitrary joint angles can then be calculated using

$$X_{stereo} = X_L \cdot B_L(\alpha_L) \cdot X_{epl2epr}^{-1} \cdot B_R(\alpha_R)^{-1} \cdot X_R^{-1},$$

where α_L and α_R correspond to the current encoder readings of the eye pan joints.

6.1.4 Evaluation

In the following, the accuracy of the calibrated kinematic model is evaluated in terms of the underlying coordinate transformations and the resulting accuracy of stereo triangulation. For the experiments, the head was equipped with 4 mm lenses. The calibration process was performed using a calibration pattern with 9×7 squares of side lenght 3.63 cm positioned at a distance of 75 cm in front of the head. This distance allowed movements of the eye pan joints in the range $\pm 10°$ and movements of the eye tilt joint in the range $\pm 15°$. For each joint, 21 positions were recorded and stored in the set of calibration data $\mathbb{C}$.

In order to determine the calibrated model, the correct weight w for the trade-off between rotational and translational error according to Eq. (6.9) needs to be chosen. Therefore, different weights were evaluated. The results indicate, that the weight is not crucial for the accuracy of the model. The optimization converged with similar residue errors for weights within the range $[0.1; 0.98]$. This indicates that as long as rotational and translational part are considered in the objective function, e.g. $w \neq \{0, 1\}$, the optimum is found. The number of iterations required for convergence was minimal when choosing a weight $w = 0.9$ which was used in the subsequent experiments.

Accuracy of the calibrated model

In a first series of experiments the accuracy of the kinematic model of the eye joints as determined with the proposed method was evaluated. Therefore, a smaller calibration pattern with 5×4 squares of side length 4.5 cm was

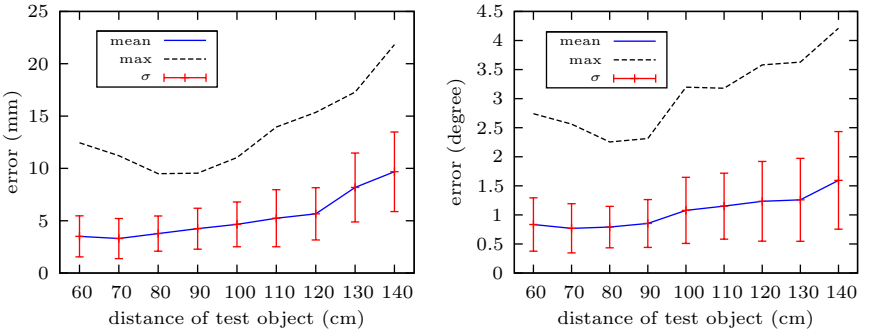

Fig. 6.5. Accuracy of the calibrated kinematic model. For different distances of the test pattern, the eyes were moved to random positions. The pose of the test pattern was determined using a model-based approach. After transformation to a common coordinate frame, translational and rotational errors were calculated with respect to a reference pose determined with zero actuation of all joints. Left: Translational error. Right: Rotational error.

used. The test pattern was positioned at distances ranging from 60 cm up to 140 cm from the eyes. For each distance, a reference pose of the calibration pattern was determined using a model-based approach with all joints in zero position. Subsequently, 100 arbitrary eye movements were performed in the extended range of $\pm 15°$ for each eye joint. For each movement, the pose of the calibration pattern was again determined using the model-based approach. The reference pose and the pose of each random actuation were transformed to a common coordinate frame using the calibrated model. In order to judge the accuracy of the model, the translational and rotational errors between the poses were determined.

The experiment was performed on 100 calibrated models. The calibrated models were generated based on 10 different intrinsic calibrations, for each intrinsic calibration 10 kinematic calibrations were performed. The results depicted in Fig. 6.5 illustrate the mean error, the standard deviation of the error and the maximum error for all distances considered in the experiment over all 100 calibrations. The minimum error is reached around the distance of 75 cm which was used during calibration. The minimum mean translation error amounts to about 2.5 mm and the minimum mean rotation error amounts to 0.8°. Both mean errors increase toward the maximum tested distance of 140 cm resulting in the maximum measured mean errors of 9.6 mm and 1.6°. The maximum error for translation and rotation is also reached at the distance of 140 cm and amounts to 22 mm and $4,3°$, respectively. Note that these errors not only include inaccurately calibrated kinematic models but also errors resulting from the model-based pose estimation step.

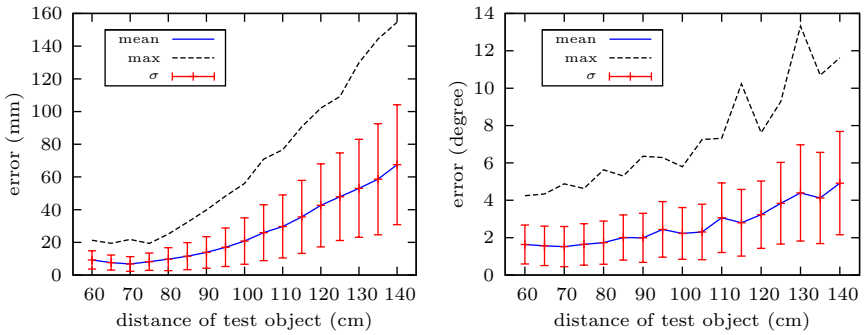

Fig. 6.6. Accuracy of the stereo calibration. The 6D pose of a test pattern was determined in the left perspective camera using a model-based approach. Furthermore, the pose of the pattern was determined using stereo triangulation. The plots show the translational and rotational error between model-based and triangulation-based pose estimation for different distances of the test pattern. Left: Translational error. Right: Rotational error.

Stereo Calibration Accuracy

The accuracy of the stereo calibration was tested in a similar way. The position of the test pattern was determined in the left camera using a model-based approach. Additionally, three corresponding corner points of the calibration pattern in the left and right image were determined and the 3D position of these points was calculated utilizing epipolar geometry based on the calibrated transformation matrix X_{stereo}. From these three points, the pose of the test pattern in the left camera was estimated. Thus, the accuracy of the stereo triangulation could be evaluated with respect to the model-based approach.

The experiment was performed using a test pattern located at distances from 60 cm up to 140 cm from the eye system. In each step 50 random eye configurations were recorded and evaluated. Fig. 6.6 illustrates the resulting translational and rotational error between the pose determined using the model-based approach and the pose calculated based on stereo triangulation. As can be seen the errors in the stereo triangulation accuracy show the same trend as the kinematic error, but are much larger. The increase of the error results from the fact, that small position errors within the kinematic calibration result in larger errors when performing stereo triangulation. The best results could again be achieved for small distances of the test pattern. At a test pattern distance of 70 cm the mean translational error was measured with 8.7 mm and the minimum mean rotational error was measured with 1.72° for the same distance.

Repeatability

In order to determine the repeatability of calibrations, the 100 kinematic calibrations were analyzed and compared. This analysis was performed for both eye pan joints. More precisely, the coordinate transformation from the camera coordinate system to the first rotation joint was examined. This transformation should ideally stay constant over the calibrations. In order to determine the error, the orientation of the rotation axis and the shortest translation from optical center to the rotation axis were calculated. Within all 100 kinematic calibrations, the standard deviation of the orientation amounts to about 0.17°. The standard deviation of the translation was measured with about 4 mm. This indicates that the orientation of the rotation axis can be recovered very accurately from the collected data, while the translation is hard to recover precisely given the 20° range of motion used for the collection of data. The problem of inaccuracies in the translation causes errors that make stable perception over several saccades difficult. How these inaccuracies are handled is discussed in the context of the correspondence problem (see Section 7.2.2).

6.2 Saccadic Eye Movement Execution

The execution of saccadic eye movements requires to direct the gaze of the active stereo camera pair toward a position in the visual field. As previously discussed, saccadic eye movement is performed in an open-loop manner. Therefore, an appropriate configuration of the eye system in terms of the joint angles $\boldsymbol{\theta}$ has to be determined which aligns both optical axis with the target position $\mathbf{x}$. In the following, the problem of directing the gaze is formulated as an inverse kinematics problem and the solution is derived on basis of the calibrated kinematic model discussed in the previous section. Further, the suitability of the calibrated model is evaluated in terms of the accuracy of executed saccadic eye movements.

6.2.1 Solving the Inverse Kinematics Problem

As already outlined in the introduction to this chapter, the problem of calculating the joint angles $\boldsymbol{\theta}$ of a kinematic chain given a target position $\mathbf{x}$ for the end-effector is referred to as inverse kinematics problem $\boldsymbol{\theta} = IK(\mathbf{x})$. In contrast to the classical inverse kinematics problem where the goal is to move a tool to a desired Cartesian position, the execution of saccadic eye movements requires to determine joint angles that direct the gaze of the camera system in terms of their optical axis toward a given target position.

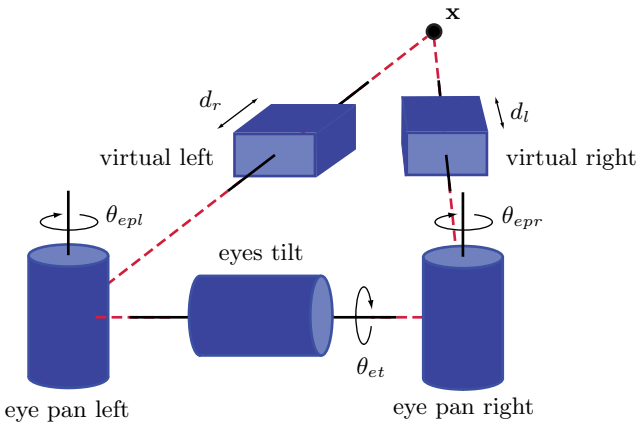

Fig. 6.7. In order to solve the problem of directing the gaze toward a given Cartesian position $\mathbf{x}$ using inverse kinematics, the kinematic chain for each eye is extended by a virtual prismatic joint.

In order to execute saccadic eye movements using inverse kinematics, the kinematic chain of the eye system is extended as depicted in Fig. 6.7. For each eye, a prismatic joint is attached to the camera coordinate system. The prismatic joint serves as virtual joint and defines a movement along the optical axis of each camera. Using the end-effector defined by the extended kinematic chain, the inverse kinematics problem can be solved using standard methods.

The inverse kinematics problem is solved for the left and the right eye separately. For each eye s, a configuration $\boldsymbol{\theta}_{\mathbf{ext,s}} = (\theta_{et,s}, \theta_{ep,s}, d)$ is calculated which points the optical axis toward the Cartesian position $\mathbf{x}$. The solution to the problem is determined using differential kinematics. The Jacobian is calculated based on the calibrated kinematic model using the geometric method [Orin and Schrader, 1984]. Since the resulting matrix is regular and in $\mathbb{R}^{3\times3}$ the inverse J^{-1} always exists. Thus, the relation between velocities in Cartesian space and joint space is defined with

$$\dot{\boldsymbol{\theta}}_{\mathbf{ext,s}} = J^{-1}(\boldsymbol{\theta}_{\mathbf{ext,s}})\dot{\mathbf{x}}.$$

By approximating the above model through linearization, the inverse kinematics problem is solved iteratively. The Cartesian error $\boldsymbol{\Delta}\mathbf{x}$ is derived using the forward kinematics of the system as defined by the calibrated model.

Once a solution has been calculated for both eyes, an approximated configuration $\boldsymbol{\theta} = (\theta_{et}, \theta_{epl}, \theta_{epr})$ is derived, where θ_{epl} and θ_{epr} are the joint angles for left eye pan and right eye pan and θ_{et} is the joint angle for the common tilt. The eye pan joints angles are directly taken from the solutions for each eye $\boldsymbol{\theta}_{\mathbf{ext,l}}$ and $\boldsymbol{\theta}_{\mathbf{ext,r}}$ using

$$\theta_{epl} = \theta_{ep,l} \; ; \; \theta_{epr} = \theta_{ep,r}.$$

The eye tilt joint angle is determined using the mean of both kinematic chains

$$\theta_{et} = \frac{\theta_{et,l} + \theta_{et,r}}{2}.$$

The calculated configuration $\boldsymbol{\theta}$ approximates the solution to the inverse kinematics problem.

6.2.2 Evaluation of Saccade Accuracy

In order to evaluate the accuracy of the saccadic eye movements, the first three DoF of the head-eye system (namely eye pan left, eye pan right and eyes tilt) were calibrated. In the experiments a test pattern was positioned at distances ranging from 60 cm to 140 cm. For each distance, arbitrary camera poses were generated by moving all three joints to random positions in the interval of $-10°$ to $10°$. The pose of the test pattern in the left camera was determined using a model-based approach. Based on the calibrated kinematic model, the inverse kinematics problem was solved and a saccade was performed in order to point the optical axes of the cameras toward a reference point on the test pattern. In order to evaluate the accuracy, the distance between the reference point of the test rig and the principal point of the camera in the image plane was measured after performing the movement.

Since left and right camera share a common tilt joint, the error in the vertical direction will never be zero. The inverse kinematics outputs the mean eye tilt angle for left and right eye to minimize the overall error. In order to compensate for this effect, the modified error

$$y_m = \frac{y_l + y_r}{2}$$

was used to derive a more expressive result for the quantiative evaluation. Based on the modified vertical error y_m and the horizontal error, the position errors for the left and the right camera were calculated.

In Fig. 6.8 the results of the experiments are illustrated. The error in the left camera decreases with increasing distance of the test pattern. This effect is caused by the fixed range for eye pan ($-20°$ to $20°$) and eye tilt ($-15°$ to $15°$) movement. The maximum of the mean error for the left camera amounts to about 2 pixel for a distance of 60 cm. The plot for the right camera differs slightly from the results of the left camera. The increased error in the right camera is caused by the additional transformation X_{stereo} which has to be considered in the differential kinematics. In subsequent experiments, where the test pattern was located in the right camera, similar results could be produced with more accurate results for the right camera and less accurate results for the left camera.

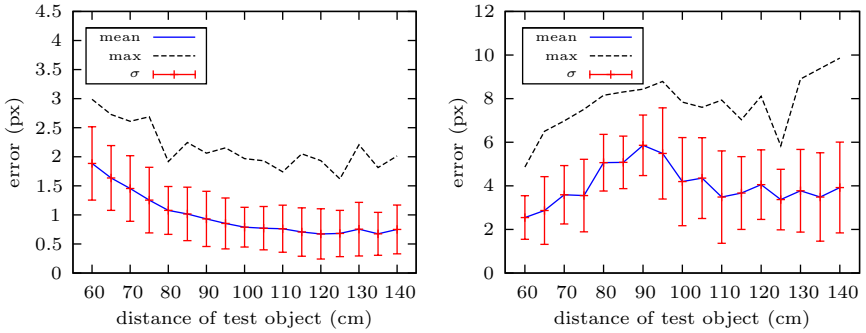

Fig. 6.8. Accuracy of saccadic eye movements. The calibrated model was used for calculation of the inverse kinematics in order to point the optical axes of the cameras toward a reference point on the test pattern at different distances. The results illustrate the distance in pixels from principal point to the reference point on the test pattern in the image planes. Left: Left camera. Right: Right camera.

6.3 Summary

The knowledge of the kinematic model of the head is a prerequisite for the execution of saccades and the calculation of stereo correspondences based on the epipolar geometry. In order to determine the pose of the camera coordinate systems which are not available from CAD models, a novel approach for the kinematic calibration of the head-eye system was introduced in this chapter. The proposed approach allows to determine a kinematic model by moving the head joints and observing a fixed calibration rig in the camera images. The underlying non-linear optimization procedure considers one joint at a time and estimates the rotational and the translational part of the unknown transformation at the same time.

Based on the kinematic model extended by a virtual prismatic joint, the execution of saccades is implemented using differential inverse kinematics. The accuracy of saccade execution showed to be accurate up to a mean position error of 6 pixels. Further, the calibrated model was evaluated in the context of stereo calibration. Based on stereo triangulation, the 6D pose of a calibration rig was reconstructed and compared to the pose estimation result using a model-based approach. The experiments show that a significant error remains which increases with the distance of the rig to the camera pair. In order to solve the correspondence problem between objects observed under different configurations of the eyes, an approach is proposed in the subsequent chapter which takes into account these calibration inaccuracies.

7

Transsaccadic Memory

Saccades performed by an active vision system provide different partial views of the scene. In order to establish a consistent model of the perceived world, a spatial scene memory is necessary which accumulates the gathered visual information "across separate glances and over time" [Melcher, 2001]. For this type of memory the term transsaccadic memory as introduced in [Irwin, 1992] is used in the following.

In this chapter, such a transsaccadic memory in the context of active visual search is introduced, which allows for the accumulation of information about instances of the target object across saccadic eye movements. In general, the transsaccadic memory encodes spatial information accompanied with the visual appearance of scene fragments. Through constant verification of the stored entities based on recognition in the foveal cameras the spatial information as well as the match of a scene fragment with the representation of the target object are updated. Together with the approach for visual attention as discussed in the next chapter, the transsaccadic memory provides a consistent visual model of the scene with respect to the target object searched for. As such, the transsaccadic memory represents the result of the memory-based active visual search approach.

7.1 Memory Organization

The implementation of a spatial representation of the scene requires to define a suitable reference frame, which is used to represent the locations of entities in space. Two different organizations of such spatial representations have been proposed in the literature: the egocentric and the allocentric organization. While both terms have not been used in a consistent fashion throughout the literature, Klatzky provides a precise definition of both types of organization [Klatzky, 1998]. Both egocentric and allocentric representations "convey the layout of points in space by means of an internal equivalent of a coordinate

system (which may be distorted or incomplete)". While entities in allocentric representations are stored with respect to a reference frame external to the perceiver, entities in egocentric representations use an origin and an orientation which is defined by the body of the perceiver. Egocentric representations are often referred to as self-to-object relations, allocentric representations as object-to-object relations [Meilinger and Vosgerau, 2010]. In human perception and spatial reasoning, both egocentric and allocentric representations are involved. In the case of active visual search, the goal is to update the content of the transsaccadic memory using the visual information captured at the current gaze direction. According to [Meilinger and Vosgerau, 2010], the task of updating involves egocentric representations. Further, the visual perception itself is performed in an egocentric manner. As a consequence, the egocentric organization is used for the representation of spatial locations in transsaccadic memory for this work.

Evidence has been found for a variety of different possible egocentric frames which are specific for different effectors in the cognitive neuroscience literature (see [Briscoe, 2008] for a review). The human brain seems to hold multiple representations which best fit the needs of a specific task. In contrast to many passive attention systems which make use of retinotopic reference frames, for the task of active visual search a head-centered frame is chosen as reference. The origin and orientation of the egocentric reference frame is defined by the base joint of the head kinematic chain.

The transsaccadic memory receives input from two different sources: the foveal percept and the peripheral percept (see Fig. 7.1). Thereby, peripheral object candidate detection only performs a rough analysis of the scene whereas the foveal object recognition allows to determine the existence of a target object at a specific location in the scene. In order to account for the different types of information provided by the two sources the transsaccadic memory is divided into two layers as depicted in Fig. 7.1. The preattentive memory is updated continuously based on the result of the object candidate detection mechanism thus keeping track of relevant candidates in the scene. In contrast, foveal object recognition requires attention in order to bring relevant locations of the scene to the view of the foveal cameras. The content of the object memory is updated based on the results of foveal vision after each saccade. For objects which are not represented in the object memory because they have not been inspected by the foveal cameras yet, the preattentive memory serves as a prior.

In the following two section, the entities and associated mechanisms of transsaccadic memory are discussed. In Section 7.2 the preattentive memory is introduced and an approach for solving the correspondence problem between the memory content and observed candidates is derived. Subsequently, the representation of the object memory and the update using foveal object recognition is discussed in Section 7.3. The links between both memory layers are made explicit in Section 7.4.

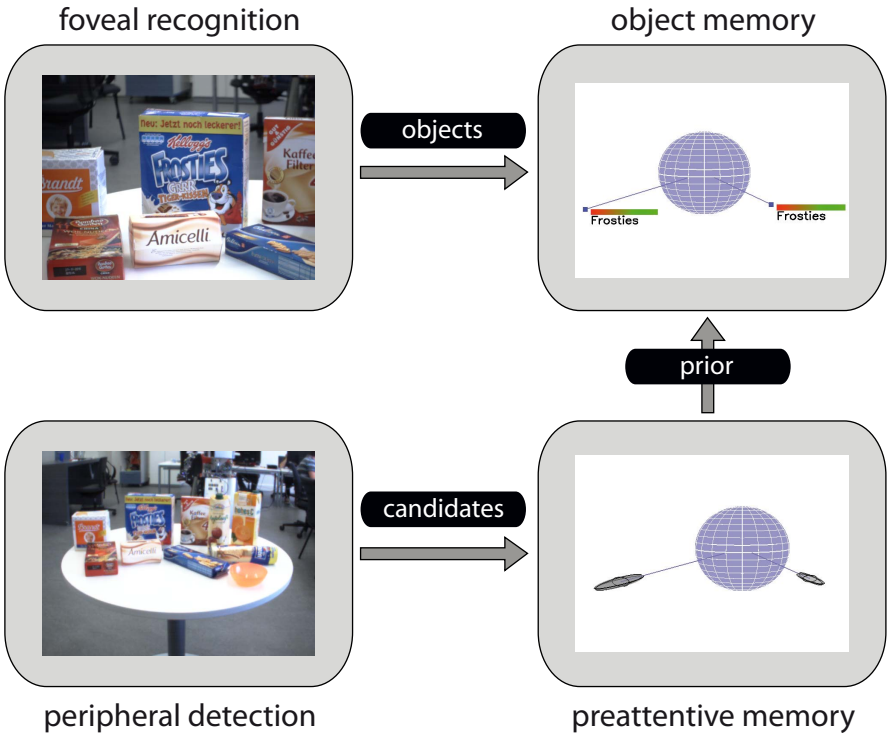

Fig. 7.1. The transsaccadic memory is subdivided into two layers: the preattentive memory layer and the object memory layer. While the preattentive memory layer accumulates object candidates resulting from peripheral object detection, the object memory content is accumulated by focussing regions of interest with the foveal cameras.

7.2 Preattentive Memory Layer

The main task of the preattentive memory layer consists in providing information about areas of the scene which are relevant to the current search task. These object candidates are determined using peripheral detection. An example of detected object candidates and the preattentive memory content is depicted in Fig. 7.2.

The object candidates constitute the starting point for the generation of sequences of saccades in order to perform object recognition. The execution of saccades requires to associate a valid target position to each object candidate. This position can then be used to configure the active system in order to direct the gaze toward the object candidate for a more detailed inspection.

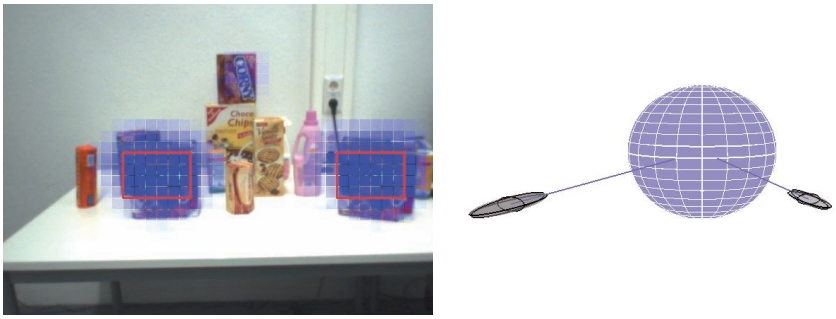

Fig. 7.2. Example of preattentive memory content. Left: Two object candidates are detected within the peripheral images in a table top scene. Right: Preattentive memory holds an entity for each of the candidate locations accompanied with its spatial uncertainty.

The object candidates extracted in the peripheral images constitute only a snapshot of the current scene, restricted by the field of view of the stereo camera pair and valid only for the current point in time. Due to inaccuracies in the calibrated model as well as inaccuracies in object detection and recognition, these positions are subject to noise. In order to establish a stable representation of the scene with respect to object candidates, methods are required that allow to solve the correspondence problem between observed and stored object candidates in order to accumulate spatial information in a consistent fashion.

In the following sections, solutions to the problems of identifying correspondences and inferring a consistent spatial memory are introduced. The problem is modeled in a probabilistic fashion and solved using Bayesian inference. The discussed model and inference constitutes an extension to previous work described in [Welke et al., 2009]. Beginning with the definition of an entity in preattentive memory in the following section, this chapter proceeds with a probabilistic model for the accumulation of spatial information in Section 7.2.2. Based on this model, the inference of spatial locations and the existence of object candidates is detailed in Section 7.2.3. The update of the memory entities using the inferred results is detailed in Section 7.2.4. Finally, the proposed approach is evaluated in the context of active visual search in Section 7.2.5.

7.2.1 Memory Entities

Entities in preattentive memory correspond to hypotheses about object instances being present in the scene. Each entity results from a match between features of the searched object and the current scene. In order to accumulate

information extracted at different gaze directions, the identity, the location and the existence of the object instance is stored with each entity. In the following, the stored information is described in detail.

- **Identity of object candidates**
  The identity of object candidates is defined by the feature $\mathbf{o}$ of the target object and the feature $\mathbf{s}$ corresponding to the match extracted in the current scene. The object feature $\mathbf{o}$ represents the search query, while the scene feature $\mathbf{s}$ encodes the visual appearance of the object candidate. Together with both features, the probability w of observing the scene feature given the object feature is stored.

- **Location of object candidates**
  Each entity in preattentive memory is accompanied with a spatial location in 3D Cartesian space. In order to encode spatial uncertainties resulting from the matching procedure, the location of an object candidate is represented using a normal distribution $\mathcal{N}(\mu, \Sigma)$ with mean $\mu \in \mathcal{R}^3$ and covariance $\Sigma \in \mathcal{R}^{3\times3}$.

- **Existence of object candidates**
  The real world is subject to change, thus object candidates might be removed or new object candidates might appear in the visual field. In order to cope with such dynamic aspects, the probability of the existence of an object candidate c is stored for each memory entity.

Together, an entity in preattentive memory M_p holds information about identity, location and existence as described above

$$M_p = (\mathbf{o}, \mathbf{s}, w, \mu, \Sigma, c)\,. \tag{7.1}$$

7.2.2 Model for Memory Update

In order to provide a stable representation of the scene with respect to object candidates, the information encoded in entities is accumulated during saccadic eye movements. After each gaze shift a new measurement of perceived object candidates is performed using the peripheral views. In order to determine, whether an observed object candidate is already represented by an entity in the preattentive memory, the correspondence between entities and perceived object candidates has to be determined. Once the correspondence problem has been solved, the information encoded in the stored entity can be updated using the measured location of the object candidate in order to derive a more concise representation of the scene.

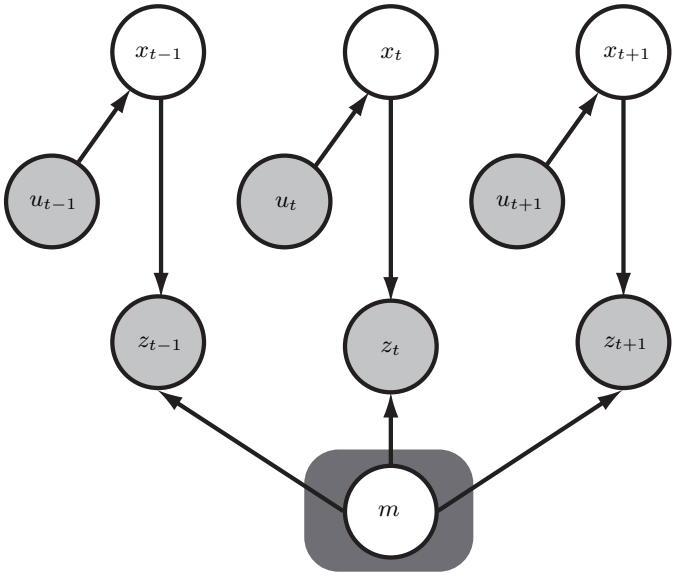

Fig. 7.3. Bayesian network for the accumulation of spatial information. The control u comprises the observed motor command and camera parameters. The system state x is modeled by accounting for the dominant noises in execution and calibration. Each observation of features in the system state x results in measurements z. The goal is to infer map m while capturing measurements during multiple saccadic eye movements.

In the following, a model is proposed which allows to solve the problem of accumulating spatial information over time within the probabilistic framework. In Fig. 7.3 a dynamic Bayesian network is illustrated which covers all involved parameters and their dependencies which are explained in the following. The goal of the approach consists in inferring the memory entities which correspond to object candidates within a spatial egocentric map m. Entities are accumulated over time, after each gaze shift a new measurement z is performed and fused into the map m. The spatial locations of measurements are brought into the common egocentric reference frame by considering the head pose according to the current system state x. The system state itself depends on the control u which includes the joint angles corresponding to the current gaze direction. While both, the control u and the measurement z are observed, the system state x and the map m are unobserved variables of the model.

The problem in consideration deals with sequences of random variables. The subscript of each random variable indicates its position in the time-series represented by the Dynamic Bayesian network. In the context of the proposed approach, the subscript t indicates a random variable associated with the t-th saccade performed during the active visual search task.

Definition of sample spaces

The derivation of an approach for the inference within the probabilistic model requires to define the sample spaces associated to each involved random variable as depicted in Fig. 7.3. Therefore, the dimensionality and semantic meaning of the underlying continuous sample spaces are defined in the following.

The control state u comprises all necessary quantities in order to determine the exact pose of left and right camera after the saccadic eye movement. The pose of the camera is defined by the current encoder readings for the eye joints and the kinematic model of the head. The encoder readings of the eye tilt and both eye pan joints are subsumed in the vector of joint angles $\boldsymbol{\theta} = (\theta_{et}, \theta_{epl}, \theta_{epr})$. The kinematic model of the eye joints is calibrated using the method introduced in Section 6.1. As has been shown in the experiments (see Section 6.1.4), the resulting calibration exhibits significant noise in the position of the rotation axes. In order to allow for robust mapping of object candidates, the position of the rotation axes is included in the control state in terms of their absolute position with respect to the preceding joint $\mathbf{l} = (\mathbf{l}_{et}, \mathbf{l}_{epl}, \mathbf{l}_{epr})$, where $\mathbf{l}_j \in \mathbb{R}^3$. Overall, the control state u is defined by

$$u = (\boldsymbol{\theta}, \mathbf{l})\,;\, u \in \mathbb{R}^{12}.$$

The control state reflects the measured quantities in terms of joint angles and axes translations, which vary from the true state due to inaccuracies in the execution of saccades and in the kinematic model. The unobserved true state of the system is represented with the random variable x. The system state resides in the same sample space as the control state u, thus $x \in \mathbb{R}^{12}$.

The measurement z represents all observed object candidates in the scene. Since multiple object candidates have to be handled, z is composed of a set of observations

$$z = (z_1, \cdots, z_K)\,,$$

where K is the number of observed candidates. Each object candidate is observed in the left and right camera at a spatial location $\mathbf{b}$ in the image plane. Furthermore, in both camera images, the object candidate is observed with the feature space representation $\mathbf{s}$. The stereo camera system is modeled as a sensor which provides 3D measurements. Thus, the measurement of an object candidate z_k is defined within 3D Cartesian space, where its position $\mathbf{a}$ results from corresponding 2D points $\mathbf{b_l}$ and $\mathbf{b_r}$. In summary, the observation z_k of an object candidate is composed of the spatial locations $\mathbf{a}$ and the two corresponding feature descriptors for left and right image

$$z_k = (\mathbf{a}, \mathbf{s_l}, \mathbf{s_r})\,.$$

Assuming a feature descriptor of size S, the sample space of the measurement then becomes $z \in \mathbb{R}^{K(3+2S)}$.

The map m encodes the locations of object candidates in an egocentric coordinate frame. The map is composed of N landmarks corresponding to object candidates in the scene. Each landmark m_n is represented with its spatial location $\mathbf{c} \in \mathbb{R}^3$ and the associated scene feature $\mathbf{s}$

$$m_n = (\mathbf{c}, \mathbf{s}) .$$

The ensemble of landmarks yields the map $m = (m_1, \cdots, m_N) \in \mathbb{R}^{N(3+S)}$.

Motion Model

The motion model comprises all relevant uncertainties which result from the motion of the eyes. The current state of the head is represented by the observed state u_t. The actual state of the eyes, modeled with the latent variable x_t, is subject to uncertainty. The motion model formulates the conditional probability $p(x_t|u_t)$ that the system is in state x_t given the observed state u_t under consideration of the dominant sources of noise.

Positioning noise

The positioning of the head is very accurate (see [Asfour et al., 2008]). However, depending on the settings of the low-level controllers, small errors in positioning remain. Furthermore, a small error in the conversion of encoder readings to joint angles has been observed. Both errors are assumed as additive noise with Gaussian uncertainty. While the positioning noise is assumed to be constant over the complete space of joint angles, the conversion noise is modeled as a normal distribution with variance proportional to the rotation of the eye from its zero position. Considering joint angle readings $\boldsymbol{\theta} = (\theta_{et}, \theta_{epl}, \theta_{epr})^T$ for the eye tilt and both eye pan joints, the uncertainty resulting from inaccurate positioning Σ_p and inaccurate conversion from encoder values to joint angles Σ_c lead to the positioning errors

$$e_{pos,p} \sim \mathcal{N}(0, \Sigma_p),$$
$$e_{pos,c} \sim \mathcal{N}(0, \Sigma_c \boldsymbol{\theta}),$$

where Σ_p and Σ_c are diagonal matrices modeling the variance for each joint independently.

Calibration noise

The eye system is calibrated with the method described in Section 6.1. For the generation of gaze directions, the resulting calibrated model showed to be sufficiently accurate in the experiments. However, the calculation of 6D poses from stereo correspondencesis is affected considerably even by small errors. Thus, the remaining inaccuracy is considered in the motion model.

As already discussed in the last paragraph, the dominant error in the calibration resides in the translational component of the coordinate transform. In order to cope with this inaccuracy, the calibration error e_{cal} that is induced to the position of rotation axes for different calibrations is modeled using additive Gaussian noise. The position of each rotation axis is exrpessed with its translation $\mathbf{l}$ relative to the preceding joint. The uncertainty about the position of the axes is modeled using the normal distribution

$$e_{cal} \quad \sim \quad \mathcal{N}(0, \Sigma_{cal}).$$

Given the above considerations, the observed state u_t comprises the currently measured actuation of the joints $\boldsymbol{\theta}_t$ and the calibrated joint axes translations $\mathbf{l}$. Using the formulated errors, the conditional probability of being in the hidden state x_t given an observed state $u_t = (\boldsymbol{\theta}_t, \mathbf{l})^T$ can be formulated by

$$p(x_t|u_t) = u_t + \begin{pmatrix} e_{pos,p} + e_{pos,c} \\ e_{cal} \end{pmatrix}.$$

Measurement Model

Both cameras of the active vision system are modeled as a single sensor that measures 3D positions of object candidates. Hence, intuitive inference of 3D maps is facilitated. The measurement model is defined by the conditional probability $p(z_t|x_t, m)$ that a measurement z_t is observed given the system state x_t and the map m. Let each measurement z_t be composed of K measured features $z_{k,t}$. We derive the above conditional probability for the measurement of a single feature $z_{k,t}$. First a model for the uncertainty of the 3D position resulting from inaccuracies in the 2D position of features in the image plane is derived. Then the resulting 3D localization uncertainty is expressed with respect to the eye state x_t.

The localization of features is affected by inherent measurement errors. While accurate 2D feature extraction methods allow subpixel accuracy, the resulting error during 3D reconstruction is, depending on the distance of the feature, of considerable amount and cannot be neglected in approaches for probabilistic mapping of features. In the following, a model of the 3D localization error is derived from measured 2D feature locations in the left and right camera images and the associated 2D localization uncertainties.

A common approach for modelling the 2D localization error e_{2d} of a feature in the image plane assumes an additive Gaussian noise with independent standard deviations σ_x and σ_y in the two image coordinates

$$e_{2d} \sim \mathcal{N}(0, \Sigma_{2d}) \; ; \; \Sigma_{2d} = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}.$$

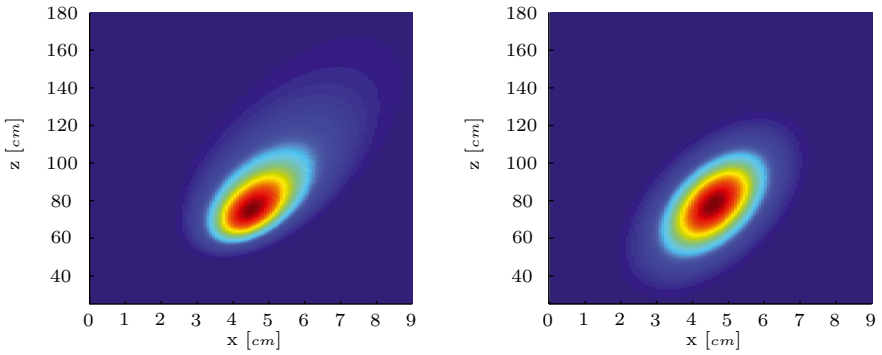

Fig. 7.4. Approximation of the 3D localization uncertainty using a normal distribution. The cut at $y = 0$ is used for visualization. For this example, the pan angles for left and right camera were set to 22.5° and 0°. A focal length of 6 mm and a baseline of 9 cm were used. The variance of 2D localization uncertainty was set to $\sigma = 16\,pt$ in the left and right camera. Left: Exact localization uncertainty. Right: Approximation of uncertainty using a normal distribution.

Given two corresponding points in the left and right image plane $\mathbf{b}_l$ and $\mathbf{b}_r$, the 2D localization errors lead to uncertainty in 3D localization. In Fig. 7.4 an example of the distribution of the 3D localization uncertainty in the xz-plane of the camera coordinate frame is illustrated. The projection to the camera plane was determined using the camera projection matrices P_l, P_r. Assuming statistical independance of the 2D localization uncertainties, the conditional probability for the sampled 3D location $\mathbf{a}_i$ being generated from a feature which has been localized at the 2D points $\mathbf{b}_{0,l}$ and $\mathbf{b}_{0,r}$ with the uncertainty Σ_{2d} is given with

$$p(\mathbf{a}_i|\mathbf{b}_{0,l}, \mathbf{b}_{0,r}) = p_{2d}(P_l\mathbf{a}_i|\mathbf{b}_{0,l})p_{2d}(P_r\mathbf{a}_i|\mathbf{b}_{0,r}).$$

The 2D uncertainty $p_{2d}(\mathbf{u}|\mathbf{b}_0)$ is defined by the normal distribution $\mathcal{N}(\mathbf{b}_0, \Sigma_{2d})$.

As can be seen in Fig. 7.4, left the resulting 3D error does not clearly follow a normal distribution. Nevertheless, the approximation with a normal distribution as illustrated in Fig. 7.4, right still yields good results in practice [Olson et al., 2003, Matthies and Shafer, 1990]. According to [Hartley and Zisserman, 2004], the parameters for the normal distribution have to be determined in a manner that preserves important properties of the true distribution, such as the increase of uncertainty with decreasing angles between the projection axes.

The calculation of a 3D position $\mathbf{a}$ from 2D correspondences involves stereo triangulation. Let G be the function that performs stereo triangulation based on the calibrated kinematic model in a given system state x_t

$$\mathbf{a} = G(\mathbf{u}_{0,l}, \mathbf{u}_{0,r}, x_t).$$

Since G is non-linear, normally distributed errors in the 2D locations do not result in a normally distributed error in the 3D localization. In order to approximate the resulting 3D uncertainty with a normal distribution, G is linearized using the unscented transform in this work. The unscented transform determines samples of the input distribution at predefined points, the so-called sigma points [Julier and Uhlmann, 1996]. In this work, sigma points are calculated and passed through the system in terms of the system function G and the approximated normal distribution is derived from the transformed sigma points. Instead of generating sigma points for the left and right 2D uncertainty separately, the sigma points are determined based on the combined 4D uncertainty. After passing the resulting sigma points through the system, the normal distribution in 3D space is recovered resulting in the 3D localization error

$$e_{3d} \sim \mathcal{N}(0, \Sigma_{3d})$$

and the recovered mean μ_{3d}.

In summary, the measurement model for a single object candidate $z_{k,t}$ conditioned on the n-th landmark m_n and the system state x_t is approximated by

$$p(z_{k,t}|x_t, m_n) \sim \mathcal{N}(\mu_{3d}, \Sigma_{3d}).$$

7.2.3 Inference of Memory Content

The goal of the inference process consists in calculating the posterior over the map given a sequence of control commands and measurements up to time t:

$$p(m|u_{1:t}, z_{1:t}). \tag{7.2}$$

In order to derive an exact solution for the map m from a sequence of observed controls $u_{1:t}$ and measurements $z_{1:t}$, marginalization over the complete sequence $x_{1:t}$ is required. This optimal solution is computationally intractable due to the high dimensionality of the problem arising from the continuous sample space of the map and the correspondence problem. Hence, the posterior has to be approximated with suitable methods.

The model derived in the previous sections has significant similarity to the simultaneous localization and mapping (SLAM) problem. In the following, we resort to methods which have been proposed in the SLAM context and apply a similar principle to the problem at hand. The SLAM problem consists in estimating the posterior over robot poses $x_{1:t}$ and the map m given the movement of the robot $u_{1:t}$ and the measurements $z_{1:t}$. In contrast to the model proposed for the accumulation of object candidates, the pose of the robot at two different points in time x_{t-1} and x_t depends, since uncertainty in the previous pose also affects the updated pose in the series. Nevertheless, some of the most mature approaches in the SLAM literature can be adapted

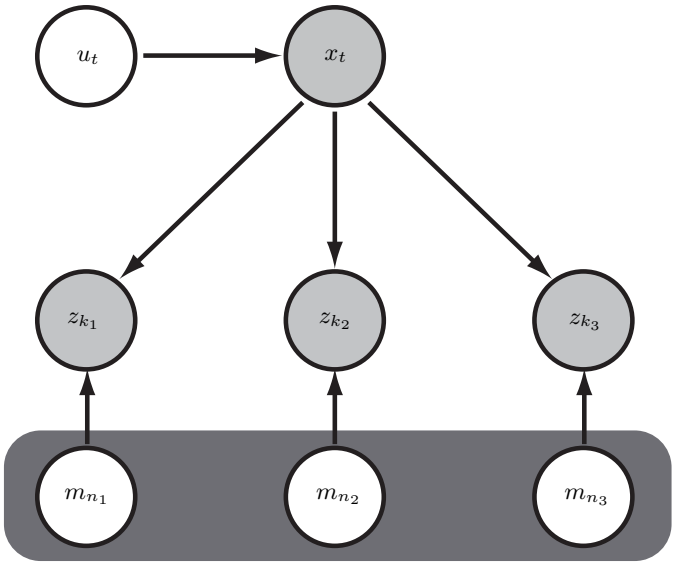

Fig. 7.5. Given the knowledge of the head pose x_t, all observed object candidates can be updated independently. Every path between two object candidates is blocked in the graphical model.

for the accumulation of spatial information of object candidates. The approach proposed in the following is derived from the FastSLAM approach according to [Montemerlo et al., 2002]. We apply principles which are at the core of the FastSLAM approach to the probabilistic model proposed in the last section.

As starting point for the inference of the desired posterior, the following factorization of the joint posterior of the map m and the sequence of head poses $x_{1:t}$ is applied

$$p(m, x_{1:t}|u_{1:t}, z_{1:t}) = p(x_{1:t}|u_{1:t}, z_{1:t})p(m|x_{1:t}, z_{1:t}). \tag{7.3}$$

The inference of the posterior can be broken down into a set of much simpler estimation problems by exploiting a conditional independence intrinsic to the model. As already proposed in [Doucet et al., 2000], we assume to have knowledge of the head poses $x_{1:t}$ over the complete time-series. In the SLAM context, it has been shown, that using this assumption, the update of object candidates within the map can be performed independently [Montemerlo, 2003]. The derivation of the independence assumes the performance of only one measurement in each time-step. This assumption is only valid, if the states are dependent over time. Since in our model this dependency does not exist, it remains to show that the knowledge of the head pose $x_{1:t}$ allows to treat the measurements of multiple object candidates independently.

The consequences of the knowledge of the poses on the estimation of the map when observing multiple object candidates can be understood considering the graphical model illustrated in Fig. 7.5. The Bayesian network models one update step of the map m. In contrast to the previous model in Fig. 7.3, the measurement z_t and the map m are broken down to single elements, each corresponding to separate object candidates in the scene. For now we assume that the correspondence problem has been solved and the correct correspondence between measured candidate and map is given by $m_i = k_i$. The goal is to further factorize the update of the map

$$p(m|x_{1:t}, z_{1:t})$$

as derived in Eq. (7.3).

As can be seen in Fig. 7.5, given the set of measurements z_t and the control u_t, the update of two different object candidates in the map depends via the path $m_{n_i} \rightarrow z_{k_i} \leftarrow x_t \rightarrow z_{k_j} \leftarrow m_{n_j}$. The knowledge of the state x_t overcomes this dependency since it blocks the path between measurements. Thus, each individual object candidate can be updated independently. Given this independence, the posterior in our model can be factorized with

$$p(m|x_{1:t}, z_{1:t}) = \prod_{n=1}^{N} p(m_n|x_{1:t}, z_{1:t}).$$

In summary, the knowledge of $x_{1:t}$ allows to break down the problem of inferring the map m into N smaller estimation problems. As shown in the SLAM context in [Thrun et al., 2005], the update for each object candidate can then be reformulated using Bayes rule

$$p(m_n|x_{1:t}, z_{1:t}) = \frac{p(z_t|m_n, x_{1:t}, z_{1:t-1})p(m_n|x_{1:t}, z_{1:t-1})}{p(z_t|x_{1:t}, z_{1:t-1})}.$$

By exploiting further independences the posterior can be simplified to

$$\begin{aligned} p(m_n|x_{1:t}, z_{1:t}) &= \frac{p(z_t|m_n, x_t)p(m_n|x_{1:t-1}, z_{1:t-1})}{p(z_t|x_{1:t}, z_{1:t-1})} \\ &= \eta p(z_t|m_n, x_t)p(m_n|x_{1:t-1}, z_{1:t-1}), \end{aligned} \tag{7.4}$$

thus each object candidate stored within the map can be updated independently using the measurement model as described above.

Solution using Rao-Blackwellized Particle Filters

In order to achieve the above factorization and derive the map posterior the knowledge of the states $x_{1:t}$ was assumed. Conditioned on the system state, the resulting estimation problem can be solved analytically. This technique is referred to as Rao-Blackwellization [Blackwell, 1947]. Having derived the posterior conditioned on the system state leaves the question how to solve our initial problem. Murphy et al. proposed the application of particle filters to estimate the posterior over $x_{1:t}$ [Murphy and Russell, 2001]. In order to infer our initial problem $p(m|u_{1:t}, z_{1:t})$ from Eq. (7.2) we make use of this Rao-Blackwellized particle filter approach. The following approach is derived from [Thrun et al., 2005] and adapted to the model proposed in this work.

Sampling from the proposal distribution

In a first step, the proposal distribution for x_t which takes only into account the current control u_t is represented using a set of samples where each sample x_t^d is drawn according to the motion model

$$x_t^d \sim p(x_t|u_t).$$

Each sample is associated to a particle in a set of D particles. The sampled pose x_t^d is appended to the history of poses $x_{1:t-1}^d$ already stored in the particle. Further, each particle contains an estimate for all object candidates in the map resulting from the previous estimation step. For each object candidate, the particle contains its position uncertainty encoded with a normal distribution $\mathcal{N}(\mu_n^d, \Sigma_n^d)$. Overall each particle d is represented using

$$Y_t^d = \left(x_1^d, \cdots, x_t^d, (\mu_1^d, \Sigma_1^d), \cdots, (\mu_N^d, \Sigma_N^d)\right).$$

Update of the maps

The maps which are stored with each particle are distributed according to

$$p(m_t^d|x_{1:t-1}, z_{1:t-1}).$$

Each map is updated with the current measurement. This involves first the estimation of the correspondence between measurements from the set z_t and object candidates in the map. The correspondence problem is solved using the maximum a posteriori (MAP) object candidate for each measurement. Furthermore, the scene feature $\mathbf{s}$ stored in the map is compared with the measured features $\mathbf{s_l}$ and $\mathbf{s_r}$ using the similarity measure $D(\mathbf{s}, \mathbf{s}_{l|r})$. Based on the MAP estimate and this similarity, the correspondence is determined.

Having determined the correspondence, the update of the object candidate's location is then performed according to equation 7.4 using one Kalman filter

per correspondence m_i and $z_{j,t}$. The Kalman filter is defined by the following process and measurement model

$$m_i = Am_i$$
$$z_{j,t} = Hm_i + v.$$

Since the above Kalman filter does not consider motion of the map elements, the process model does not involve a position update. Assuming no motion, the position uncertainty of the predicted map location m_i and its uncertainty do not change, thus $A = I$. The spatial uncertainty of an object candidate after the prediction step is then equal to the spatial uncertainty within the map. In order to derive the correction step, the matrix H and the measurement noise R with $p(v) \sim \mathcal{N}(0, R)$ need to be determined. Since measurement and map lie within the same reference frame, the matrix H is chosen to be identity. The 3D uncertainty of localization yields the measurement noise $R = \Sigma_{3d}$. Overall, the Kalman filter performs the update step

$$\hat{m_i} = m_i + K(z_k - m_i)$$
$$\hat{\Sigma_i} = (I - K)\Sigma_i$$

with the Kalman gain

$$K = \Sigma_i(\Sigma_i + \Sigma_{3d})^{-1}.$$

Importance Resampling

The update of the maps has been performed using the proposal distribution. Assuming that particles have been distributed according to $p(x^d_{1:t-1}|z_{1:t-1}, u_{1:t-1})$, the sampling from the motion model $p(x_t|u_t)$ results in a distribution according to

$$p(x^d_{1:t}|z_{1:t-1}, u_{1:t}).$$

Given the initial factorization in Eq. (7.3), the target distribution for particles is defined as

$$p(x^d_{1:t}|z_{1:t}, u_{1:t}).$$

In the resampling phase an importance weight is calculated for each particle, which corrects the distribution of the particles in accordance to the target distribution. The importance weight in particle filtering is calculating using the quotient of target and proposal distribution

$$w^d = \frac{p(x^d_{1:t}|z_{1:t}, u_{1:t})}{p(x^d_{1:t}|z_{1:t-1}, u_{1:t})} = \eta p(z_t|x^d_t).$$

This probability and thus the importance weight can be expressed in terms of innovation derived from the difference of the predicted measurement and the observed measurement. For a more in-depth discussion and the derivation of the previous equation the reader is referred to [Montemerlo, 2003].

Extension to changing scenes

In the model for memory update as introduced in Section 7.2.2 the map m was assumed to be static. This restriction poses two problems for the active visual search: First, a humanoid robot usually acts in environments, which are highly dynamic. An approach which does not account for changing scenes is not likely to be applicable in real world scenarios. A more serious issue arising from the assumption of a static map m consists in its inability to generate new representations of object candidates on the fly. When starting a new visual search task, no prior is given for the map m. Thus, while the scene itself is not changing the absence of prior knowledge requires to model change in order to allow for the generation of new object candidates in the preattentive memory.

In order to enable visual search in changing scenes, for each object candidate the probability of its existence c is stored as already discussed in Section 7.2.1. As proposed in [Thrun et al., 2005], the probability is stored as log odds ratio. For the generation of new object candidates in the map, the MAP likelihood determined during the calculation of correspondences is used as similarity measure. If the MAP likelihood lies below a threshold p_{min_c}, a new landmark is generated with mean and variance corresponding to the measurement uncertainty μ_{3d} and Σ_{3d}. The probability of existence is initialized to c_{new}.

Each time a new measurement is performed, the probability of existence is updated accordingly. The approach discussed by Thrun et al. increases the probability by a fixed amount c_{obs}, if an object candidate has been observed within the set z. The probability for object candidates which are visible to the robot but are not included in the set of measurements are decreased by the amount c_{miss}. In order to model the possibility of change in the scene this approach is extended. The possibility of existence for each object candidate which is not visible in the current measurement is reduced by the probability of unobserved change c_{unobs}. Finally, in order to provide fast reaction to change, the probability of existence is restricted to a maximum c_{max}. While using a maximum for the probability does not agree with the probabilistic framework, it accounts for the fact that given a changing scene a defined amount of uncertainty about the content remains.

7.2.4 Recovery of Memory Entities

The content of the preattentive memory consists of memory entities as defined in Section 7.2.1. Each memory entity corresponds to one object candidate within the map m. In order to update the content of the memory entities based on the inferred map m, two problems need to be addressed: First, the expectation value of the posterior over m has to be recovered from the set of particles and second the correspondence between object candidates within

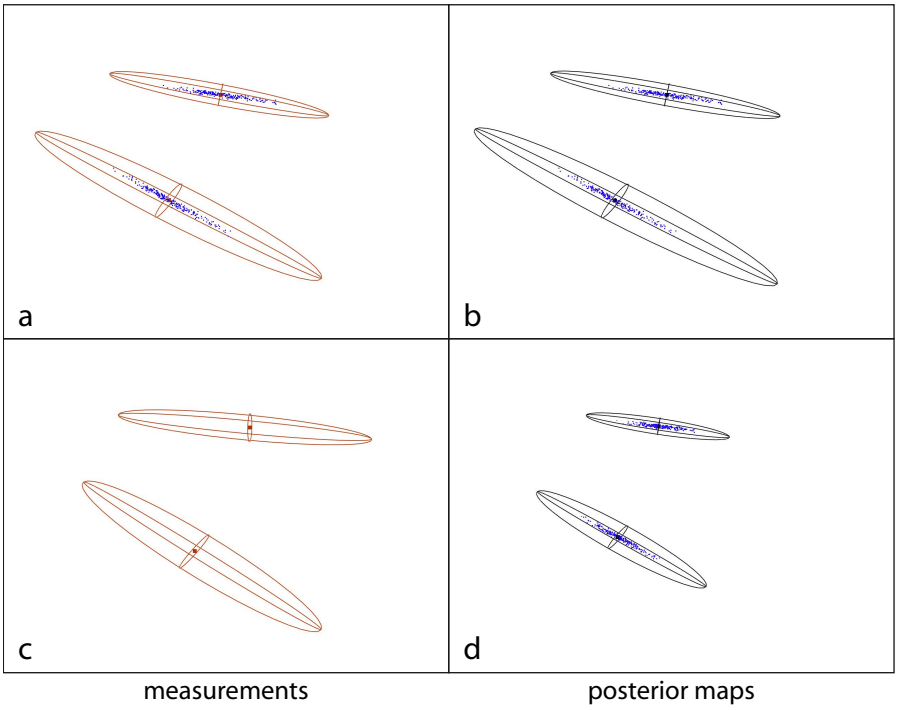

Fig. 7.6. Recovery of the posterior map. A mapping task with two object instances is used as example. a: The first observation and the resulting distribution of particles. b: The posterior map is determined by the expectation value over all particles. c: Second performed measurement of the object instances. d: Resulting posterior map after integrating the second observation.

the map and memory entities has to be determined in order to update the memory content.

The recovery of the map involves the calculation of the expectation value over the inferred joint distribution over m and x_t. A common approach is the calculation of the means of the particles Y_d weighted by the importance weight w_d

$$E[m, x_t] = \frac{1}{N} \sum_{j=1}^{N} w^d(m^d, x_t^d). \tag{7.5}$$

For the head pose in terms of the joint angles included in the system state x_t the previous equation can be applied directly. The derivation of the expectation value over maps is more involved. In order to apply Eq. (7.5) to the maps represented with each particle the correspondence between object candidates stored within different particles has to be determined. Since the

correspondence of measurements and object candidates is estimated per particle, object candidates stored within two different particles can result from a different set of measurements. Each particle represents a different estimation of correspondences between object candidates and measurements. Thus, the expectation value over maps can only be recovered approximatively.

In order to build correspondence between object candidates within different map estimates, we assume that the correspondences between measurements and candidates are solved identically in all maps. Therefore, a unique identifier is assigned to each measured object candidate z_k. The measurements of object candidates in the scene are numbered consecutively, resulting in a unique identifier per measurement. A new object candidate in the map inherits the identifier of the measurement that is responsible for its creation. In the recovery of the expectation value, object candidates from different particles are assumed to correspond if the assigned unique identifiers match. Further, only object candidates are considered in the calculation of the expectation value, where a sufficient amount of particles includes estimates with the same unique identifier. For each object candidate meeting this requirement, the expectation value is calculated according to Eq. (7.5). The recovery of the expectation values is illustrated in Fig. 7.6.

Having determined the expectation value the preattentive memory is updated. In order to build correspondences between memory entities and object candidates again the previously introduced unique identifier is drawn on. If no memory entity exists with the unique identifier of an object candidate contained in the set, a new entity is created. Further, memory entities are removed if their unique identifier is not present in any of the maps represented by the particles.

Each memory entity is updated based on the information available from the object candidate and the search task. The object candidate provides the estimate of location $\mathcal{N}(\mu, \Sigma)$ and existence c. The most recent measurement which has been associated to the object candidate provides the signature $\mathbf{s}$ and the match w. The feature of the target object $\mathbf{o}$ is provided by the search task itself.

7.2.5 Experimental Evaluation

All experiments were carried out on the Karlsruhe Humanoid Head, equipped with lenses of focal length $f = 4\,\text{mm}$. A series of saccadic eye movements was performed using random movements for the eye joints. The eye pan movements were selected within the range $\theta_{epl}, \theta_{epr} \in [-10°, 10°]$, the eye tilt movements were generated within the range $\theta_{et} \in [-10°, 10°]$. For all experiments the same motion model was used. The joint angle conversion error and the positioning noise were determined empirically. The positioning noise was set to the

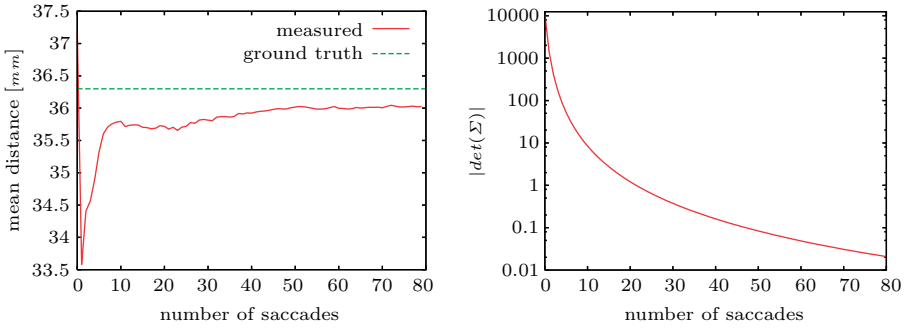

Fig. 7.7. Results of the mapping using chessboard corner points after 80 saccades. Left: Mean distance between neighbored corner points during 80 saccadic eye movements. The ground truth was measured with 36.3 mm. Right: Mean of the uncertainty ellipsoid volume for all chessboard corners during 80 saccadic eye movements in log scale. Reprinted from [Welke et al., 2009] ©2009 IEEE.

conservative value $\sigma_p = 0.15°$ per joint. Additionally, a joint angle conversion error of $\sigma_c = 0.1$ was assumed. The calibration error has been analyzed in Section 6.1.4, the observed standard deviation amounts to about $\sigma_{cal} = 4.0\,\text{mm}$. For all experiments $D = 200$ particles were used.

In the following, three experiments are described and discussed. While the first two experiments validate the mapping of features using ground truth, the third experiment demonstrates the performance of the approach in conjunction with object candidate detection as described in Section 5.1.

Experiment 1: Mapping of corner features

In order to test the convergence of the map towards the observed scene, a chessboard rig with known size was deployed for the first experiment. As ground truth, the distance of chessboard corners was drawn on. For chessboard patterns out-of-the-shelf corner point detectors can be applied with subpixel accuracy (see [Mallon and Whelan, 2007]). For this experiment we did not use signatures to describe the features, the correspondence are only built based on their spatial location and uncertainty. Therefore, the similarity measure is set to $D(\mathbf{s}, \mathbf{s}_y) = 1$ for the calculation of correspondences. Under consideration of the experiments performed by Mallon et al., a standard deviation $\sigma_x = \sigma_y = 0.5$ pixel for the uncertainty of corner point localization is realistic and has been used for this experiment. Overall 80 random saccadic eye movements were performed and the covariance matrix Σ was recorded for all object candidates. Further, the distance between neighbored corner points in the inferred map was calculated for each saccade.

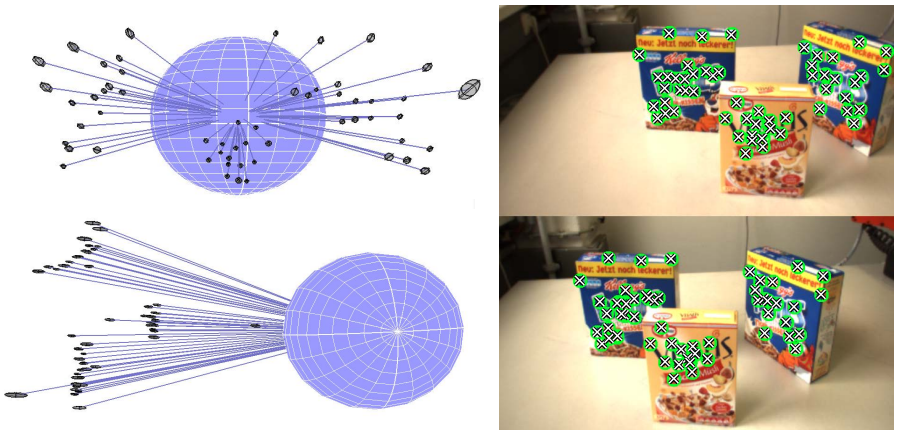

Fig. 7.8. Results of the mapping using SIFT features. Left: Best map of SIFT features after 20 saccadic eye movements in the egocentric coordinate frame. The resulting landmarks lie along the visible plane of the object. Right: Reprojection of the map onto the images of left and right camera. Reprinted from [Welke et al., 2009] ©2009 IEEE.

The 48 landmarks could be tracked through all 80 saccadic eye movements. Fig. 7.7, right illustrates how the mean volume of the uncertainty ellipsoid $|det(\Sigma)|$ converges over the iterations of the particle filter. After 80 saccadic movements, the volume of the ellipsoid amounts to about $0.021\,\mathrm{mm}^3$. In Fig. 7.7, left the mean distance of neighbored corner points over the 80 saccadic eye movements is illustrated. The manually measured distance of corner points on the chessboard amounts to 36.3 mm. The mean distance as calculated from the particles converged to about 36.0 mm. The difference between the mean distance and the ground truth of about 0.8% results from unmodeled phenomena such as inaccurate intrinsic camera parameters.

Experiment 2: Mapping of texture features

In the second experiment the focus is shifted to a more realistic mapping task. The goal consists in the mapping of corner features of textured objects. In order to calculate object features, the Harris-SIFT approach as introduced in Section 5.2 was deployed. For the experiments a standard deviation of $\sigma_x = \sigma_y = 1.5$ pixel was used for the uncertainty of 2D localization. For the similarity of SIFT descriptors $D(\mathbf{s}, \mathbf{s}_y)$, the Euclidean cross-correlation was used.

Since no ground truth is available concerning the absolute or relative position of these features, examples of resulting mappings are provided in Fig. 7.8, left.

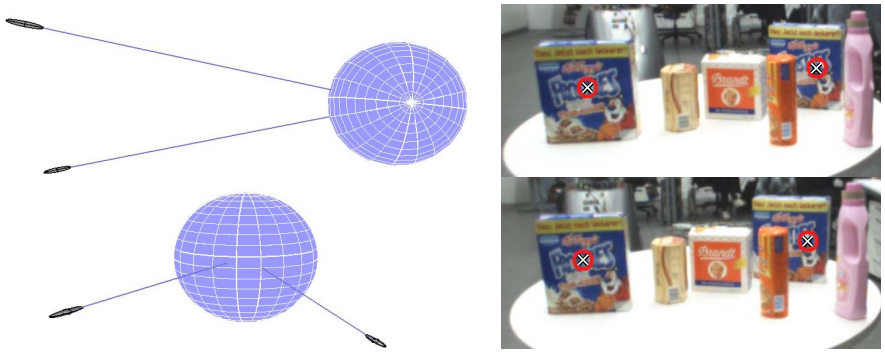

Fig. 7.9. Results of the mapping of object candidates using LCCH features. Left: Best map of object candidates after 80 saccadic eye movements in the egocentric coordinate frame. Right: Reprojection of the map onto the images of left and right camera. Reprinted from [Welke et al., 2009] ©2009 IEEE.

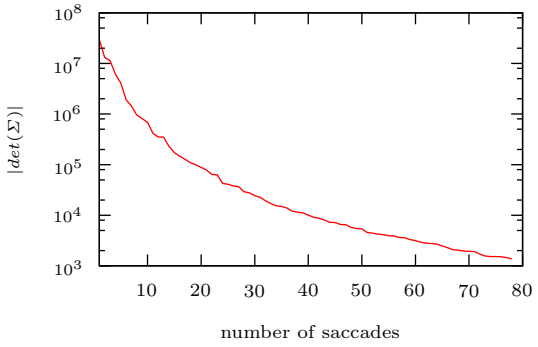

Fig. 7.10. Mean volume of the uncertainty ellipsoid all landmarks during 80 saccadic eye movements in log scale. The uncertainty is successively reduced with each new observation.

The figure illustrates how all three objects produced a set of 3D landmarks which have only a small deviation from the common plane after 20 saccadic eye movements. Four outliers have been mapped which result from erroneous correspondences between the left and right images. In Fig. 7.8, right the projection of the resulting map to the images of the left and right camera are illustrated.

Experiment 3: Object candidate mapping

In the third experiment, the mapping performance of the approach using peripheral object candidate detection as discussed in Section 5.1 was evaluated.

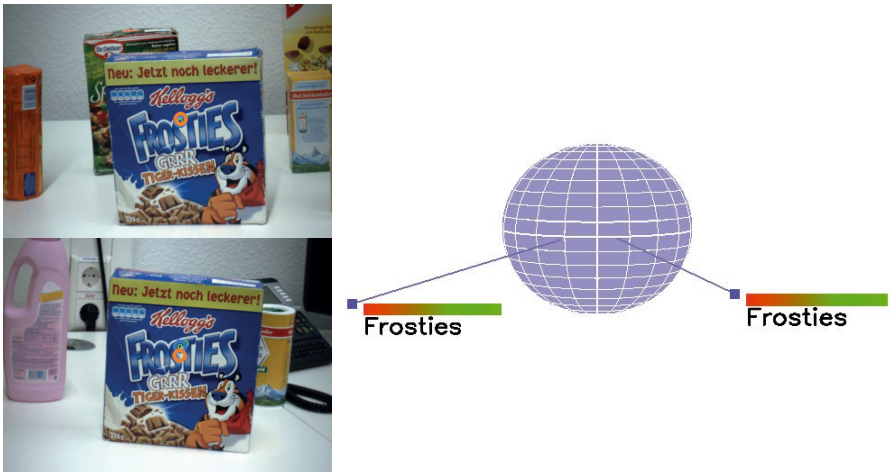

Fig. 7.11. Example of object memory content. Left: Foveal images corresponding to the fixation points of two object candidates. Right: Object memory holds an entity for each of the objects accompanied with its probability of existence.

Thus, the experiment reveals the feasibility of the approach for the mapping of object candidates in preattentive memory. The detection of object candidate was performed with the window based technique discussed in Section 5.1.3, based on the LCCH descriptor. The size of the detected regions was used as estimate for the standard deviations of the 2D localization uncertainty σ_x and σ_y. As similarity measure $D(\mathbf{s}, \mathbf{s}_y)$ the histogram intersection was used.

The experiment has been carried out using two instances of the target object in the scene. Further, a set of distractor objects was positioned in the scene. Overall 80 saccadic eye movements were performed and the mean volume of the uncertainty ellipsoid $|det(\Sigma)|$ was recorded for all landmarks. In Fig. 7.10 the decrease of the volume during the sequence of saccadic eye movements is illustrated. Compared to the previous experiments, the resulting uncertainty is much higher due to the high 2D localization uncertainty. Nevertheless, the object candidates could be reliably mapped through the complete course of the experiment resulting in a decrease of uncertainty.

The resulting map of two object candidates after 80 saccadic eye movements is illustrated in Fig. 7.9, left. As can be seen in the back projection of the map (see Fig. 7.9, right), both produced landmarks correspond well to the spatial location of the searched cereal box in the peripheral views.

7.3 Object Memory Layer

The object memory layer of the transsaccadic memory holds information about objects recognized in the scene. In contrast to the preattentive memory layer which collects information from the peripheral cameras, the object memory layer collects information from the foveal cameras as observed during different fixations. Using the detailed views of the foveal camera system, object recognition is performed and a consistent representation of the scene with respect to the query object is accumulated over different fixations. Examples of foveal views and the corresponding content of the object memory layer is depicted in Fig. 7.11.

7.3.1 Object Memory Entities

Entities in the object memory correspond to objects which have been recognized using foveal object recognition. Similar to the preattentive memory layer, information about the location and the existence of objects are stored with each entity. Additionally, each entity is associated to corresponding object candidates in preattentive memory. In the following, the stored information is describe in detail.

- **Identity of the object**
  In order to provide a solution to the object search problem which reflects the uncertainties in the object recognition step the probability of identity p_i is stored for each recognized object.
- **Location of the object**
  In contrast to preattentive memory, the spatial location in object memory is encoded with the joint angles $\boldsymbol{\theta} = (\theta_{epl}, \theta_{epr}, \theta_{et})$ of the active system. The joint angles $\boldsymbol{\theta}$ correspond to the pose of the eyes that centers the object instance in both camera images. The corresponding Cartesian position $\mathbf{x}_{\boldsymbol{\theta}}$ can be recovered from the joint angles using the calibrated kinematic model (see Section 6.1). Since the calibration is subject to inaccuracies, the joint angles provide a more reliable answer to the object search problem.
- **Link to preattentive memory**
  For each recognized object instance, the links to one or multiple preattentive memory entities are stored. The link l_i allows to associate objects with candidate regions in the scene and is necessary for the generation and removal of entities as discussed in Section 7.4.

In summary, an entity in object memory M_o holds information about location and identity of object instances and the link to corresponding preattentive memory entities

$$M_o = (\boldsymbol{\theta}, p_i, (l_1, \ldots, l_m)) \,. \tag{7.6}$$

7.3.2 Object Memory Update

Entities in the object memory are updated using measurements from the foveal camera pair. Once the fixation point has been reached after the execution of a saccadic eye movement, the foveal images are processed and an update of the object memory content is performed.

Before the object memory is updated, entities that are currently visible to the foveal cameras are determined. Therefore, the spatial location $\mathbf{x}_{\boldsymbol{\theta}}$ is projected to the left and right camera plane using the calibrated kinematic model. The visibility check is based on the distance from the resulting 2D positions $\mathbf{u_l}$ and $\mathbf{u_r}$ to the image centers $\mathbf{c_l}$ and $\mathbf{c_r}$:

$$vis_{l|r} = e^{-\frac{(\mathbf{u}_{l|r}-\mathbf{c}_{l|r})^T \Sigma_{vis} (\mathbf{u}_{l|r}-\mathbf{c}_{l|r})}{2}},$$

where Σ_{vis} is a diagonal matrix of variances. If the visibility for one of the cameras falls below the threshold vis_{min}, the corresponding entity is not considered during the update. An update of the probability of existence and the location is performed for all visible entities.

Update of the identity

The believe of identity for an object p_i is given by the conditional probability

$$p_i = p(O = 1|Z_{f,1:t}),$$

where $Z_{f,1:t}$ represents all observations of the object using the foveal cameras up to the current time t. The probability of identity is updated recursively with each measurement using the Bayes filter

$$p(O = 1|Z_{f,1:t}) = \eta p(Z_{f,t}|O = 1)p(O = 1|Z_{f,1:t-1}). \tag{7.7}$$

In order to perform the Bayes update, the sensor model $p(Z_f|O = 1)$ needs to be specified. Since the underlying object recognition approach as introduced in Section 5.2.2 is not formulated in a probabilistic manner, a mapping from the score of object matching s_f is performed. For this purpose, a sigmoid function is used to achieve a mapping in the range of $]0; 1[$. The resulting sensor model is then given by

$$p(Z_f|O = 1) = \frac{1}{1 - e^{-a(s_f - b)}},$$

with scale a and the offset b. The scale a, the offset b and the marginalization constant η are chosen in a way that a fixed threshold can be established for the presence of an object in the scene. In contrast to preattentive memory entities, the probability of existence is not directly available from the filtering step. In order to make a decision for the existence of the object, this threshold on the identity is drawn on.

Update of the position

The position of objects is refined in a closed-loop fashion using the 2D localization resulting from foveal object recognition. The goal is to successively correct the gaze direction in order to fixate the center of the object with the foveal cameras. For this purpose, the error in the image plane between the center $\mathbf{c}$ and the object center $\mathbf{u}$ is calculated. The location of the object center is recovered from the Hough space and the error is calculated using

$$\epsilon = \mathbf{u} - \mathbf{c},$$

where $\mathbf{u}$ corresponds to the mean position of all entries in the maximum bin of the Hough space. Similar to the approach proposed in [Ude et al., 2006], a simplified kinematic model is used to iteratively reduce the errors for left and right camera ϵ_l and ϵ_r. The 3D position $\mathbf{x}_\theta$ is updated using

$$\mathbf{x}_{\theta,1} = \mathbf{x}_{\theta,0} + \begin{pmatrix} w_x & w_y & w_z \end{pmatrix} \begin{pmatrix} \frac{\epsilon_{r,x}+\epsilon_{l,x}}{2} \\ \frac{\epsilon_{r,y}+\epsilon_{l,y}}{2} \\ \epsilon_{r,x} - \epsilon_{l,x} \end{pmatrix}, \tag{7.8}$$

where w_x, w_y and w_z are gains for the movements in all three Cartesian directions. The closed-loop refinement of positions successively determines the joint angles $\boldsymbol{\theta}$ that allow to bring an object instance to the center of the foveal cameras.

Due to inaccuracies in object candidate detection multiple entities might be generated for one single object instance in the scene. In order to detect such cases, where multiple object candidates point to the same object instance, the distance between object memory entities is monitored after each position update. Cases where two entities point to the same object instance are detected based on their distance. If the distance falls below a threshold, the object memory entities are merged. The resulting merged object memory entity is initialized with the mean position, maximum probability of identity and the references to all preattentive memory entities as stored with the original entities.

7.4 Interplay between Memory Layers

In the previous sections the preattentive memory layer and the object memory layer were discussed separately, each in the context of its associated update mechanisms. While the update of both memory layers is performed independently, their interplay is an essential aspect of the transsaccadic memory. Based on the definition of the entities of preattentive and object memory, their interaction is discussed in the following.

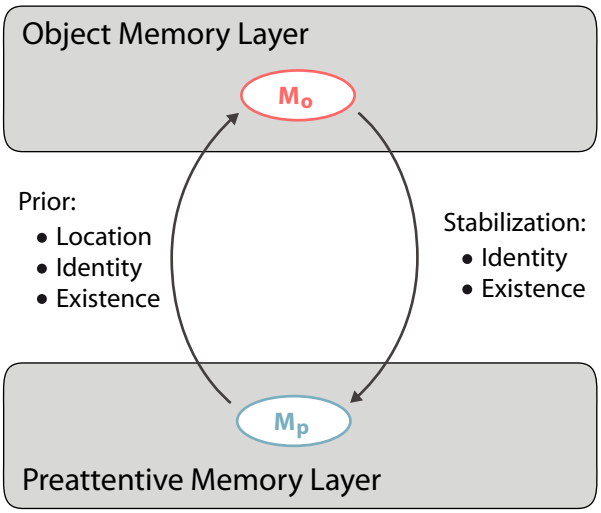

Fig. 7.12. Interplay between entities in preattentive and object memory. Both entities are connected via a bidirectional link. The preattentive memory entities provide prior knowledge for the object memory entities. In the opposite direction, the object memory entity stabilizes linked preattentive memory entities.

In Fig. 7.12 an example of the interaction between entities in preattentive memory M_p and object memory M_o is illustrated. Both entities are connected via a bidirectional link. This bidirectional link inherits many properties of the Coherence Field (see Section 2.2). The up-link from the preattentive memory entity serves as prior for the properties of the object memory entity. In the opposite direction, the preattentive memory entity is stabilized by the object memory entity. These mechanisms are discussed in detail in the following.

7.4.1 Preattentive Memory Entity as Prior

The starting point for the detection of previously unseen object instances is provided by the peripheral object detection mechanism. Successively, the correspondence problem is solved in order to generate or update entities in preattentive memory layers. Each entity in the preattentive memory layer corresponds to a potential instance of the target object. In order to ascertain the existence of the target object a saccade has to be performed in order to perform foveal object recognition. For object candidates that have not been verified by fixating, the saccade execution as well as the update of its existence probability requires prior knowledge. This prior knowledge is provided by the preattentive memory entity.

Once a saccade is executed toward an object candidate which has not been previously fixated, an object memory entity is generated which inherits the properties of the corresponding preattentive memory entity. While identity and location are inherited explicitly in terms of the encoded mean of the normal distribution μ and the probability of identity w the existence is implicitly encoded by the generation of an object memory entity. The initial position is used as starting configuration for the closed loop refinement using the foveal percept in Eq. (7.8). The probability of identity provides the prior for the Bayes update of the object identity according to Eq. (7.7).

7.4.2 Stabilization of Preattentive Memory Entities

Entities in preattentive memory are coupled directly to the peripheral object candidate detection mechanism. If a new object candidate is detected, it is immediately mapped in the preattentive memory layer and serves as prior for the object memory entity as discussed above. Now the opposite case of disappearing object candidates is considered. The disappearance of object candidates can result from several reasons. First, in dynamic scenes the object candidates might be removed or moved to another location in the scene. Second, due to the actuated eyes not all candidates are visible to the peripheral cameras all the time. The probability of existence might drop to zero if the object has not been visible for a long period of time. Finally, the coarse analysis of the scene is subject to noise resulting in a failure of candidate detection.

In order to prevent inconsistencies of the transsaccadic memory in the above cases, the object memory entity is used to stabilize the linked preattentive memory entities. Once an object memory entity has been updated, the corresponding preattentive memory entities inherit the existence and identity properties. While the object identity p_i is directly available from the memory entity, the existence is derived from thresholding as described in Section 7.3.2. As consequence of the stabilization, entities of preattentive memory which have been fixated once can only be removed by performing another fixation and verification. Only if this verification fails, the object memory entity and the linked preattentive memory entities are removed.

7.5 Summary

The transsaccadic memory was proposed in this chapter as spatial representation of the scene with respect to the target object. A hierarchical organization using two layers has been chosen which follows the principle of foveated vision. The preattentive memory layer accumulates object candidates, which result from peripheral vision. Its content is inferred using a probabilistic model which

allows to accumulate the location, the identity, and the existence of object candidates in a consistent fashion. Further, the correspondence problem is solved using the proposed model.

Updates of the object memory layer are accomplished using foveal object recognition in a closed-loop manner. The resulting representation provides the joint angles that allow to direct the gaze toward target object instances. Each object memory entity has preattentive memory entities linked via a bidirectional channel. Entities in preattentive memory serve as prior for the object memory. On the other hand, object memory entities stabilize linked preattentive memory entities in order to increase the robustness of the approach.

The objective of the memory-based active visual search approach consists in providing a consistent representation of the scene with respect to the target object in the object memory layer. The consistency of entities in the object memory layer are ascertained by executing saccades toward the corresponding location and performing foveal object recognition. The decision for the sequence of saccades is crucial for the memory consistency and is determined using methods of visual attention as discussed in the subsequent chapter.

8

Visual Attention

The decision where to look next is an essential aspect of memory-based active visual search. Following the principles of human visual perception, this process of deciding for the next fixation is usually termed visual attention. The application of visual attention mechanisms allows to restrict the search space thus making the problem of visual search computationally tractable. Having introduced the object recognition and detection approaches, the methods for execution of saccades, and the transsaccadic memory, the visual attention mechanism closes the action-perception cycle in terms of the decision for saccade targets. According to the objective of this thesis, the selection of targets should allow the recognition of all target object instances and should ascertain a consistent representation of the scene with respect to the target object.

As illustrated in Fig. 8.1, attention is generated based on the content of the transsaccadic memory. Additionally, events are detected based on the previously discussed update processes and are reported to the attention mechanism. These events include the validation of object memory entities using foveal object recognition and the detection of change based on the preattentive memory content. Based on this information, an attentional mechanism is introduced in the following that maximizes the probability of recognizing an object and that at the same time retains a consistent representation of the scene in the object memory layer.

According to the review of attentional mechanisms in Section 3.2, the Bayesian Strategy showed to be well-suited for technical implementations since it allows the integration of different sources for saliency. In the following, the formalization of saliency according to the Bayesian Strategy is introduced and extended by the requirement of a consistent memory in Section 8.1. Subsequently, a general model for the inference of inconsistency is discussed in Section 8.2. Finally, the model is further refined in the context of memory-based active visual search in Section 8.3.

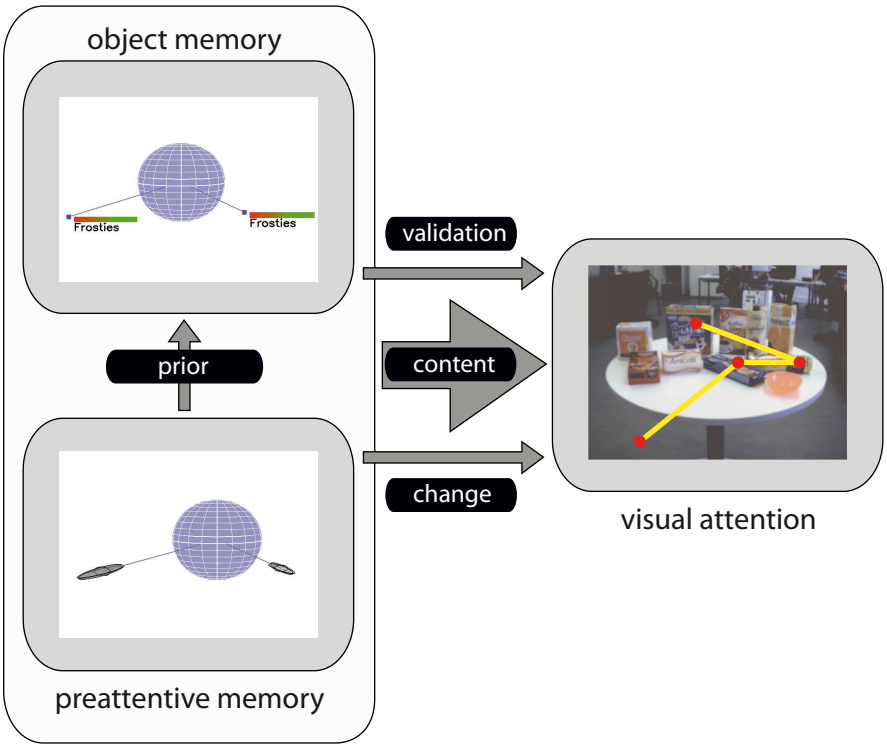

Fig. 8.1. The attentional mechanisms closes the perception-action cycle by generating target positions for the saccade execution based on the transsaccadic memory. Thereby, the memory content as well as the events of validation and change form the basis in order to decide for gaze directions that maximize the object recognition probability and the transsaccadic memory consistency.

8.1 Probabilistic Saliency and Active Visual Search

The approach for memory-based active visual search detects and recognizes object instances in a top-down fashion. The decision for applying only top-down cues is based on the fact that top-down search provides the most reliable answer to the search problem. Thereby, only top-down knowledge about the target appearance is exploited. While the search task is solvable using only this knowledge, the inclusion of other sources such as position priors for objects in the scene or bottom-up depth information allow a further restriction of the search space. Although not covered in this thesis, the goal is to provide an open approach which allows for the seamless inclusion of such factors.

The Bayesian Strategy allows to seamlessly integrate bottom-up factors with top-down guidance and scene priors [Torralba, 2003]. Thus, its formalization

of saliency is used as basis for this work. In the following, the basic definition is introduced and the active saliency is defined which extends the Bayesian Strategy by taking into account memory consistency.

8.1.1 The Bayesian Strategy

In the Bayesian Strategy as introduced in [Torralba, 2003] saliency is formulated in terms of the probability of detecting an object in the image. Given the image J, the probability of detecting an object O at the spatial location X is expressed using the conditional probability $S = p(O = 1, X|J)$, where S is the measure of saliency. The image J is decomposed into two observations: the local features F which are extracted at the spatial location $X = x$ and the global features G which are extracted from the complete image and represent the scene gist. By replacing the image J with the observations F and G the saliency measure according to Torralba becomes $S = p(O = 1, X|F, G)$.

One great appeal of the Bayesian Strategy consists in the fact that this conditional probability can be sub-divided into different factors and thus becomes

$$p(O = 1, X|F, G) = \frac{1}{p(F|G)} p(F|O = 1, X, G) p(X|O = 1, G) p(O = 1|G).$$

Each of the factors represents a different semantically meaningful aspect of saliency. The first factor $\frac{1}{p(F|G)}$ does not depend on the object and thus is a pure bottom-up factor. Features which are unlikely to be observed given the current scene will produce high saliency. The factor $p(F|O = 1, X, G)$ encodes the probability of observing local features F given the object O, the spatial location X, and the scene gist G. As such it corresponds to the top-down knowledge of the target appearance. The factor $p(X|O = 1, G)$ provides the possibility to include scene priors based on the spatial location depending on the scene gist G. This distribution can be learned e.g. from past experiences. Finally, the factor $p(O = 1|G)$ describes the likelihood of finding a target in the scene.

The benefit of the above formalization has been shown in a series of publications which deal with different aspects of the model. One line of research showed the feasibility of the model in the field of contextual guidance using scene priors [Oliva et al., 2003, Torralba et al., 2006]. Also, implementations of the bottom-up aspect of the model have been demonstrated in recent research (see e.g. [Zhang et al., 2008]).

8.1.2 Extension to Active Visual Search

In order to make use of the resulting saliency data for the generation of saccadic eye movements, a sequence of salient locations has to be determined

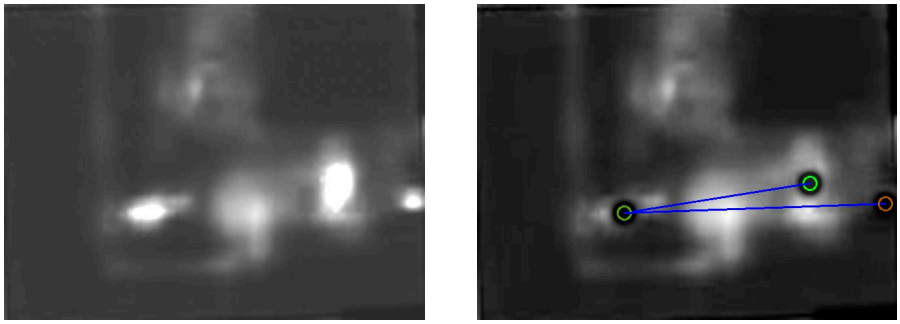

Fig. 8.2. The inhibition of return mechanisms allows to generate sequences of saccades on saliency maps. Left: Saliency map calculated with the approach proposed in [Ude et al., 2005]. Right: The inhibition of return mechanism inhibits the saliency in a region around attended locations equal to the size of the field of attention.

which are to be fixated by the active vision system. In the most common approaches, spatial locations are defined in the image plane $X = x = (u, v)$ and the saliency is determined for each pixel in the image resulting in a saliency map as depicted in Fig. 8.2, left. An intuitive way for determining a sequence of gaze shifts based on such saliency maps consists in calculating a list of positions corresponding to the n spatial locations with maximum saliency and visiting them in the order of saliency magnitude. The success of this approach is limited considering the distribution of saliency within saliency maps. As can be seen in Fig. 8.2, left maxima in the saliency map usually cover spread areas and the intuitive approach will report many spatially adjacent locations as saccade targets.

The most common approach for generating a sequence of saccades from saliency data which are spatially separated is based on the inhibition of return (IOR) mechanism as proposed in the Itty-Koch-Niebur model of visual attention (see Section 3.2). In this model inhibition is performed on the saliency map once a salient point is fixated in an area with the size of the field of attention (see Fig. 8.2, right). In conjunction with a winner-take-all network composed of leaky-integrate-and-fire neurons, the proposed model has been successfully applied in the context of bottom-up attention. The resulting order of fixations within the saliency map does not only depend on the amount of assigned saliency but also on the time of the last fixation performed at that location. Various modifications of the Itty-Koch-Niebur model have been proposed, nevertheless related models are still used in many state-of-the-art systems (see e.g. [Gratal et al., 2010]).

In the following, an approach for the generation of sequences of saccades is proposed which extends the Bayesian Strategy by an inhibition of return mechanism as introduced in [Welke et al., 2011]. This mechanism results from

the formulation of saliency which takes into account foveal and peripheral vision and the application of a transsaccadic memory in visual search. The basic requirement of a consistent memory representation of the real world stands at the core of the proposed saliency measure.

Similar to the Bayesian Strategy, saliency is defined as the probability of detecting an object O at a spatial location X. In order to include the requirement of a consistent memory, the random variable I is introduced which encodes the distribution of inconsistency between real world and memory representation. Given a measurement of the world J, saliency is then defined by

$$s_a = p(O = 1, X, I|J),$$

which will be referred to as *active saliency*.

Further, using an active system which provides peripheral and foveal vision, the world is not perceived using a single image as in the case of the original model. Rather the perception of the world is defined by the peripheral percept Z_p and the foveal percept Z_f. Object recognition is performed based on the foveal images, while peripheral vision in this model reports changes of the environment which might affect consistency of the memory. Consequently, the peripheral observation is then defined by the observation of change Z_c. The foveal object recognition results in a measurement Z_f which updates the believe of object existence and location $bel(O = 1, X)$. By updating the memory, foveal vision validates the visible memory entities. This validation process is represented with the random variable V. Reformulating the active saliency using the introduced observations yields

$$s_a = p(O = 1, X, I|Z_f, Z_c, V).$$

Assuming perfect knowledge of the current gaze direction, OX and I are conditionally independent given all measurements. Further, exploiting the independences from observations and hidden variables as imposed by the above model, the active saliency under consideration of peripheral and foveal perception and memory is given with

$$\begin{aligned} p(O = 1, X, I|Z_f, Z_c, V) \\ = p(O = 1, X|Z_f)p(I|Z_c, V). \end{aligned} \tag{8.1}$$

The first factor of the above model essentially corresponds to the Bayesian strategy as discussed in Section 8.1.1. The second factor allows to integrate the requirement of the consistency of transsaccadic memory and forms the basis for the inhibition of return mechanism. A detailed model for the second factor is defined in the following section, before the application of the active saliency in memory-based active visual search is discussed in Section 8.3.

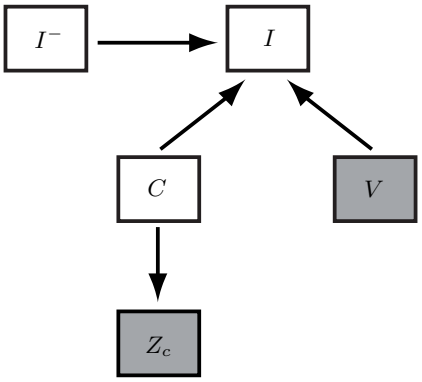

Fig. 8.3. Graphical model of the inconsistency filtering. Inconsistencies I are updated over time under consideration of the measurement of change Z_c and the validation V.

8.2 Model of Memory Inconsistency

There are generally two cases, where a memory that reflects the surrounding world becomes inconsistent: either the world changes or the memory itself is subject to decay. For the memory-based active visual search approach, perfect memory is assumed. The focus will be put on the case of a world that changes by the interaction of other agents or by its own dynamics. Nevertheless, the same model could also be applied for the latter case, where inconsistency arises from memory decay or a combination of both.

8.2.1 Probabilistic Inconsistency Update

Inconsistencies are filtered over time, a filtering step involves the integration of the observed change and occurred validation in order to derive the conditional probability $p(I|Z_c, V)$. In Fig. 8.3 the graphical model of all random variables involved in the update and their conditional dependencies are illustrated. The inconsistency of entities in memory is encoded with the random variable I. The model assumes a system, which is capable of validating entities in memory by means of focusing attention and ascertaining of consistency. The validation is modeled using the random variable V. Further, a change sensor is assumed which is capable of measuring if change in the world has occurred. The observed change is modeled using the random variable Z_c. While validation V and the change sensor measurement Z_c are observed, the goal is to infer the inconsistency I of memory entities based on priors for the inconsistency I^- and change C.

The following two paragraphs further define the model in Fig. 8.3 in terms of the sample spaces of all involved random variables and the factorization of the joint distribution.

Definition of spaces

All variables introduced in the following are defined for N memory entities. For each memory entity a binary random variable is used to encode its state.

- I: Encodes inconsistencies between memory and real world. Inconsistency either takes the value consistent (con) or inconsistent ($\neg con$), thus $I \in \{con, \neg con\}^N$.
- I^-: Inconsistencies at time $t-1$. $I^- \in \{con, \neg con\}^N$
- C: Change that occurred in the world as relevant for the memory. The world was either static ($\neg ch$) or changed (ch) with respect to the memory entity, thus $C \in \{ch, \neg ch\}^N$.
- Z_c: Change of the world as measured by the change sensor. $Z_c \in \{ch, \neg ch\}^N$.
- V: For each entity the variable V encodes whether validation has been performed for the respective entity in order to ascertain its consistency with the real world. $V \in \{val, \neg val\}^N$.

Factorization

Reviewing the graphical model depicted in Fig. 8.3, the following factorization of the joint distribution is given

$$p(I, I^-, C, V, Z_c) = p(I^-)p(I|I^-, C, V)p(Z_c|C)p(C)p(V).$$

As discussed above, each random variable encodes the distribution for all memory entities. In order to keep the formalization general, we do not state whether memory entities are stored in a occupancy grid manner, as is the case in probabilistic saliency map, or in a landmark based manner, as is the case in transsaccadic memory introduced in Section 7.3. For both representations, a common simplification is the assumption of independence between entities. Thus, the joint distribution over all variables can be factored to

$$p(I, I^-, C, V, Z_c) = \prod_{i=1,\ldots,N} p(I_i, I_i^-, C_i, V_i, Z_{c,i}).$$

According to Fig. 8.3, the inconsistency for each memory entity i can then be factored using

$$p(I_i, I_i^-, C_i, V_i, Z_{c,i}) = \underbrace{p(I_i^-)p(I_i|I_i^-, C_i, V_i)}_{\text{Prediction}} \underbrace{p(Z_{c,i}|C_i)p(C_i)p(V_i)}_{\text{Correction}}. \quad (8.2)$$

The model parameters and the approach for inferring the posterior over I_i are introduced in the following for the update of a single entity i.

8.2.2 Model Parameters

According to the factorization in Eq. (8.2), the required parameters can be assigned to the prediction step and the correction step. While the prediction step is defined by only one factor, the correction step involves the definition of two factors corresponding to the change sensor model and the change prior, respectively.

Prediction model

The prediction model is defined by the conditional probability of inconsistency I_i given the change C_i, the validation V_i and the inconsistency from the last filtering step I_i^-. This dependency is expressed by the conditional probability table

$$p(I_i = \neg con | I_i^-, C_i, V_i) = \begin{cases} 0, & \text{if } (I_i^- = con \wedge C_i = \neg ch), \\ 1 - p_v, & \text{if } (V_i = val \wedge I_i^- = \neg con \wedge C_i = \neg ch), \\ 1, & \text{else.} \end{cases}$$

In the above definition, the first statement covers cases where the consistency from the previous step is preserved since no change happened. Validation leads to consistency of the memory if no change happened and the memory was inconsistent, as stated in the second case. The parameter p_v allows to define a confidence for the validation success. In all other cases, the resulting inconsistency equals to one, thus change always overrides validation. This is a necessary statement since the order of the performed validation and measured change is not provided and thus change might have occurred after validation has been performed within the last time interval.

Change sensor model

The change sensor is modeled using the forward model $p(Z_{c,i}|C_i)$. We define the sensor model for change in a general manner using its sensitivity $w_{c,1}$ and its specificity $w_{c,0}$. The corresponding conditional probability table is given by

$$p(Z_{c,i} = ch|C_i) = \begin{cases} 1 - w_{c,0} & \text{if } C_i = \neg ch, \\ w_{c,1} & \text{if } C_i = ch. \end{cases}.$$

Change prior

Another parameter required to fully define the probabilistic model is the prior probability of change $p(C_i)$. The prior allows to encode the likelihood of change in the scene p_c. This probability can be used as top-down cue in order to instruct the system to perform more validations in cases where the scene is changing rapidly. The change prior is then defined by

$$p(C_i = ch) = p_c.$$

In summary, the model provides four free parameters which can be used to influence the inference of inconsistencies. The sensitivity $w_{c,1}$ and specificity $w_{c,0}$ of the change sensor as well as the validation probability p_v depend on the system built around the model. These parameters will be defined for the application of active visual search in Section 8.3. The change prior p_c allows to tune the model according to the volatileness of the current situation.

8.2.3 Inference of Inconsistencies

The inference of the inconsistency of memory entities is formulated based on the model defined in Section 8.2.1 and its parameters discussed in Section 8.2.2. Using the prior believe of inconsistency of a memory entity $bel(I_i^-)$, the observation of validation $V_i = v_i$ and of measured change $Z_{c,i} = z$, the posterior believe $bel(I_i)$ can be calculated by performing the marginalization

$$p(I_i|Z_{c,i} = z, V_i = v_i) = \frac{\sum_{I_i^-, C_i} p(I_i, I_i^-, C_i, V_i = v_i, Z_{c,i} = z)}{\sum_{I_i, I_i^-, C_i} p(I_i, I_i^-, C_i, V_i = v_i, Z_{c,i} = z)}. \tag{8.3}$$

The posterior believe is calculated using the factorization from Eq. (8.2).

8.3 Active Saliency in Visual Search

In summary, the model for active saliency has been formalized according to Section 8.1.2 with the factorization

$$s_a = p(O = 1, X, I | Z_f, Z_c, V) = p(O = 1, X | Z_f) p(I | Z_c, v). \tag{8.4}$$

Within this section, the complete model is put together in the context of active visual search. In Section 8.3.1, the first factor will be derived from the content of the object memory. The general model for filtering of inconsistencies discussed in the previous section will be further specified in the context of active visual search in Section 8.3.2.

8.3.1 Saliency from Top-Down Search

As already discussed, the first factor in Eq. (8.4) corresponds to the Bayesian Strategy to saliency. The Bayesian Strategy allows to integrate bottom-up, top-down and contextual guidance. The approach for memory-based active visual search discussed in this work is solely based on top-down information. Nevertheless, the compatibility with the Bayesian Strategy allows to seamlessly integrate contextual guidance e.g. using a world model or the integration of bottom-up information.

The top-down knowledge about searched objects in the scene is held in the object memory. As discussed in Section 7.3 the object memory is organized as a landmark-based map, where landmarks are represented by memory entities. Each landmark corresponds to a 3D location of the object which is updated in a closed-loop fashion. Thus, the position X is handled by the closed-loop controller and is not part of the inference of saliency. Thus, the simplified Bayesian Strategy can then be estimated using the Bayes filter

$$p(O = 1 | Z_f) = \eta p(Z_f | O = 1) p(O = 1).$$

The above posterior corresponds to the probability of identity as already introduced in Section 7.3.1 which is updated with each foveal measurement of an object instance. Therefore, we use the probability of object identity p_i as stored with each object memory entity M_o as first factor for the saliency measure.

8.3.2 Saliency from Inconsistencies

The saliency from inconsistencies is determined based on the model derived in Section 8.2. Similar to the calculation of top-down saliency, the calculation

of inconsistency is performed per landmark. For each memory entity the inconsistency is updated according to Eq. (8.3) separately. The update involves the observation of change Z_c and the validation step, which depends on the validation certainty p_v.

Preattentive memory as change sensor

Change is detected based on the preattentive memory content. As discussed in Section 7.2 the preattentive memory holds a landmark-based map of object candidates within the observed scene. The candidates are tracked over time by building correspondences between object candidates observed under different gaze directions of the peripheral cameras. An approach for handling changing scenes has been discussed in Section 7.2.3. The resulting ability to seamlessly add appearing object candidates and to remove disappearing candidates forms the basis for the change detection as used for the inference of inconsistencies.

Each time a new landmark is inserted or a landmark is removed from preattentive memory, the world as observed by the peripheral cameras is subject to change. The change is associated to all entities in the object memory that correspond to the changed entity in preattentive memory. For the sensor we assume perfect specificity $w_{c,0} = 1$. In order to take into account that change might occur which is not visible to the peripheral object detection an imperfect change sensor is modeled. Therefore, a sensitivity of $w_{c,1} = 0.9$ is used to express the possibility of unobserved change.

Validation certainty

The validation of the object memory content is performed by directing the gaze to memory entities and performing object recognition based on the foveal images. The most reliable verification of the content is achieved, if the object of interest is centered in the foveal images. The more distant the object is located from the center of the image the more likely parts are not visible to the camera resulting in an inferior validation performance. The model for the validation certainty takes into account this property of decreasing validation performance toward the borders of the foveal images. Let the 2D location of the object within left and right image be defined by $\mathbf{u}_l$ and $\mathbf{u}_r$. The validation certainty for each image is then determined using

$$p_{v,l|r} = e^{-\frac{(\mathbf{u}_{l|r}-\mathbf{c}_{l|r})^T \Sigma_v (\mathbf{u}_{l|r}-\mathbf{c}_{l|r})}{2}},$$

where $\mathbf{c}_l$ and $\mathbf{c}_r$ are the centers of left and right image and Σ_v is a diagonal matrix of variances. The overall validation certainty p_v is defined by the product of the validation certainties of both cameras

$$p_v = p_{v,l} p_{v,r}.$$

Given the above model of validation uncertainty introduces correction saccades in the saliency mechanism. If an object is not fixated properly, the validation is not executed with the necessary certainty and the associated entities will still exhibit significant active saliency. Since the position is updated using a closed-loop control scheme, which guides the gaze of the system toward the center of the object, the next saccade is more likely to fixate the center of the object and will result in an increased validation certainty.

8.4 Summary

The introduction of the visual attention mechanism in this chapter rounds up the memory-based active visual search approach by closing the perception-action cycle. In accordance with the objective of this thesis, the drive for the guidance of attention stems from the requirement of recognizing instances of the target object and at the same time retaining consistency of the transsaccadic memory. These two aspects are encoded in the active saliency measure as formalized in this chapter. Based on the notion of observed change and validation performed by directing the gaze toward memory entities, the resulting inhibition of return mechanism integrates seamlessly with the Bayesian Strategy to visual attention. By integrating with the Bayesian Strategy, the proposed active saliency is not only applicable for top-down search tasks but is open for the inclusion of all factors which are covered in the formalization of the Bayesian Strategy.

9

Evaluation

The previous five chapters discussed all relevant mechanisms in order to perform memory-based active visual search. Where applicable, the individual approaches were already evaluated in terms of their feasibility and reliability. In this chapter, the complete approach is evaluated including all discussed subproblems. Thereby, the focus is put on the ability of identifying objects in the scene and retaining a consistent memory of object instances using the Karlsruhe Humanoid Head.

First, in Section 9.1 the experimental setup used in all experiments is described including the hardware configuration, choices of methods and parameters. Subsequently, in the first set of experiments the capability to identify object instances in the scene is evaluated and discussed in Section 9.2. The consistency of the resulting memory is subject to the experiments in Section 9.3. Thereby, the accuracy of the location of entities as well as the ability to react to changing scenes is evaluated. In Section 9.4 the runtime of the system is analyzed before the results are discussed in Section 9.5.

9.1 Experimental Setup

All experiments were carried out on the Karlsruhe Humanoid Head as introduced in Section 4.1. The fixation of targets as generated by the attention mechanism was performed using the three DoF of the eye system. For each of the pan joints and the common tilt joint a range of motion within the interval $[-25°, 25°]$ was used. All cameras were configured to use fixed gain, exposure and white balance in order to avoid shifts in color space due to automatic adjustment of these parameters. The peripheral camera pair was equipped with lenses of 4 mm focal length and the foveal camera pair with 16 mm lenses. The large focal length of the foveal camera pair restricts the depth of sharpness. For the visual search tasks, the focus of the cameras was adjusted in order to

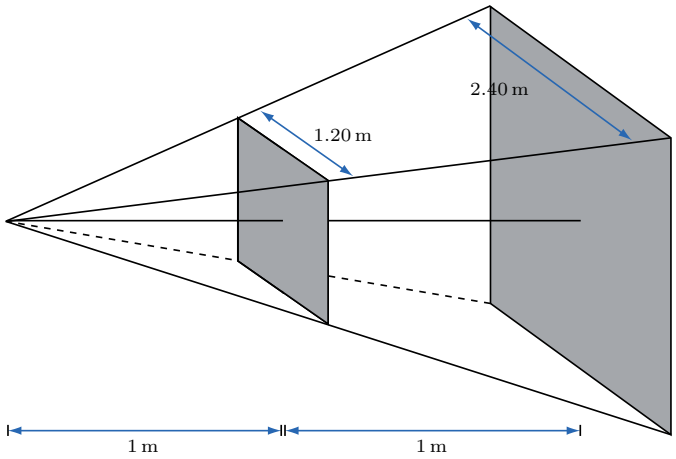

Fig. 9.1. Resulting frustum which is covered by the active camera system using the foveal cameras. The depth of the frustum is restricted by the depth of sharpness, all other borders correspond to joint limits of the eye joints. The system was setup to cover a frustum ranging from 1 m to 2 m in depth. An area of 1.20 m × 1.20 m at the near plane and 2.40 m × 2.40 m at the far plane can be covered by the active system.

provide sufficient sharpness within the range from 1 m to 2 m distance from the cameras. The resulting frustum which defines the area in space where objects can be recognized using the foveal cameras is depicted in Fig. 9.1. The region of the near plane at 1 m which can be observed with the active cameras amounts to 1.20 m × 1.20 m. In a distance of 2 m, the corresponding region amounts to about 2.40 m × 2.40 m. For all experiments, the target objects were positioned within this frustum.

Several parameters have been defined in the previous chapters. In the following, the most important parameters are described as used throughout the experiments. The choice of parameters is motivated by the experiments performed for the different subproblems in the previous chapters.

The detection of object candidates in the peripheral camera pair was performed using the NLCCH descriptor based on the Hue channel. As already evaluated in Section 5.1.3, reasonable detection performance can be achieved using 30 clusters. The matching of descriptors was performed based on a grid of 40 × 30 windows, matches were accepted if their intersection exceeded a threshold of 0.5. The optimal grow factor for the calculation of object candidate regions $p_{grow} = 0.85$ as evaluated in Section 5.1.3 was used for the experiments. Regions were accepted if their activation was within 35% of the maximum activation. For the recognition of objects in the foveal camera pair, a minimum cornerness of $C_{min} = 0.01$ was used for the detection of Harris

interest points. A resolution of 5×4 bins was used for the Hough space in order to recognize objects.

The camera system was calibrated offline as described in Section 6.1. The inaccuracies of the translational part of the calibrated model was evaluated with $\sigma_{cal} = 4\,\text{mm}$. This inaccuracy together with the positioning noise $\sigma_p = 0.15°$ and the conversion error $\sigma_c = 0.1$ form the parameters for the mapping of preattentive memory entities as discussed in Section 7.2.5. In order to provide robust inference, 150 particles were used in the Rao-Blackwellized particle filtering step.

The gaze sequence generated by the attention module depends on the parameters of the sensor model for observing change and the priors for validation and change (see Section 8.2). For the change sensor, a specificity of $w_{c,0} = 1$ and a sensitivity of $w_{c,1} = 0.9$ was used in order to account for the possibility of unobserved change. For the validation prior p_v, a validation certainty with variances $\sigma_x^2 = 150$ and $\sigma_y^2 = 100$ pixels was used. The change prior was set to $p_c = 0.05$. A saccade was generated once the active saliency s_a exceeded the threshold $s_{a,max} = 0.8$.

9.2 Active Visual Search

In the active visual search series of experiments, the goal was the detection of the target object in a scene containing several objects. Detection was considered successful, if the object could be brought to the view of the foveal cameras and the recognition based on the foveal images succeeded. In order to determine success of foveal recognition, a threshold for the probability of existence amounting to 0.4 was used.

Before the active visual search is evaluated in a table top setup (Section 9.2.2) and in a kitchen setup (Section 9.2.3), the trained target objects used as search queries are introduced in the next section.

9.2.1 Object Set

As targets for the active visual search tasks objects were selected which exhibit all necessary visual properties in order to enable detection and recognition based on the proposed approaches. More precisely, typical objects found in a kitchen were chosen which offer a significant amount of textural information. Furthermore, all selected objects were box shaped — the planarity of objects allows a wider range of out-of-plane rotational invariance using the proposed Hough space approach. In Fig. 9.2 the 10 selected objects used in the active visual search experiments are illustrated. The object set includes instances

1: spaghetti 2: butter cookies 3: cereals 4: choco cereals 5: corny

6: pasta 7: orange juice 8: coffee 9: vanilla cookies 10: vitamin juice

Fig. 9.2. 10 objects were used as target object during the active visual search experiments. For training, the objects were manually segmented and processed by the feature extraction methods in an offline step.

which exhibit very similar color signatures in order to demonstrate the ability of the system to cope with multiple hypotheses.

All objects were captured with the peripheral cameras and manually segmented offline. The resulting views were processed with the descriptors corresponding to peripheral and foveal vision approaches. In the case of the NLCCH descriptor, the objects were downscaled by a factor of three before calculating the descriptor.

9.2.2 Table Top Setup

For the first series of active visual search tasks a table top setup was chosen. The objects were positioned on top of a table, a second layer in height was added using a simple board. The head was positioned at a height of about 1.40 m slighty looking down at the scene. The wall behind the table has a distance to the head of about 2 m and thus marks the end of the depth of sharpness.

The 10 objects from the trained object set were distributed in the frustum in 10 different arrangements. For each arrangement, a search task was executed for each object resulting in overall 100 tasks. The search was stopped, once the target object could be recognized in the foveal cameras. In order to assess the benefits of the proposed approach, a random observer was implemented

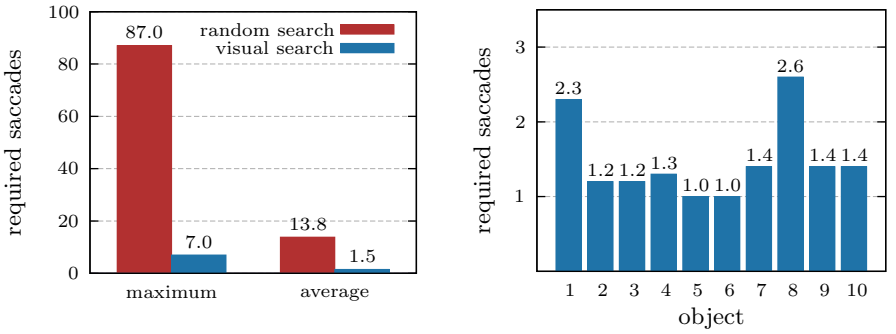

Fig. 9.3. Results of 100 search tasks in the table top setup. Left: The maximum and average number of saccades required until the target object is recognized in the foveal camera pair. The memory-based active visual search approach outperforms a random observer by the factor of 9.4 on average. Right: Average number of saccades per object. For objects with unique color signatures only one saccade is required. For some objects multiple hypotheses have to be scanned before the target object is recognized.

for comparison. For this purpose, random positions were generated within the frustum and the gaze of the foveal cameras was pointed toward these random positions. Again the recognition was performed based on the foveal images and the search was stopped once the object could be recognized.

All 100 search tasks could be accomplished without any false positives or false negatives. In order to assess the performance of active visual search, the number of saccades required to fixate and recognize the target objects were recorded for all search tasks and were compared to the performance of the random observer. The results are depicted in Fig. 9.3, left. The proposed active visual search approach took 1.47 saccades to recognize the target object on average. The random observer on the other hand required 13.84 saccades to fixate and recognize the object. Thus, the resulting relative improvement in the search performance amounts to a factor of 9.4 on average. A similar improvement in search performance is visible in the maximum number of saccades required to recognize the target where the ratio between random and visual search is 87 to 7.

In the evaluation, the visual search task was considered successful, if the object could be recognized within the foveal cameras. Using the proposed approach for foveal vision, the object can be recognized even with only a small subregion visible to the cameras. In order to further compare the random observer and the visual search approach, the sum of distances of the target object to the centers of left and right image was measured for each successful recognition. Due to a constant offset in the vertical direction between left and right camera image resulting from imprecise mounting of the cameras, the minimum reachable sum of distances amounts to about 155 pixels. In the

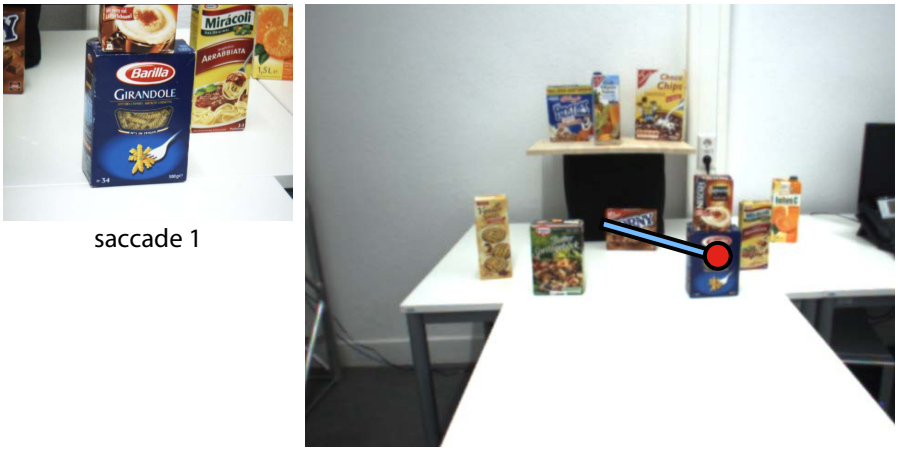

Fig. 9.4. Scan pattern in the pasta box search task. The pasta box (object 6) could be brought to the foveal cameras with a single saccade in all test cases.

case of active visual search, the mean sum of distances for all search tasks amounts to about 181.9 pixels. In contrast, the random observer brought the object to a position of the foveal images at a sum of distances of 342.4 pixels. This difference shows, that the proposed update mechanism of preattentive memory provides a good estimate for the initial saccade landing point. The refinement of this initial position toward the optimal position in successive verifications steps is evaluated in Section 9.3.2.

As already introduced in the last section, the selected object set contains similar objects in terms of their color signature in order to demonstrate how the system deals with multiple hypotheses. In order to further investigate the search performance, Fig. 9.3, right provides a closer look at the average number of required saccades for each object. Two objects (e.g. objects 5 and 6) could be brought to the foveal camera within a single saccade in all trials. These objects have a unique signature in terms of their NLCCH descriptor. For other objects, such as objects 1 and 8, several saccades were required until successful recognition. An example for both cases is illustrated in Figs. 9.4 and 9.5. The pasta box (object 6) could be brought to the fovea with only one saccade. In contrast, in order to fixate the spaghetti box (object 1), the system successively generates saccades to regions of the scene that are similar to the NLCCH descriptor of the target object until the active visual search task succeeds. On each saccade, one object candidate of preattentive memory is ruled out until the target object is reached.

Fig. 9.5. Scan pattern in the spaghetti box search task. The spaghetti box (object 1) does not have a unique signature among the object set. Three saccades are required before the object is recognized.

9.2.3 Kitchen Setup

In a second series of experiments, a more involved scenario was used for the active search task. The target objects were distributed in a typical kitchen scene including different places such as the sideboard, the cupboard, and the fridge as illustrated in Fig. 9.6. The head was positioned at a height of about 1.70 m looking straight at the scene which was set up at a distance of about 1.90 m from the cameras. The search task in this scene is more involved since objects are visible at a very small scale only and the lighting varies in different locations of the enhanced visual field. The object set was retrained in the kitchen environment in order to allow detection in the presence of significantly different lighting conditions compared to the table top setup.

Again the experiment was repeated 10 times for all 10 target objects using different arrangements of the objects. The results are illustrated in Fig. 9.7.

Fig. 9.6. Example of the kitchen setup. The target objects were distributed on the sideboard and in the cupboard and fridge.

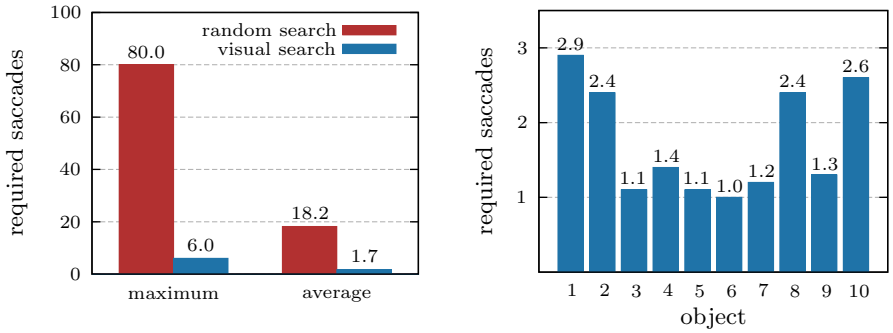

Fig. 9.7. Results of 100 search tasks in the kitchen setup. Left: The maximum and average number of saccades required until the target object is recognized in the foveal camera pair. The memory-based active visual search approach outperforms a random observer by the factor of 10.7 on average. Right: Average number of saccades per object. In contrast to the table top setup, the objects 2 and 10 are affected significantly by the changing lighting conditions.

In the kitchen setup, the average number of saccades required to recognize the target object in the foveal camera pair amounts to 1.67. Similar to the table top setup, the visual search approach outperforms the random observer by a factor of 10.7. The analysis of the number of saccades for each object is shown in Fig. 9.7, right. For most objects the results are very similar to the table top setup. For the butter cookies box (object 2) and the vitamin juice (object 10) about one saccade more was required on average. The color descriptors of these objects are more affected by the different lighting conditions resulting

episode 1 episode 2 episode 3

Fig. 9.8. The experiment for consistency validation comprises three different episodes. In each episode the scene was changed by adding or removing an instance of the cereal box (object 3) marked by the red border.

from locating them in the fridge, on the cupboard, or on the sideboard. The accuracy of the saccade landing position in the experiment was again measured using the sum of distances between center of the foveal images and object center. For the active visual search, an average sum of distances of 180.13 pixels was measured. The quality of the landing point of the random observer was measured with an average sum of distances from the image center of 343.48 pixels.

9.3 Memory Consistency

In the previous section the capability of the system to search for query objects in the presence of distractors was evaluated. For each search task, only one valid target object was present in the scene. Furthermore, the scene was kept constant for both setups. In this section, the focus is put on the ability of the system to retain a consistent memory of the scene with respect to the target object. The requirement of consistency implies the evaluation of two different aspects of the memory. First, the ability to generate and remove memory entities on the fly if change in the scene happens is evaluated. Successively, the ability of the system to refine the spatial location of stored entities towards the desired joint angles is demonstrated.

9.3.1 Consistency of Memory Entities

The goal of the memory-based active visual search task is to retain a consistent memory representation of the scene with respect to the target object. In order to show the feasibility of the proposed model, a search task was chosen which allows to illustrate the properties of the generated gaze sequence and the resulting foveal verifications. Therefore, two instances of the cereal box (object 3) were successively brought into the view of the peripheral cameras.

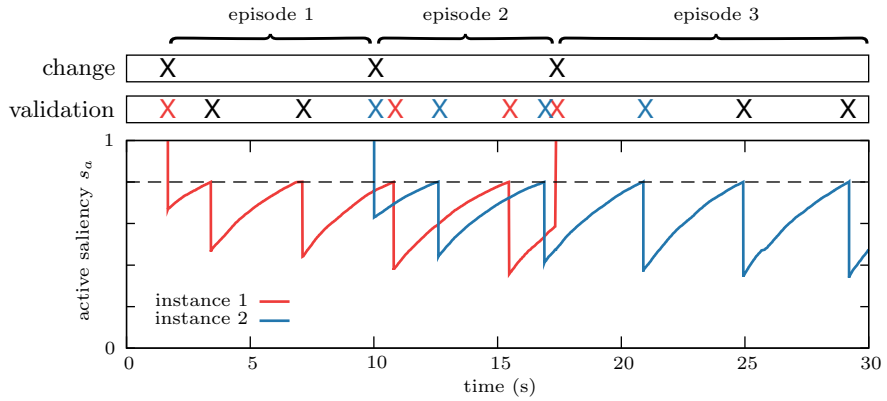

Fig. 9.9. During the memory consistency experiment, two instances of the cereal box (object 3) were visible to the system. Depending on the current measurement of change and performed validation, the active saliency is inferred. Once the active saliency exceeds the threshold $s_{max} = 0.8$ a saccade is executed.

Figure 9.8 shows snapshots of the scene from the course of the experiment. The experiment consisted of three episodes corresponding to the snapshots: In the first episode, one object instance was brought to the view of the system. For the second episode, an additional instance was positioned in the scene. Finally, the first instance was removed from the scene in the third episode. Using the proposed model, the active saliency s_a for the memory entities corresponding to visible object instances was continuously updated. Once the saliency exceeded the threshold $s_{max} = 0.8$ a saccade towards the corresponding memory entity was executed and foveal object recognition was performed in order to validate the object memory content.

The time course of the active saliency as monitored during the experiment is illustrated in Fig. 9.9. The active saliency as defined in Section 8.3 is recorded for the memory entities which correspond to the two object instances. Each update of inconsistency is performed using a measurement of change Z_c and an observation of validation V. Time steps where either change was detected or validation was performed are marked with crosses in the corresponding bars in Fig. 9.9. In the case of measured change, the inconsistency of the corresponding memory entities rises to one and the active saliency increases according to its definition. Validation is performed once the active saliency of a memory entity exceeds the threshold s_{max}. In the following the different episodes of the experiment and their counterparts in Fig. 9.9 are explained.

Each episode of the experiment starts with the adding a new object or removing and object from the scene which results in the detection of change. Since the resulting active saliency exceeds the threshold s_{max}, an immediate validation is scheduled. In the first two episodes where object instances have been

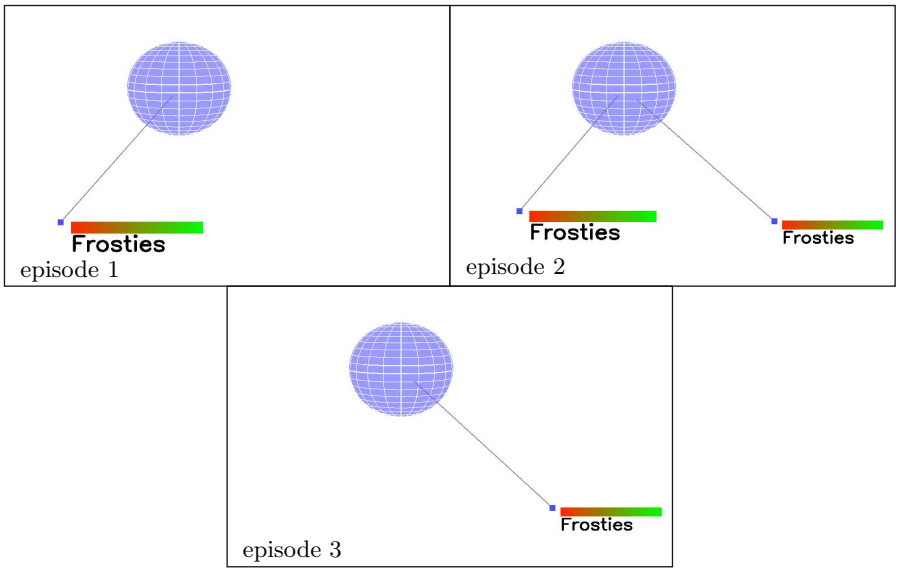

Fig. 9.10. Content of object memory in the three episodes of the experiment. During the first two episodes, two object instances are successively brought into the visual field of the active head. In the third episode, one instance is removed from the visual field. The system succeeds in generating a consistent memory of the scene with respect to the object through the detection of change and the generation of a gaze sequence according to the active saliency.

added to the scene, the validation reduces the saliency of the corresponding memory entities. In the third episode where one object instance has been removed, the validation results in removal of its memory entity. In the course of the experiment, the memory entities are validated frequently. The frequency is defined by the prior of change p_c, the sensitivity of the change sensor, and the success of the last performed validation. Each time the active saliency exceeds the threshold s_{max}, validation rescheduled performed. The increasing success of validation in terms of lower remaining active saliency results from the closed-loop control scheme used for active visual search. The closed-loop control moves the object closer to the center of the cameras in each validation step thus allowing for a more accurate validation in terms of the validation certainty p_v. This behavior is further evaluated in the context of consistency of locations in the subsequent section.

The object memory content in all three episodes is depicted in Fig. 9.10. The first detection of change leads to the validation and storage of the cereal box object. After adding another instance, the object memory is updated resulting in two memorized instances. Finally, one instance is removed which also results in the removal of the corresponding memory entity after foveal validation.

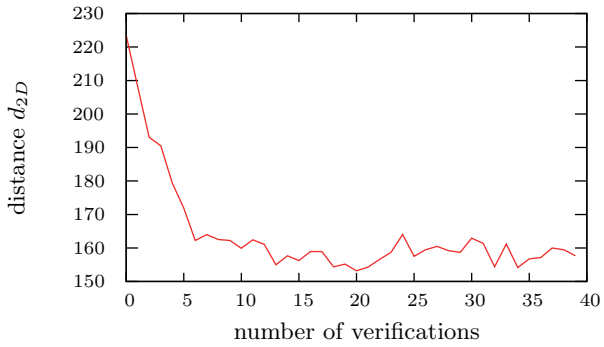

Fig. 9.11. Sum of distances of the center of left and right foveal camera and the object. The location stored for each object memory entity is refined in order to bring the corresponding object instance to the center of the images in a closed-loop fashion. The minimum of about 155 pixels is reached within 5 foveal validations of an object memory entity. The remaining distance results from a fixed offset in panning of left and right foveal camera.

9.3.2 Consistency of Locations

The preceding section focused on the ability of the system to generate memory entities for each instance of the target object present in the enhanced visual field of the active head. It remains to show that the associated information about the object location is stored in a consistent fashion. As already discussed in Section 7.3.1, the object memory stores the position information for object instances as the motor code $\boldsymbol{\theta}$ which points the foveal cameras towards the detected object instances. This motor code is refined in a closed-loop fashion using foveal object recognition; the information about the stored position for each entity is consistent if the closed-loop approach converges.

In order to demonstrate the behavior of the closed-loop approach, the sum of distances of the left and right foveal camera centers to the object center d_{2d} is measured over successive validations in the previous experiments as depicted in Fig. 9.11. Each foveal object recognition corrects the joint angles in order to bring the center of the foveal cameras closer to the object instance. Starting at a sum of distances of 225 pixels on landing of the first saccade, the distance is reduced to 160 pixels within 5 validations. The distance then settles between 155 and 165 pixels. The remaining distance of about 155 pixels results from an offset of the left and right foveal camera panning. Since the relative pan is not controlled, the resulting joint angles constitute the optimal trade-off between right and left distance. Since the object position is subject to noise, the sum of distances does not stabilize but varies within a range of about 10 pixels.

9.4 Runtime Analysis

All proposed approaches in this work were evaluated according to their suitability for the problem of memory-based active visual search so far. Although real time performance of the derived algorithms has not been a requirement for this work, their applicability in real world scenarios depends heavily on the runtime. The proposed system couples perception and action, the resulting loop has to be executed with a sufficient cycle time in order to allow fast reaction to changes. In order to demonstrate the applicability, an analysis of the runtime is presented and discussed in the following.

For the analysis, the runtime was measured separately for the different components. As depicted in Table 9.1 the main components for the runtime analysis are foveal and peripheral vision, preattentive and object memory, and visual attention. For each component, the most time consuming operations were identified and their runtime was analyzed. For the runtime evaluation, a scene containing two cereal boxes was used as reference. The runtime for the individual operations depends on the structure of the scene and the search task. In order to make the results more expressive, Table 9.1 contains a column which shows the dependency between the runtime and the most influential parameter of the operation. The dependency is either linear, denoted with the letter L or quadratic, denoted with the letter Q. Furthermore, the actual value of the parameter is given in brackets which has been observed during the experiment. The runtime analysis has been performed on a 3GHz Intel(R) Core(TM)2 Quad system. The components were running in parallel, each at its own cycle time, exploiting the parallel cores of the processor. In the following, the runtime for all components and the corresponding operations is discussed in detail.

The main operations in peripheral vision include the matching of the query features in the current input images, the extraction of candidate regions from the matches based on the additive region image, and the calculation of stereo correspondences. The feature matching is performed on a 30×40 cell grid. Search windows of a size ranging from 2×2 cells up to 5×5 cells were compared with the query feature. The search is performed on the left and right image. For each image overall 19200 matches between query feature and scene image regions were calculated. In the reference scene, this matching procedure took overall 328 ms. Based on the matches, the vote matrix and maxima are determined. For each maximum, a region growing is performed on the grid. In the reference scene, two such regions are extracted per peripheral image corresponding to the two instances of the cereal box. This operation took 11 ms. For both instances, stereo correlation is accomplished within overall 178 ms. The computation time of the stereo correlation depends on the size of the objects quadratically, which in the reference scene amounts to 1470 pixels on average. Including miscellaneous operations, the peripheral vision calculations took overall 541 ms on the reference scene.

description	runtime	dependency
peripheral vision	541 ms	
feature query	328 ms	# images (L) [2] # windows (L) [19200]
additive region image	11 ms	# images (L) [2] # object regions (L) [2]
stereo correspondences	178ms	# object regions (L) [4] # pixels per object (Q) [1470]
preattentive memory	40 ms	
forward model	32 ms	# object measurements (L) [2] # particles (L) [150]
landmark update	7 ms	# object measurements (L) [2] # object candidates (L) [2] # particles (L) [150]
foveal vision	706 ms	
feature extraction	117 ms	# images (L) [2] # scene features (L) [1733]
feature matching	584 ms	# images (L) [2] # scene features (L) [1733] # model features (L) [1355]
Hough space	1 ms	# images (L) [2] # matched features (L) [825]
object memory	< 1 ms	
existence update	< 1 ms	-
position update	< 1 ms	-
visual attention	< 1 ms	
saliency update	< 1 ms	# memory entities (L) [2]

Table 9.1. Runtime analysis of the different components and the major operations. For each operation, the runtime is determined on a reference scene. The dependency between the runtime and the most influential parameter of the operation is shown in the third column, where the letter L denotes a linear dependency and the letter Q denotes a quadratic dependency. The actual value of the parameters in the reference scene are given in brackets.

The object measurements form the basis for the inference of the spatial map in the preattentive memory layer. For each object measurement, a forward model is calculated taking into account the configuration of the current particle. For 150 particles and two object measurements, the forward model calculation took 32 ms. For each particle, the resulting 3D uncertainty is subject to the maximum-likelihood estimation of corresponding preattentive memory entities. Therefore, the likelihood has to be determined for each measurement and for each stored object candidate. The estimation of correspondences and the resulting update of the map took 7 ms. Overall, the preattentive memory update took 40 ms.

The overall runtime of peripheral vision and map update defines the delay for detecting change in the scene. In all experiments in this evaluation, two measurements of a preattentive memory entity were necessary in order to reach the required log-oddsa ratio for change detection. Thus, two passes of peripheral vision and map update were required. The resulting delay of change detection amounts to about 1162 ms. This delay in the detection of change is also visible in Fig. 9.9.

Foveal object recognition is only performed once the foveal camera is pointed toward an object instance via saccadic eye movements. The involved main operations include the extraction of Harris-SIFT features, the matching of the scene and the query features and the recognition of the object based on the Hough space approach. The extraction of 1733 scene features in the left and about the same amount in the right foveal image took 117 ms. The matching operation involved the scene features and 1355 model features and could be performed within 584 ms for both images. For each of the about 850 matches between scene and model features per image, voting is performed in the Hough space. This operation took approx. 1 ms. In summary, the foveal vision for one foveated object instance took 706 ms. Having performed recognition, the object memory is updated. This involves the filtering of the object‘s existence and the update of the location using the closed-loop approach. Both operations were executed within less than 1 ms each.

The visual attention component calculates the active saliency for all entities of the object memory. The update of the active saliency is performed at a cycle time of 10 ms and was executed within less then 1 ms.

9.5 Summary

In the previous sections, the two aspects of visual search and memory consistency were evaluated and the runtime of the system was analyzed. In order to assess the ability of the system to recognize the target object in the scene, overall 200 search tasks were executed in 2 different realistic setups using 10 different objects. All search tasks could be reliably accomplished by the memory-based active visual search approach within less then 2 saccades on the average. Compared to a random observer the improvement of search efficiency amounts to a factor of about 10. For the validation of the active saliency approach to visual attention, a setup using two object instances in a changing scene was used. The exemplary gaze sequence illustrates how the active saliency is able to deal with memory inconsistencies. The reaction to changes could be shown by adding and removing objects in the scene. The desired behavior of foveal verification in cases of inconsistencies could be demonstrated. Finally, the consistency of the locations encoded in the object memory entities was measured. The closed-loop approach successively brought the target object instances to the center of the foveal cameras resulting in a consistent motor code representation of the spatial location.

10

Conclusion

The objective of this thesis was to provide humanoid robots with the capabilities to actively guide their gaze in order to build an inner model with respect to relevant objects. Thereby, the focus was put on purposeful guidance of the gaze using a foveated vision system for object detection and recognition and the accumulation of gathered information about objects in a consistent representation. In this chapter, the contribution of this work is summarized and extensions and future work are discussed.

10.1 Contribution

The necessity for approaches to solve the active visual search problem stems from the active and foveated nature of camera systems integrated in state of the art humanoid platforms. By enhancing the field of view using moving cameras and by equipping the system with a high resolution fovea in order to perform a local and detailed scene inspection, the perceived world is decomposed into distinct snapshots taken from different viewing directions. By purposefully guiding the gaze, the presented approach of memory-based active visual search provides an answer to the object search query in terms of transsaccadic memory. This memory constitutes a consistent representation of the scene with respect to the target object.

Different aspects of the overall problem have been addressed including object detection and recognition, calibration and saccade control, memory, and visual attention. The key scientific contributions of this thesis include:

- **Transsaccadic memory**
  The integration of memory and active visual search extends the answer to the problem from a single fixation to the content of a memory. This thesis introduced a transsaccadic memory which supports the accumulation of

visual information during saccadic eye movements. The underlying hierarchical layout was chosen in accordance with the peripheral-foveal principle of the vision system. The preattentive layer accumulates hypotheses about object candidates based on the peripheral images while the object memory layer fuses recognition results from foveal vision. The output of the visual search procedure is provided by the object memory layer which is updated in a closed-loop fashion. Each object entity in object memory stores properties such as the spatial location and the existence probability of an object instance.

- **Solution to the correspondence problem**
  The accumulation of spatial information from different gaze directions requires to solve the correspondence between perceived entities. A solution to the correspondence problem between measurements performed at different eye configurations has been introduced in this thesis which takes into account uncertainties from vision and saccade execution. The approach is based on the Rao-Blackwellized Particle Filter which allows the inference of the spatial memory in a probabilistic fashion. The proposed motion model takes into account the current eye configuration and uncertainties resulting from the offline kinematic calibration. Combining the offline calibration and an online estimation of remaining errors provides a robust estimation for the 3D Cartesian position of object candidates independent of the appearance of the current scene. The 3D localization uncertainty is approximated using a normal distribution which is derived from the 2D localization errors in the left and right image based on the unscented transform. The problem of building correspondences between 3D measurements and preattentive memory entities resulting from past measurements is solved using the maximum-likelihood estimate.

- **Active saliency**
  In order to guide the active camera system, an approach for the calculation of saliency has been proposed which considers transsaccadic memory as integral part of active visual search. The proposed active saliency measure extends the Bayesian Strategy to visual attention in order to include the requirement of a consistent memory. Memory consistency and the probability of detecting an object together constitute the drive for actively fixating elements of the scene. The general formalization of active saliency defines a probabilistic model for the inference of memory consistency from change in the scene and validation performed by fixating elements. Based on this model, an implementation taking into account transsaccadic memory has been derived. The preattentive layer serves as prior for the existence of objects and at the same time reports relevant changes. The object memory layer refines the existence probability with respect to the foveal observations. Each foveal observation leads to a validation of object memory.

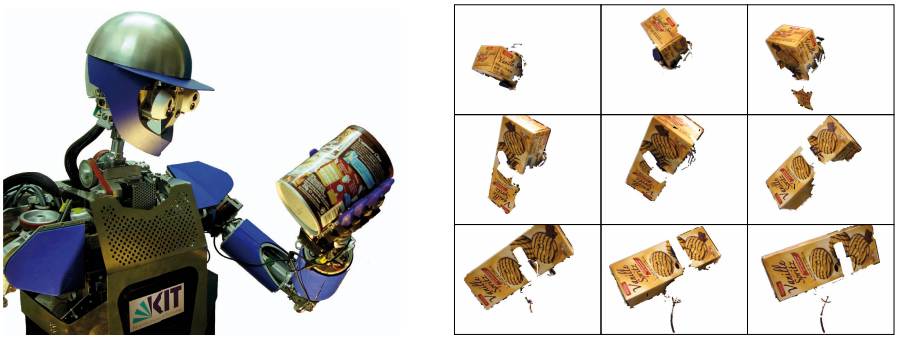

Fig. 10.1. Acquisition of multi view representations of objects held in the hand of ARMAR-III. Figure-ground segmentation and hand-object segmentation is performed in order to extract the views of an unknown object. The views are stored in an aspect graph representation [Welke et al., 2010].

The proposed approach for memory-based active visual search provides the means to extract and retain information about relevant objects in the scene. Thereby, object perception is considered as a continuous process with the responsibility to provide a consistent representation in terms of scene knowledge. As such, the presented mechanisms form an integrated perceptual component which provides this representation to other processes such as task execution and scene interpretation in an autonomous manner.

10.2 Discussion and Outlook

This thesis concentrates on relevant mechanisms in order to allow the execution and evaluation of memory-based active visual search on an anthropomorphic system. Some aspects which were not critical for the design of the core approaches have been left out for future work.

- **Simultaneous search for multiple targets**
  In this thesis, the capability of the system to retain a consistent memory of instances of a single target object was evaluated and discussed. Usually several objects have to be considered in order to provide a memory with task relevant spatial information in the scene. The extension to multiple target search is straight forward using the proposed memory-based active visual search approach.
- **Multi-view object representations**
  The representation of target objects used for the experiments comprises only one view for each object which was trained manually. In order to provide the necessary invariance to object rotations as required for many

real world tasks, such a representation is not sufficient. In order to extend the appearance-based representation scheme for the inclusion of multiple views of an object, the aspect graph representation has been proposed [Welke et al., 2008a]. The aspect graph encodes the set of views using a spherical graph where each node corresponds to one view. The acquisition of such representations on humanoid robots should be accomplished in an autonomous manner thus equipping the system with the ability to incrementally build its world knowledge. An approach for the acquisition of such multi-view representations by actively rotating the object in the hand of the robot has been presented in [Welke et al., 2010] (see Fig. 10.1). Preliminary results show that these representations are feasible for the application in active visual search.

- **Extension to multiple object types**
  The implementation of the memory-based active visual search relies on two different visual modalities in order to detect and recognize objects: color and texture. Robust search for objects is possible only if they exhibit enough texture and offer a significant color signature which is the case for many kitchen objects. In general not all real world objects fall into this class. The extension to the search for more object types can be accomplished by extending the approach using more detectors and descriptors in order to represent the object. The object recognition, currently based solely on texture, can be e.g. extended by the inclusion of edge, color and shape information. This would allow recognizing a wider class of objects. The object detection which is currently based on color signatures alone could be extended by more sophisticated color indexing techniques based on the HMMD color space [Salembier and Sikora, 2002]. The most elaborated way to achieve generality in terms of object recognition and detection consists of learning object representation from simple features. A hierarchical approach for learning object representations is presented in [Fidler et al., 2009] which allows to detect and recognize a wide class of objects. The integration of such a hierarchical approach would require to adapt the memory structure and accompanied processes to the hierarchies inherent in the object representation.

A

EarlyVision Libraries and Tools

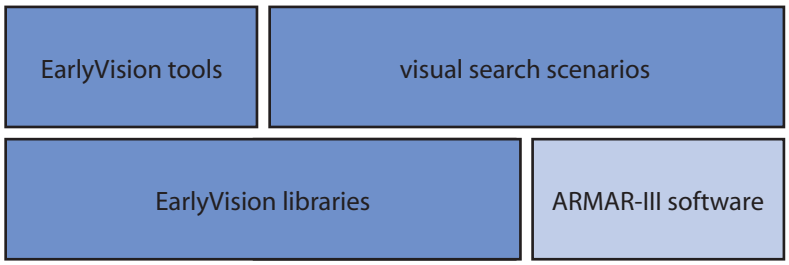

Fig. A.1. The software components developed for this thesis include the EarlyVision set of C++ libraries and a set of tools and applications. Applications that require the real robot are implemented as scenarios based on the ARMAR-III software framework and making used of the EarlyVision libraries.

This appendix gives on overview of tools and software components developed in order to realize the memory-based active visual search approach. The software has been developed in C++ for linux platforms.

An overview of developed components is illustrated in Fig. A.1. The EarlyVision libraries constitute the core of the software comprising all relevant algorithms. The libraries build on the Integrating Vision Toolkit[1] (IVT) which provides fundamental image processing and advanced 2D and 3D object recognition techniques. While parts of the developed software components such as threading and camera capture module have been integrated into the IVT, the EarlyVision libraries provides functionality and data structures which are specific to active visual search. The EarlyVision tools provide examples and applications which make use of the EarlyVision libraries.

[1] `ivt.sourceforge.net`

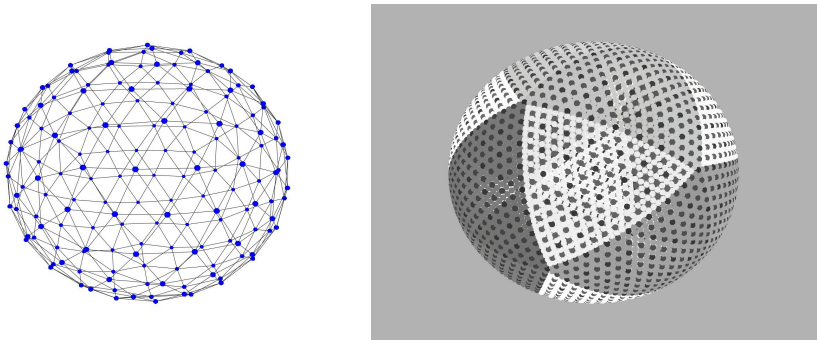

Fig. A.2. Visualization of the spherical graphs. Left: The basic spherical graph implementation offers nodes on the unit sphere and undirected edges. Right: The intensity graph offers additional properties of nodes such as a float valued activation.

The implementation of software components within the ARMAR-III framework is accomplished using so-called scenarios. In addition to the applications provided by the EarlyVision tools, scenarios have been implemented whenever access to the robot hardware or software framework was necessary.

In the following two sections an overview of the functionality and the data structures that have been developed and made available in the EarlyVision libraries is provided.

A.1 The Base Library

The base library `evbase` contains data structures, algorithms, and software components which are used by all other libraries in EarlyVision. In the following, a brief overview of the library content is provided. The organization roughly reflects the file system layout of the library.

- **data structures**
  An implementation of spherical graphs has been realized (see Fig. A.2, left). In order to generate such graphs, geometrical as well as numerical methods for the distribution of nodes have been implemented. Based on the spherical graph, the intensity graph extends the nodes by additional properties (see Fig. A.2, right). For the generation of gaze sequences a winner-take-all network has been implemented. The network allows supports 2D saliency maps and spherical saliency maps based on the intensity graph (see Fig. A.3).

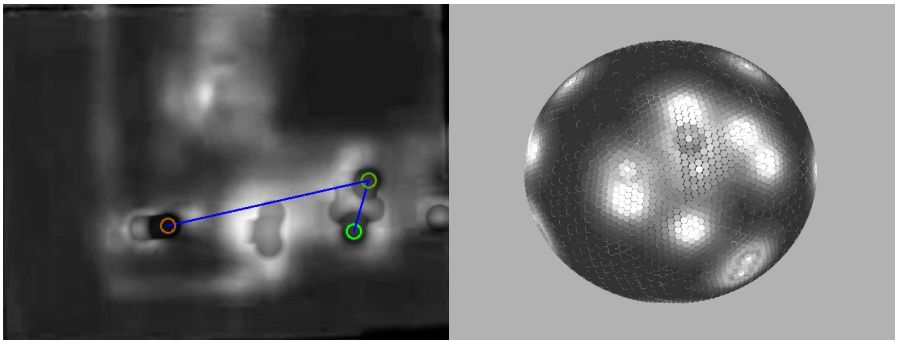

Fig. A.3. The implementation of the WTA network supports 2D saliency maps and spherical saliency representations. Left: Sequence of attended locations on a 2D saliency map. Right: Normally distributed saliency peaks in the spherical WTA implementation.

- **messaging**
  In order to provide a communication layer for multi-threaded applications, a messaging mechanism has been implemented which supports blocking as well as non-blocking communication within a single process.
- **clustering**
  Several clustering approaches have been implemented. The approaches include unsupervised and incremental clustering methods such as growing neural gas, incremental growing neural gas, and BIRCH.
- **features**
  The implementation of the different CCH descriptors developed for this work is realized in the base library. The implementation includes the RFCH, the LCCH, and the NLCCH descriptors. For the extraction and matching of the CCH features, methods for single images and for stereo images are provided.
- **vision**
  A set of basic vision algorithms is implemented including integral images, background subtraction based on Eigenbackgrounds, and particle filter based localization of geometrical models.
- **probabilistic inference**
  The base library provides tools for discrete and continuous probabilistic inference. Modeling step, marginalization, and inference is supported for discrete random variables. Tools for the inference based on anisotropic normal distributions are provided including the unscented transform and Kalman filtering (see Fig. A.4).

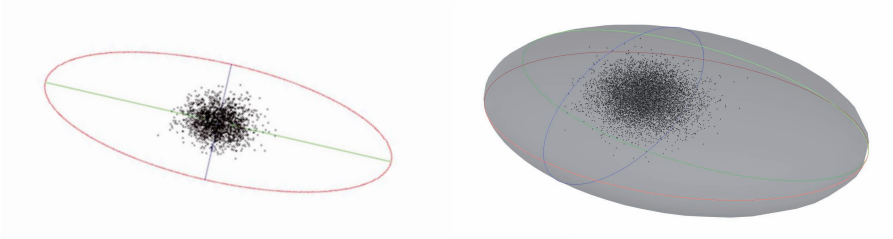

Fig. A.4. The base library provides methods to handle multivariate anisotropic normal distributions. For the 2D and 3D case tools for the visualization are provided.

- **user interface**
  A graphical user interface (GUI) is provided by the base library which builds on the IVT mechanisms. The user interface implementation adds supports for e.g. multi-threaded OpenGL and palettes. For most of the above algorithms and data structures methods for visualization are provided based on the graphical user interface.
- **tools**
  A collection of tools has been implemented including time stamps, logging, and XML configuration file handling.

The dependencies to other libraries are kept as small as possible. Dependencies to QT, OpenGL and libdc1394 are inherited from the IVT. The base library itself depends on libxml2 for XML parsing.

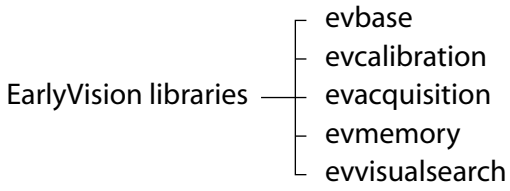

Fig. A.5. Overview of libraries provided by EarlyVision.

A.2 Visual Search Specific Libraries

Based on the `evbase` library a number of libraries containing visual search specific software components have been implemented. The organization of the libraries corresponds to the scope of the provided methods as illustrated in Fig. A.5. The content of the libraries is discussed briefly in the following.

- `evcalibration`
  This library contains the kinematic calibration of the active system as well as all methods to perform saccadic eye movements and stereo triangulation using moving eyes.
- `evacquisition`
  The acquisition library includes all methods required for the autonomous acquisition of multi-view object representations on ARMAR-III.
- `evmemory`
  The memory library implements all memory types required to realize the memory-based active visual search approach.
- `evvisualsearch`
  The visual search library contains all software components which together form the processing chain of the memory-based active visual search.

In the subsequent Appendix B, the application of these libraries in the realization of the proposed approach is discussed in detail.

B

Realization of Active Visual Search

The following sections cover the realization of the memory-based active visual search approach proposed in this thesis. All software components and tools are either part of the EarlyVision library or implemented as scenarios within the ARMAR-III software framework. The processing chain of the memory-based active visual search approach is outlined in Fig. B.1.

The sensory data including camera images and joint angle readings acquired from the Karlsruhe Humanoid Head is stored in the sensory buffer. Based on the peripheral images and the current head pose as stored in the sensory buffer, object candidates are extracted by the mapping component. The verification component performs object recognition based on the foveal images. The results of object detection and recognition are stored in the respective layer of transsaccadic memory. Making use of the information stored in the memory, the attention component determines the target and the time of a saccade. Once the saccade has been initiated, the head control calculates the target joint angles which correspond to the desired focus of attention.

In the following Section B.1, the realization of memories is discussed in more detail. Subsequently, in Section B.1 the software components are subdivided and described in more detail.

B.1 Memories

All memories are implemented within the EarlyVision library `evmemory`. The library also provides tools for the inspection and visualization of the memory content. Additionally, several ARMAR-III scenarios have been implemented in order to realize tools that require the hardware of the robot.

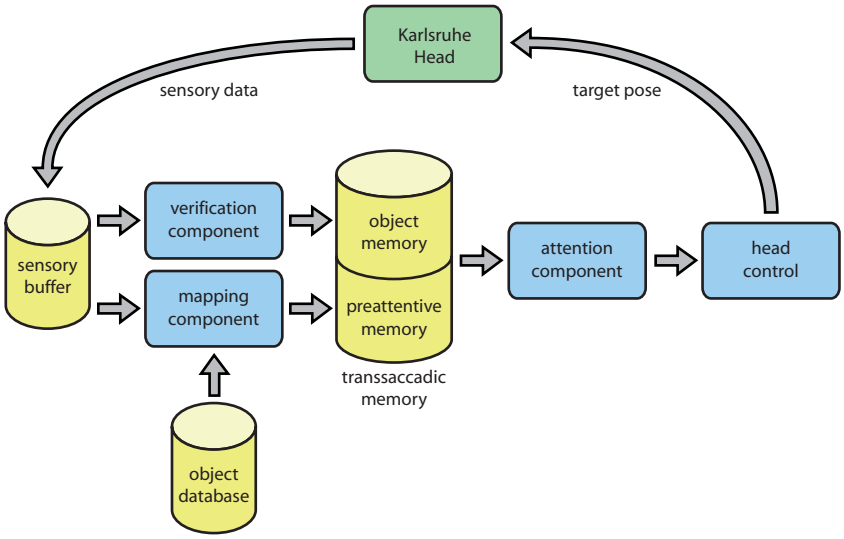

Fig. B.1. Software components and memories developed for the realization of the memory-based active visual search approach. From the sensory data provided by the Karlsruhe Humanoid Head, the proposed approach determines the target poses for saccades during the course of an active visual search task. The realization comprises three different types of memory: the sensory buffer, the object database, and the transsaccadic memory. The mapping and verification components update the content of the preattentive and object memory layer of transsaccadic memory based on the sensory buffer content and a representation of the target object stored in the object database. Based on the transsaccadic memory content the attention module generates sequences of saccades which are executed by the head control component.

B.1.1 Sensory Buffer

The sensory buffer stores the sensory data acquired from the Karlsruhe Humanoid Head. The content of the sensory buffer is made available for the mapping and verification components including:

- camera images
  The images captured from peripheral and foveal stereo camera pair are stored in the sensory buffer. In order to capture the images synchronously, the capture module relies on the IEEE 1394 synchronization mechanisms. An example of the camera images is depicted in Fig. B.2.
- joint angles
  The joint angles of the head joints are stored in the sensory buffer in order to allow transformation between the moving camera coordinate frames and a fixed base coordinate system (see Fig. B.3).

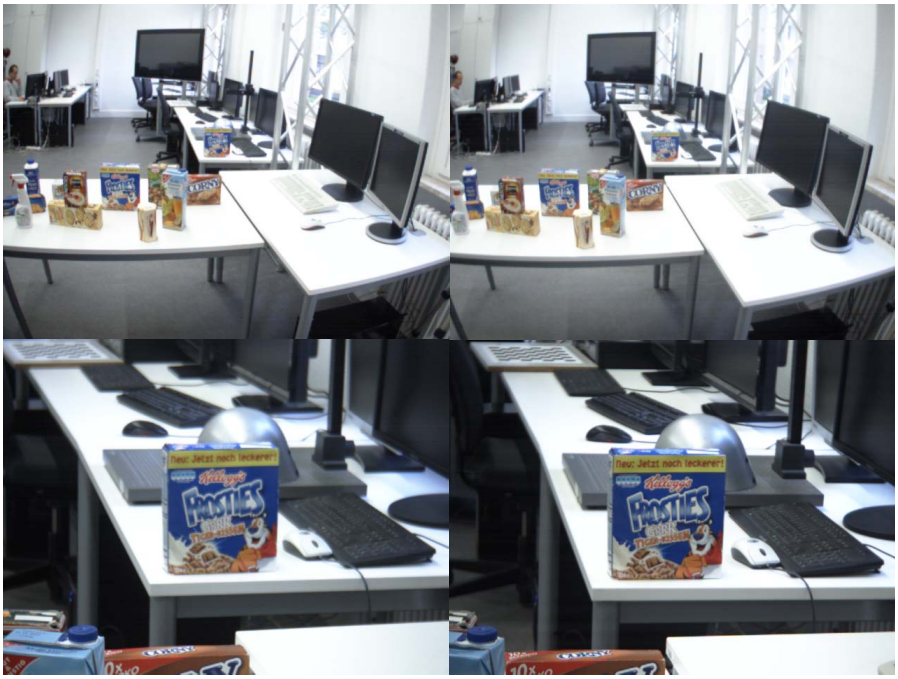

Fig. B.2. The images captured by the peripheral and foveal camera pair are stored in the sensory buffer. Top: The peripheral images provide a wide view of the scene. Bottom: the foveal images allow a detailed inspection of scene fragments.

- time stamp
  A time stamp is calculated for the current sensory data based on the time stamp clock register mechanism of the operating system.

The memory content is refreshed each time new camera images are available. On retrieval of the images, the current head joint readings are stored and the time stamp is updated. The sensory buffer does not include past sensory data. Each memory refresh results in discarding its previous content.

In order to allow access to the sensory buffer from the camera capturer, the mapping, and verification component thread safe access and cloning of its content is implemented.

B.1.2 Object Database

The knowledge about the appearance of target objects is kept within the object database. As already discussed in Chapter 5, the proposed approach makes use of an appearance-based object representation scheme. The major drawback of the application of appearance-based representations consists

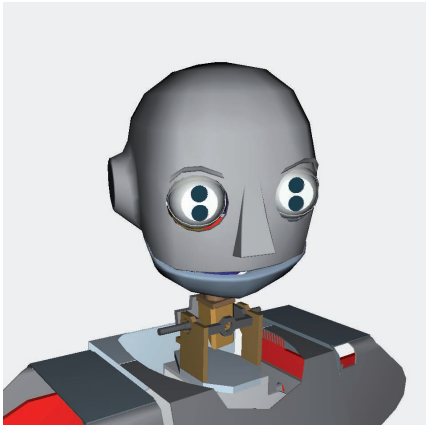

Fig. B.3. The joint angle readings of the ARMAR-III head-eye system are stored in the sensory buffer. The stored joint angles are synchronized with the input images in order to assure consistency of the sensory buffer.

in their instability over the viewing sphere of an object. Depending on the invariance toward rotation of the applied feature extraction methods, an appearance-based descriptor is only valid within a local neighborhood of the trained viewing direction. In order to recover the ability to search and recognize the target object from all possible viewing directions, the object database implements techniques that allow to combine a set of object views into an appearance-based multi-view object representation.

Figure B.4 illustrates the acquisition of multi view object representations and their storage in the object database. The views of an object collected during the acquisition process are stored in the aspect pool. In order to retain the spatial relations between collected views, each view is associated with a viewing direction. The collection of viewing directions is stored in the aspect graph representation as illustrated in Fig. B.5. The aspect graph is a spherical graph where each node is associated with a single view of the object as stored in the aspect pool. The Delaunay triangulation of the spherical graph encodes the neighbor relation of views. For each target object, a separate aspect graph is built. In order to detect and recognize the target object, the views stored in the aspect pool are processed by feature extraction methods. The current implementation of the memory-based active visual search calculates the CCH features as introduced in Section 5.1 and a set of SIFT features per view as described in detail in Section 5.2. The resulting features are associated to the corresponding nodes in the aspect graph and stored in the feature pool. In order to retrieve a compact representation of the target objects, grouping based on similarities in feature space is performed as described in [Welke et al., 2007].

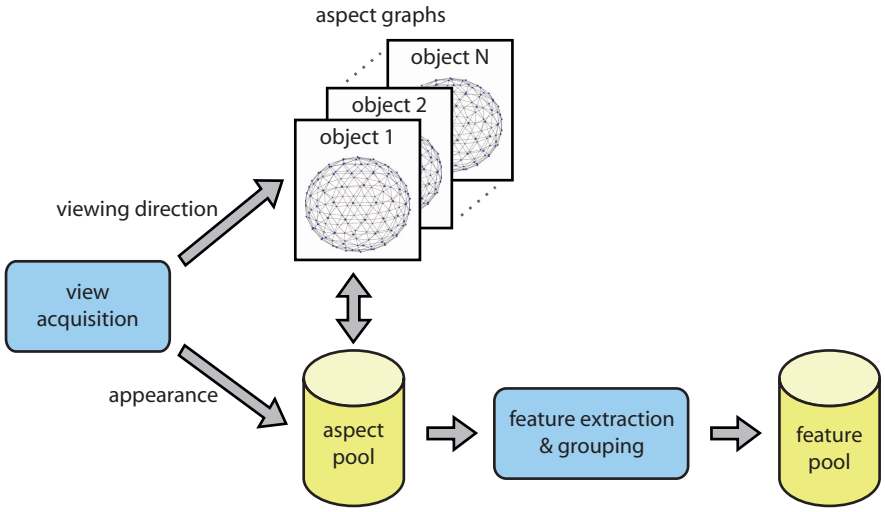

Fig. B.4. Process chain for the acquisition of multi view object representations. The view acquisition process generates a set of views of the objects together with the viewing direction relative to a fixed object coordinate system. The resulting views are stored in the aspect pool, while the set of viewing directions is stored in an aspect graph representation. The views stored in the aspect pool are processed by the feature extraction methods and grouping based on similarities in feature space is performed resulting in a compact representation which is stored in the feature pool.

In order to acquire the necessary views accompanied with the corresponding viewing directions three different view acquisition processes have been implemented. The processes and developed tools are briefly discussed in the following.

Object representations covering the complete viewing sphere can be generated offline using the Interactive Object Modeling System available at the Karlsruhe Institute of Technology [Becher et al., 2006]. The views are acquired by rotating the object on a turntable and taking images from a camera mounted on a robotic arm with one DoF. From the encoder readings of the arm and the turntable, the direction of each captured view is available and can be exploited to generate the aspect graph representation. Due to joint angle limitations of the arm one acquisition process does not cover the complete viewing sphere. The complete sphere as illustrated in Fig. B.5, left is composed by registering the results from two scans, where the second scan is performed while the object being flipped. The EarlyVision tools provide applications for the generation of aspect graphs from data of the modeling center as well as tools for the registration of aspect graphs. For the generation of the complete object database content including feature graphs, aspect and object pools a command line tool has been developed.

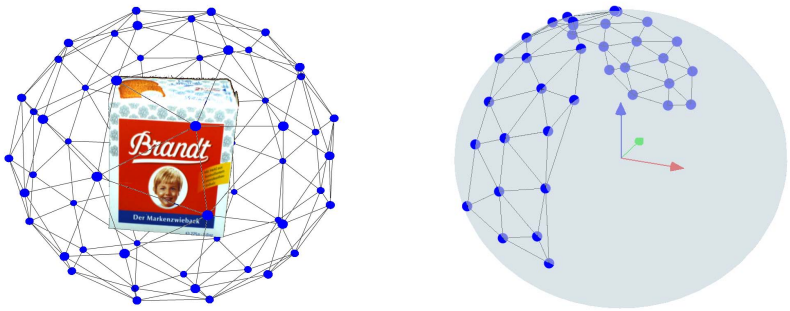

Fig. B.5. Each node of the aspect graph corresponds to a viewing direction of the object. Delaunay triangulation is performed in order to express the neighbor relation of views. Left: Aspect graph acquired by an off line acquisition process. Right: Partial aspect graph acquired autonomously on the humanoid robot.

The acquisition of representations of objects on a humanoid robot has been proposed in [Welke et al., 2010] and implemented on the robot ARMAR-IIIb. In order to avoid the offline acquisition process and to equip the humanoid system with the ability to acquire object representations in an autonomous manner, multi view representations are acquired from an object held in the hand of the robot. By exploiting the redundancy of the anthropomorphic arm different viewing directions of the object are generated. The resulting camera images are subject to figure-ground segmentation based on the Eigenbackground approach. Furthermore, the configuration and pose of the five-fingered robot hand is estimated visually. For this purpose proprioceptive sensory data is exploited in order to restrict the search space in the image plane. The estimation of configuration and pose of the hand allows to derive a viewing direction and to segment the robot hand as well as the object. The results of background subtraction, hand pose estimation, and additionally of disparity calculation are fused in an occupancy grid in order to yield the final segmentation of the object. Using the viewing direction and the segmented object view, the target object representation is established. The complete acquisition process is realized as a scenario within the robot software. The graphical user interface as depicted in Fig. B.6 allows to observe and control the acquisition process.

In order to facilitate the training of new objects, a scenario has been implemented which allows to capture a single view from an object currently visible to the camera and to generate an object database representation based on this view. The scenario further allows to test all relevant methods for object detection and recognition and as such constitutes the most relevant tool for development, testing, and evaluation. The object representations used for the experiments in this thesis have been acquired using this tool. The user inter-

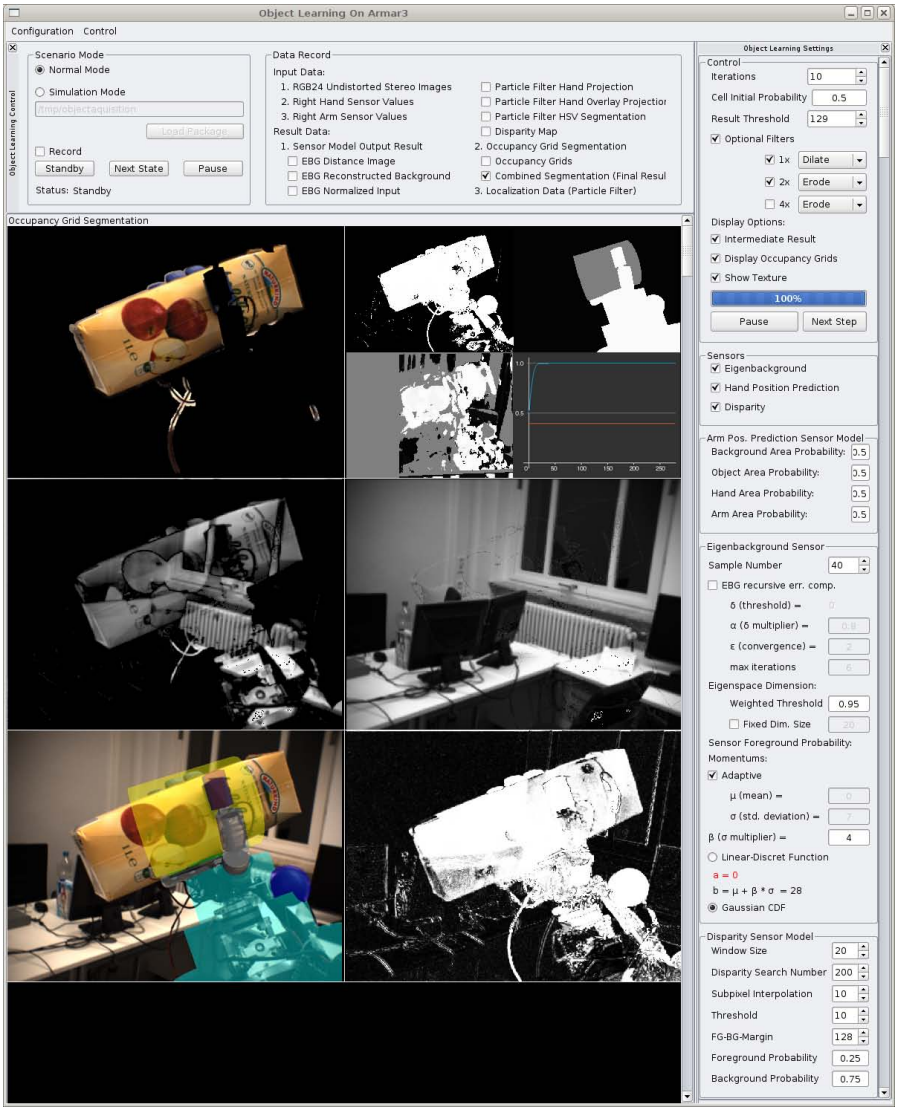

Fig. B.6. The autonomous acquisition of multi view object representations is implemented as a scenario for ARMAR-IIIb. The scenario implements the generation of representations by rotating the object. During rotation, figure-ground and hand-object segmentation is performed in order to produce segmented views.

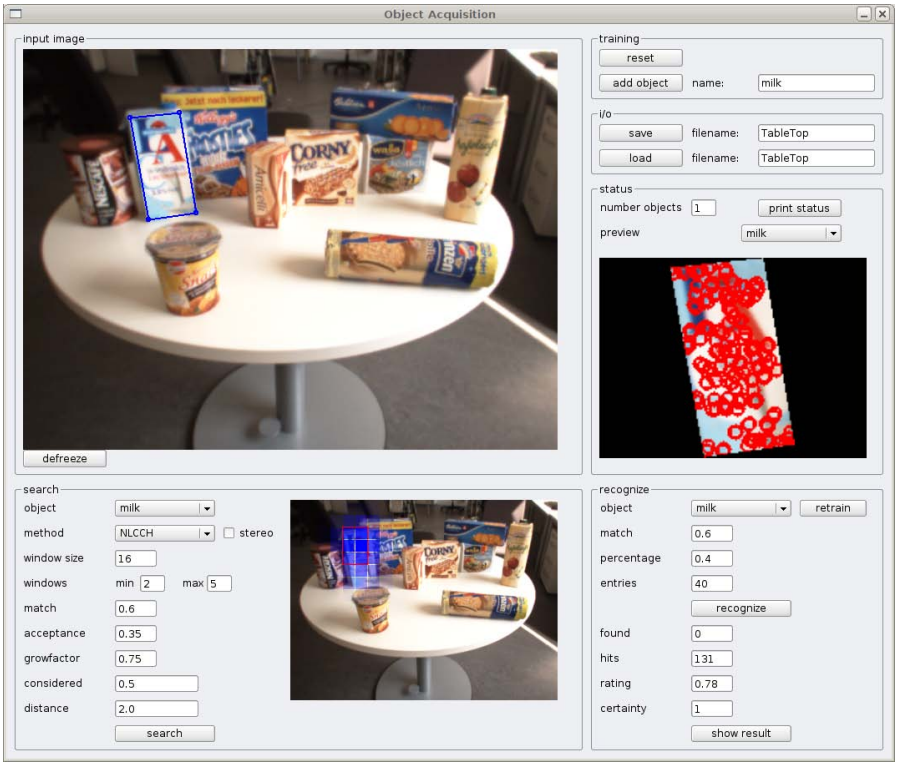

Fig. B.7. A scenario is provided which allows to acquire representations of target objects based on a single view. The user interface offers possibilities for testing the acquired representation with the relevant visual search components.

face of the tool is depicted in Fig. B.7. The silhouette of an unknown object in the current input image can be selected by the user. A one button procedure allows to train its appearance. The interface allows the inspection of all objects stored in the object database and their associated features. Based on the current view, the object candidate detection and the foveal recognition can be tested. Thereby, all relevant parameters can be adjusted by the user.

B.1.3 Transsaccadic Memory

The transsaccadic memory holds information about objects and object candidates as collected during the visual search process. In Chapter 7 a detailed discussion of involved processes and the organization of transsaccadic memory has been provided. In the following, the focus is put on aspects of the implementation and tools developed for the inspection of the memory content.

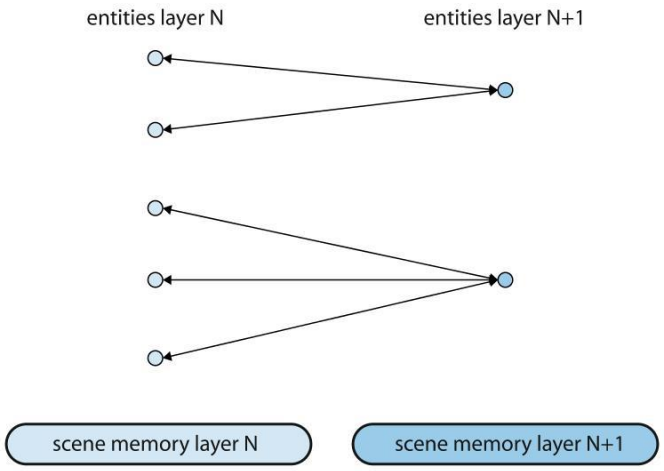

Fig. B.8. A hierarchical egocentric scene memory provides the basis for the realization of the transsaccadic memory. The memory is composed of multiple layer, each layer stores the associated entities. Entities from successive layers are linked in a bidirectional manner in order to support propagation of properties. The link is realized as N to 1 mapping where one entity of the higher level can be linked to multiple entities of the lower level. All scene memory entities share a common set of properties.

The implementation of the hierarchically organized transsaccadic memory follows the principle illustrated in Fig. B.8. For this purpose, a generic hierarchical egocentric scene memory has been implemented. The memory consists of an arbitrary number of layers where each layer contains associated memory entities. The entities of each layer are associated to the successive layer. The link is implemented as N to 1 connection, where one entity in the higher layer can be associated to a number of entities in the preceding layer. Each memory layer contains a list of entities in order to allow access to entities based on an identifier. The link mechanism allows to propagate properties from lower level entities to higher level entities which is important e.g. in the propagation of updates without the necessity to search each layer excessively.

A basic set of properties has been defined for entities from all layers. These properties are inherited by the transsaccadic memory layers. The properties common to all scene memory entities include:

- unique identifier
  Each entity has an identifier which is unique in the corresponding scene memory layer. Together with the identifier of the corresponding layer, the entity can be uniquely identified in the memory.

- associated object
  A scene memory entity is always associated to a target object which has been subject to active visual search. The association links the object representation from the object database with the scene memory entity.
- existence probability
  The probability of existence reflects the amount of accordance between the perceived scene and the object representation.
- validity
  The validity of the entry indicates whether it represents the target object or not. The validity is derived based on the existence certainty.
- position
  Each entity is associated with a spatial location represented in a common egocentric coordinate frame.
- time stamps
  Each entity holds different time stamps in order to assure consistency in asynchronous memory updates. The time stamps include the time of creation of the entity and the time of the last performed update.

The `evmemory` component of the EarlyVision library provides a general implementation of the scene memory layer and the scene memory entity. For the realization of the transsaccadic memory, the implemented mechanisms where used in order to realize a hierarchical memory based on two layers: the preattentive memory layer and the object memory layer on a higher level. As already discussed in Chapter 7 this organization is derived from the peripheral-foveal design of the active camera system.

The preattentive memory holds information extracted by the peripheral object candidate detection approach as discussed in Section 5.1. The unique identifier is derived by incrementing a counter for each detected object candidate in the scene. The associated object directly results from the search task. The existence certainty and the validity of the entry result from the matching between object representation and scene fragments based on the developed CCH descriptor. In order to reflect the uncertainties in the detection process and to allow to identify correspondences between observed and stored object candidates, the structure of the memory is extended by the position uncertainty in terms of an anisotropic normally distributed probability distribution. In order to inspect and visualize the content of the preattentive memory layer, an OpenGL based viewer has been developed and is part of the EarlyVision library. As depicted in Fig. B.9, the visualization disclose the properties of the memory entities such as unique identifier, existence certainty, validity, and position. Optionally, the associated uncertainty can be visualized. Each preattentive memory entity is linked to exactly one object memory entity following the principle illustrated in Fig. B.8.

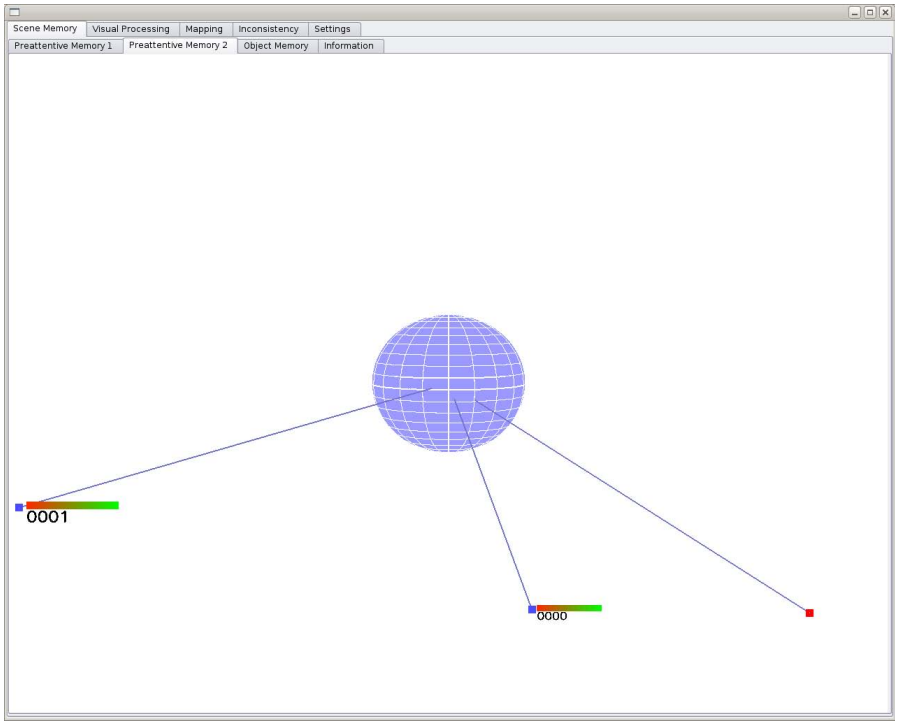

Fig. B.9. The visualization of the preattentive memory allows to inspect important properties of the preattentive memory entities. The spatial location, a unique identifier, and the probability of existence are displayed. The color of a node indicates the validity of the object candidate.

The object memory layer holds information which results from the verification of object candidates using peripheral object recognition as discussed in Section 5.2. Each object memory entity is associated to one or a number of preattentive memory entities. Most properties of object memory entities are inherited from the associated preattentive memory entities. The inherited properties include the identifier, the associated object, and the time stamps. Further, on creation of the object memory entity the position and activation are set to the corresponding properties of the linked preattentive memory entity. The position, existence certainty, and validity properties are updated based on the result of the peripheral object recognition result. As for the preattentive memory layer, an OpenGL based visualization of the memory content is provided by the EarlyVision library. The viewer allows the output of properties of object memory entities including its position and identity as illustrated in Fig. B.10.

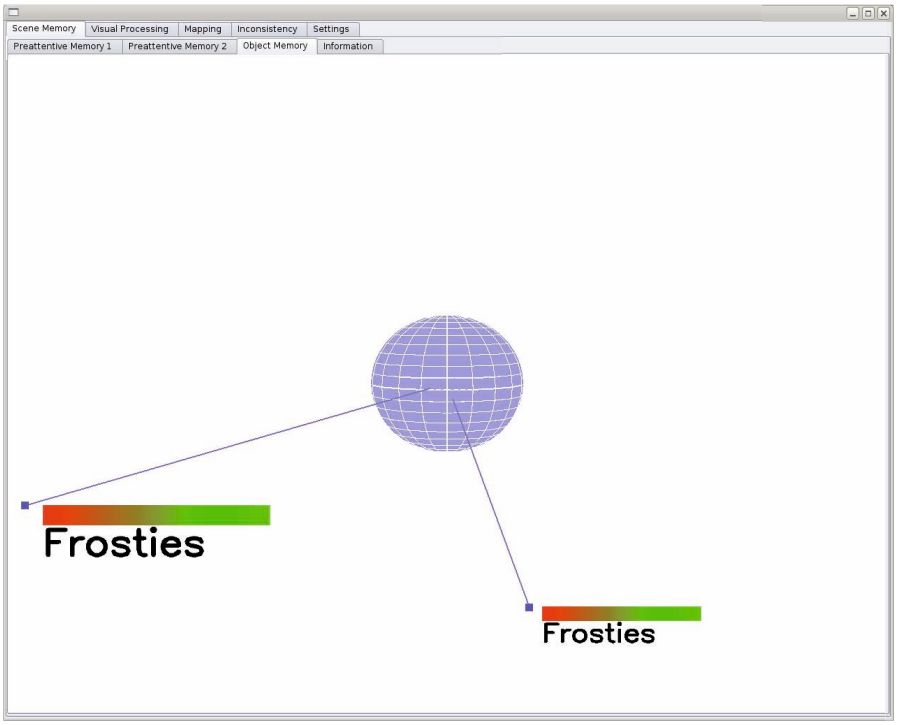

Fig. B.10. The visualization of the object memory illustrates the current interpretation of the scene by the system. Each entity of the object memory layer is displayed accompanied with the probability of existence and the object name.

B.2 Visual Search Components

The developed memory-based active visual search software is decomposed in four components as illustrated in Fig. B.1. Each component is implemented as separate thread running at its own cycle time. In order to communicate among the components, the messaging protocol of the `evbase` library is used (see Section A.1). All communication, blocking as well as non-blocking is realized by the messaging system. A central instance allows to inspect the flow of messages and determine the current state of the system.

All components introduced in the following are implemented in the `evvisualsearch` EarlyVision library. Additional functionality and tools are provided in separate libraries, as scenarios, or as part of the EarlyVision tools.

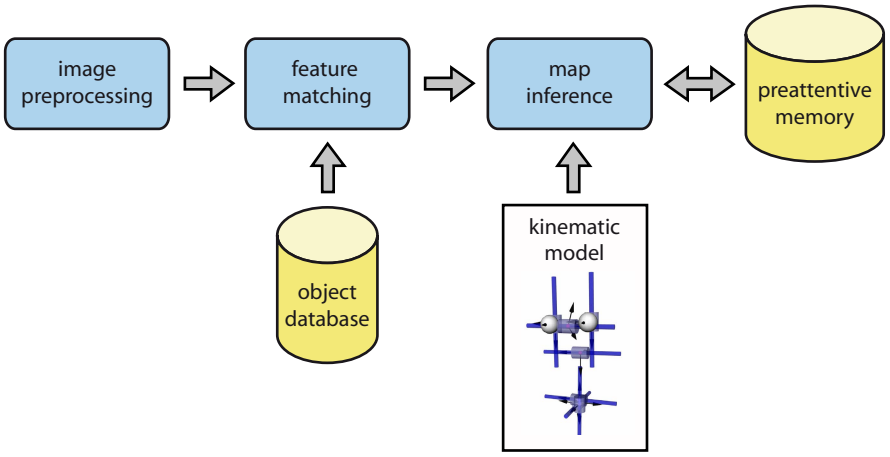

Fig. B.11. Structure of the mapping component. The mapping component updates the content of the preattentive memory layer based on the peripheral views. The processing chain includes the steps of preprocessing, feature matching, and map inference.

B.2.1 Mapping Component

The mapping component encapsulates the complete chain from peripheral image preprocessing to the update of the preattentive memory layer. The goal is to detect candidates of the target object in the peripheral views and to track these candidates over multiple saccadic eye movements in order to solve the correspondence problem. In order to achieve this goal, the processing steps illustrated in Fig. B.11 have been realized. In the first step, the peripheral stereo views are preprocessed. Using the target object representation stored in the object database, the feature matching process generates hypotheses about object locations in the left and right peripheral image. The preattentive memory layer which holds previously observed object candidates is updated using the observed object candidates based on the current pose of the head. Thereby, uncertainties during perception and execution of saccades are considered.

In the image processing step, the peripheral views are undistorted using the calibrated intrinsic camera parameters. The intrinsic camera calibration is performed using the methods provided by the IVT. Optionally, a simple lighting model covering constant shifts in illumination over the complete view is applied to the images. This step is only necessary if the target objects have been acquired off-line in order to compensate for different lighting conditions.

The feature matching identifies regions of the peripheral images that likely contain the target object using the CCH descriptor as discussed in Section 5.1.

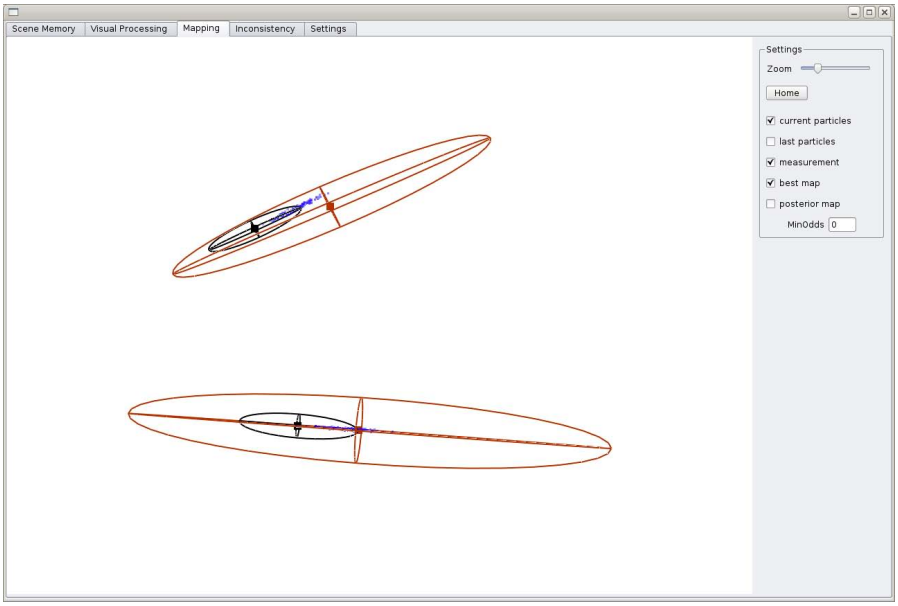

Fig. B.12. The EarlyVision library provides tools to inspect the procedure of object candidate mapping. An OpenGL based viewer allows to visualize the current object candidates, the inferred map, and the particle distribution.

If the NLCCH descriptor is used, the representation of the target object is already required during the extraction of the features, since the quantization itself is part of the representation. The result of the feature matching step consists of 2D regions in the left and right peripheral image which correspond to object candidates.

In the map inference step the transsaccadic memory content is updated based on the detected object candidates. As discussed in detail in Section 7.2, the update is performed in a probabilistic manner based on the Rao-Blackwellized particle filter. Thereby, the 2D object regions from the feature matching step are used to generate an uncertain estimate of the 2D location. Successively, 3D observations are derived from the uncertain 2D locations in the left and right image by taking into account inaccuracies within the calibrated kinematic model and in the execution of saccades. The observed 3D object candidates are fused with the content of preattentive memory based on a maximum a-posteriori approach. Based on the maximum a posteriori (MAP) estimate, the correspondence problem is solved which results either in the generation of a new entity in preattentive memory or in the update of an already stored entity.

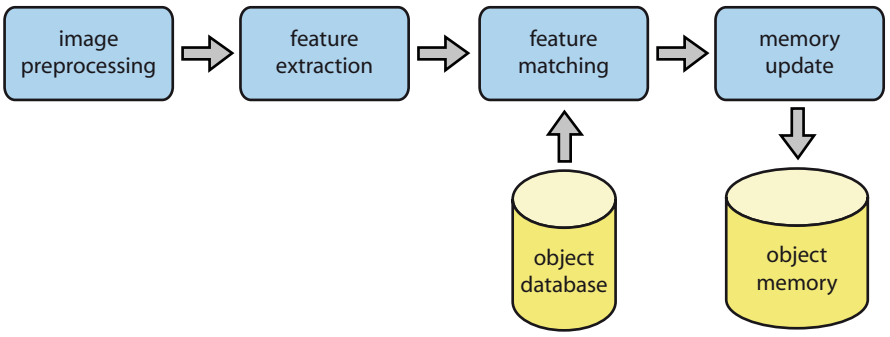

Fig. B.13. Structure of the verification component. The verification component comprises all steps from foveal object recognition to the update of the object memory layer. The processing steps include image preprocessing, feature extraction, feature matching, and memory update.

The mapping component is realized as a single independent thread. The thread runs at the maximum cycle time which depends on several parameters such as the number of object candidates in the scene (see Section 9.4).

The EarlyVision library provides methods for the visualization of the mapping procedure as depicted in Fig. B.12. The OpenGL based viewer allows to inspect the spatial layout of the extracted object candidates. The particle distribution, the performed observation, the best rated and the posterior map together with the accompanied uncertainties can be inspected in order to verify the underlying processes.

B.2.2 Verification Component

The verification component comprises all processes for updating the object memory content based on the foveal object recognition method as illustrated in Fig. B.13. Similar to the mapping component, the input images are initially preprocessed. In the subsequent step, the SIFT feature extraction is performed. The resulting set of SIFT features is subject to comparison with the target object representation in the feature matching step. Based on the recognition result the content of the object memory layer is updated.

The image preprocessing step performs undistortion of the input images. The undistorted images are converted into grayscale in preparation of the feature extraction step.

In the feature extraction step, SIFT features are extracted from both preprocessed images as described in Section 5.2. For each detected corner point, one or a set of SIFT descriptors are generated which are stored with their

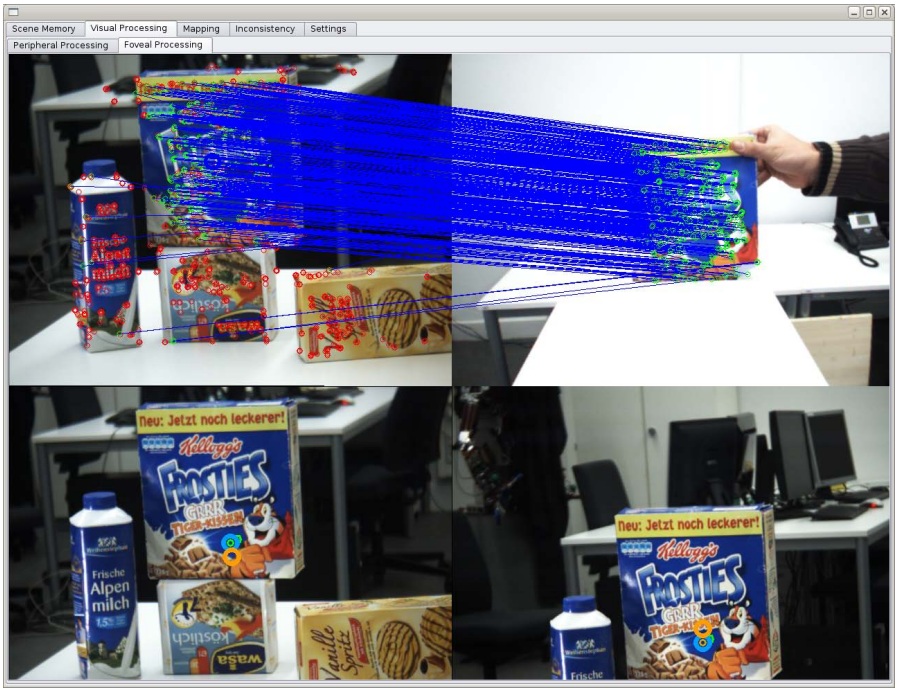

Fig. B.14. The EarlyVision library provides tools to inspect the procedure of object candidate verification. The above example of the user interface shows the foveal views of the stereo cameras and the matching of target object representation and current view.

associated orientation and location. For the extraction of SIFT features, the implementation of the IVT is used.

The extracted SIFT descriptors are subject to comparison with the target object representation in the feature matching step. Thereby, only features are considered that are associated to the view in the aspect graph corresponding to the entry in the object memory. The current view of the object results from object candidate detection and is stored in the transsaccadic memory. The correspondences resulting from the feature comparison form the input for the Hough Space verification. Using the stored orientation and location of the features, each correspondence votes in the Hough Space vote matrix. Based on the entry with maximum votes, a quality measure for the match is determined and the 2D position is derived from all corresponding features.

Having performed the feature matching, the content of the object memory layer is updated based on the matching result. Therefore, a normalized value for the quality of the match is derived which is mapped to a sigmoidal function. Using the resulting match certainty, the existence probability of the

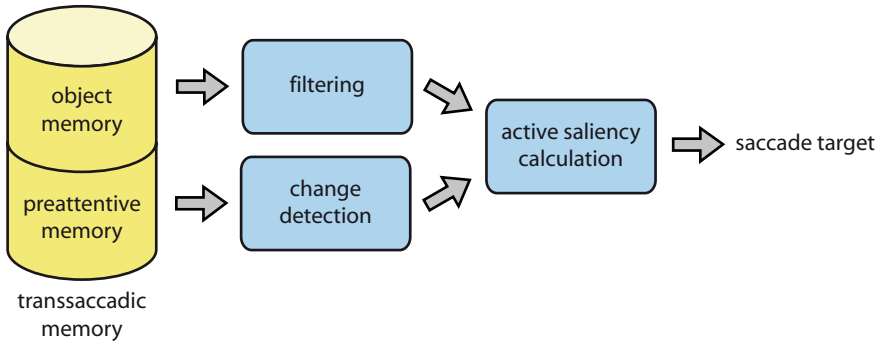

Fig. B.15. Structure of the attention component. Using the proposed active saliency measure, saccade targets are calculated based on the transsaccadic memory content. From the preattentive memory layer, detected change is reported to the active saliency calculation. Further, the probability of existence and location is derived from the object memory layer. Entities that do not correspond to the current search target are filtered.

corresponding entity in the object memory is updated using Bayes filtering. The location of the object memory entity is update in a closed-loop manner based on the 2D location in the left and right image.

The verification component is realized as a single independent thread. The implementation offers two different ways how verification can be executed. By default, the verification is invoked once a saccade to a new target position has finished, resulting in an update of the object memory layer. In order to allow tracking of a recognized target object, the verification can optionally be executed continuously after the fixation of the target object.

The EarlyVision library provides a set of tools to visualize and inspect the processing steps of the verification component. In Fig. B.14, the default visualization is depicted. The visualization provides the foveal views of the camera system and the extracted SIFT features accompanied with correspondences between target object representation and scene.

B.2.3 Attention Component

The attention component generates saccades from the content of the transsaccadic memory. As illustrated in Fig. B.15 the calculation of the active saliency detailed in Section 8.3 is based on the detection of change in the preattentive memory layer and the spatial information from object memory which is filtered for the currently searched object. The output of the attention component is the 3D location of the saccade target.

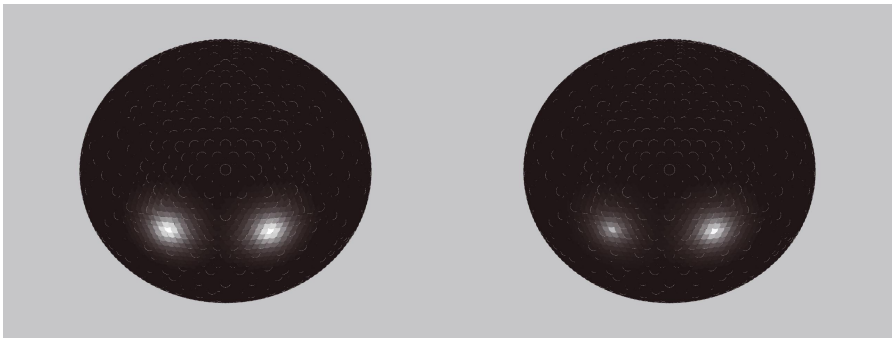

Fig. B.16. For the purpose of visualization, the saliency distribution is projected to a spherical intensity graph. Left: Probability of existence for two object memory entities. Right: Calculated active saliency after considering the consistency of the transsaccadic memory.

The change detection is performed based on the solution to the correspondence problem of preattentive memory. Change to object candidates in the preattentive memory layer is reported if a new entity is generated or if an entity within preattentive memory is subject to removal. Together these two events constitute the information necessary for the change sensor as defined by the active saliency model. The location and probability of existence of the objects is derived from the refined position stored in the object memory entities. The entities are filtered in order to consider only object memory entities which correspond to the current search task. Further, each update of the object memory results in a validation of the object memory content which forms the basis for the observed validation in the probabilistic model.

The active saliency is calculated based on the measurement of change, the location, and the probability of existence for each object memory entity. The active saliency calculation step updates the active saliency measure stored with each object entity based on the proposed probabilistic model. If the active saliency exceeds the attention threshold, a new saccade is initiated resulting in the verification of the corresponding object memory entity.

The attention component is implemented as a separate thread and runs at a fixed cycle time of 1 ms leading to about constant intervals in the inference of the active saliency. The EarlyVision library provides methods which allow to inspect and visualize the saliency distribution in space. As depicted in Fig. B.16, a 2D mapping to a spherical intensity graph forms the basis for the visualization. The object match and the active saliency are visualized for each object memory entity together with the position uncertainty stored within the associated preattentive memory entity.

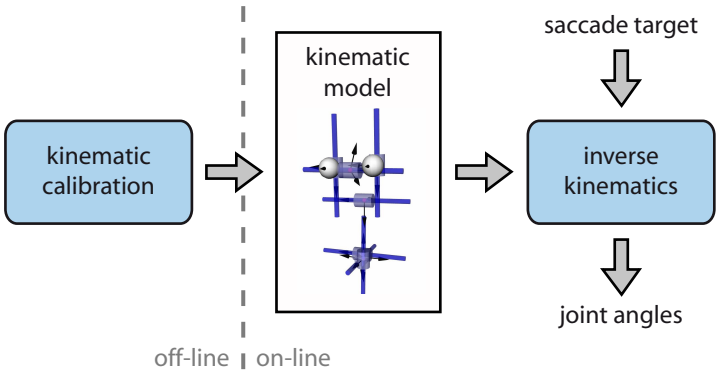

Fig. B.17. The head control component determines joint targets based on the saccade target generated by the attention component. This inverse kinematics problem is solved based on the kinematic model of the head-eye system. In order to estimate the unkown transformations in the kinematic model, an off line kinematic calibration procedure is proposed.

B.2.4 Head Control

The head control component derives the target joint angles which are used to actuate the motors of the head-eye system from the saccade target generated by the attention component as illustrated in Fig. B.17. This problem is usually referred to as inverse kinematics problem. In order to achieve an accurate solution to the inverse kinematics problem, an accurate kinematic model of the head has to be available. This model is determined using an off line visual aided kinematic calibration.

The proposed approach for calibrating the head-eye system is discussed in detail in Section 6.1. Based on the observation of a calibration pattern while moving the eyes, the pose information of the camera system is collected. Based on this data, a non-linear optimization is performed in order to estimate the coordinate transformations that define the kinematic model. Using the calibrated model, the inverse kinematics problem is solved using differential kinematics.

The kinematic calibration approach is part of the EarlyVision `evcalibration` library. This library comprises the optimization procedure and the kinematic model of the head. In order to assist in the calibration of the system, a scenario is provided as depicted in Fig. B.18. The scenario allows to collect the data required to calibrate the eye and the head DoF. Further, the foveal and peripheral extrinsic camera parameters can be estimated. Optionally, a base coordinate system calibration to a specified coordinate frame can be performed.

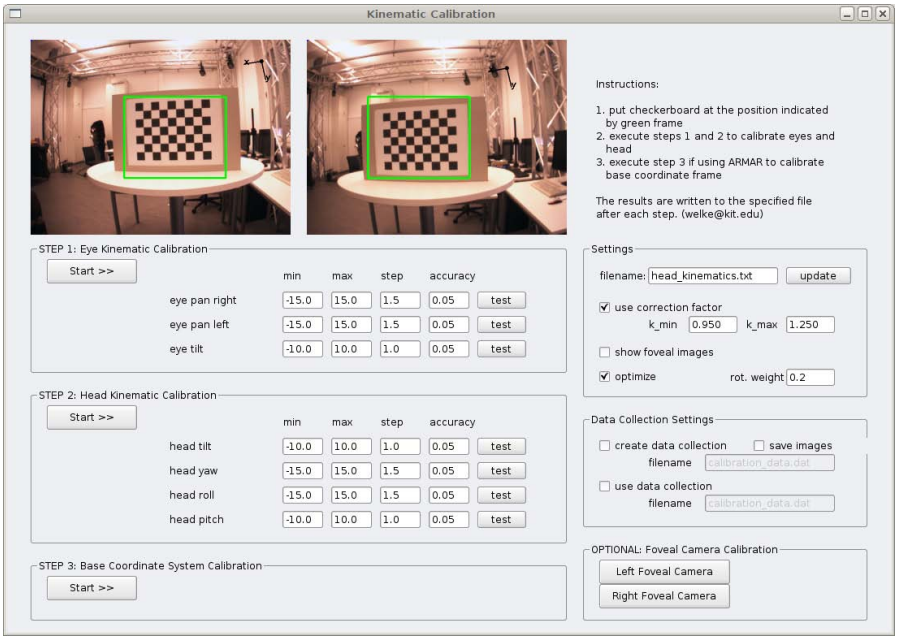

Fig. B.18. The kinematic calibration is implemented as ARMAR-III scenario. It allows to estimate all transformations of the head-eye system. Further, the extrinsic camera parameters of peripheral and foveal cameras can be estimated.

List of Figures

References

Aloimonos, 1993. Aloimonos, Y. (1993). Active Vision Revisited. In *Active Perception*, pages 1–18.

Aloimonos et al., 1988. Aloimonos, Y., Weiss, I., and Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision*, 1(4):333–356.

Amit, 2002. Amit, Y. (2002). *2d Object Detection and Recognition: Models, Algorithms, and Networks*. MIT Press, Cambridge, MA, USA.

Andreopoulos et al., 2011. Andreopoulos, A., Hasler, S., Wersing, H., Janssen, H., Tsotsos, J.K., and Korner, E. (2011). Active 3d object localization using a humanoid robot. *IEEE Transactions on Robotics*, 27(1):47–64.

Aryananda, 2006. Aryananda, L. (2006). Attending to learn and learning to attend for a social robot. In *IEEE International Conference on Humanoid Robots (Humanoids)*, pages 618–623.

Aschwanden and Guggenbühl, 1992. Aschwanden, P. and Guggenbühl, W. (1992). *Experimental Results from a Comparative Study on Correlation-type Registration Algorithms*, pages 268–282. Wichmann.

Asfour et al., 2006. Asfour, T., Regenstein, K., Azad, P., Schröder, J., Vahrenkamp, N., and Dillmann, R. (2006). ARMAR-III: An integrated humanoid platform for sensory-motor control. In *IEEE International Conference on Humanoid Robots (Humanoids)*, pages 169–175.

Asfour et al., 2008. Asfour, T., Welke, K., Azad, P., Ude, A., and Dillmann, R. (2008). The Karlsruhe Humanoid Head. In *IEEE International Conference on Humanoid Robots (Humanoids)*, pages 447–453.

Atkeson et al., 2000. Atkeson, C.G., Hale, J.G., Pollick, F., Riley, M., Kotosaka, S., Schaal, S., Shibata, T., Tevatia, G., Ude, A., Vijayakumar, S., Kawato, E., and Kawato, M. (2000). Using humanoid robots to study human behavior. *IEEE Intelligent Systems and their Applications*, 15(4):46–56.

Atkinson and Shiffrin, 1968. Atkinson, R.C. and Shiffrin, R.M. (1968). *The Psychology of learning and motivation: Advances in research and theory (Vol. 2)*, chapter Human memory: A proposed system and its control processes. New York: Academic Press.

Awh et al., 2006. Awh, E., Armstrong, K.M., and Moore, T. (2006). Visual and oculomotor selection: links, causes and implications for spatial attention. *Trends in Cognitive Sciences*, 10(3):124–130.

Azad, 2008. Azad, P. (2008). *Visual Perception for Manipulation and Imitation in Humanoid Robots*. PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany.

Azad et al., 2009. Azad, P., Asfour, T., and Dillmann, R. (2009). Combining harris interest points and the sift descriptor for fast scale-invariant object recognition. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 4275–4280.

Backer et al., 2001. Backer, G., Mertsching, B., and Bollmann, M. (2001). Data- and model-driven gaze control for an active-vision system. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(12):1415–1429.

Bajcsy, 1988. Bajcsy, R. (1988). Active perception. In *Proceedings of the IEEE*, pages 996–1005.

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

Ballard, 1991. Ballard, D.H. (1991). Animate vision. *Artificial Intelligence*, 48(1):1–27.

Bay et al., 2006. Bay, H., Tuytelaars, T., and Van Gool, T. (2006). Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417.

Becher et al., 2006. Becher, R., Steinhaus, P., Zöllner, R., and Dillmann, R. (2006). Design and implementation of an interactive object modelling system. In *Proceedings of ISR 2006 and Robotik 2006*, pages 22–27.

Becker, 1989. Becker, W. (1989). *The Neurobiology of Saccadic Eye Movements*, chapter Metrics, pages 13–67. Elsevier.

Bishop, 2007. Bishop, C.M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

Bjorkman and Eklundh, 2004. Bjorkman, M. and Eklundh, J.-O. (2004). Attending, foveating and recognizing objects in real world scenes. In *British Machine Vision Conference (BMVC)*.

Bjorkman and Eklundh, 2006. Bjorkman, M. and Eklundh, J.-O. (2006). Vision in the real world: Finding, attending and recognizing objects. *International Journal of Imaging Systems and Technology*, 16(5):189–208.

Bjorkman and Kragic, 2004. Bjorkman, M. and Kragic, D. (2004). Combination of foveal and peripheral vision for object recognition and pose estimation. In *IEEE International Conference on Humanoid Robots (Humanoids)*, pages 5135–5140.

Blackwell, 1947. Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18(1):105–110.

Bradski, 2000. Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Breazeal et al., 2001. Breazeal, C., Edsinger, A., Fitzpatrick, P., and Scassellati, B. (2001). Active vision for sociable robots. *IEEE Transactions on Systems, Man and Cybernetics*, 31(5):443–453.

Breazeal and Scassellati, 1999. Breazeal, C. and Scassellati, B. (1999). A context-dependent attention system for a social robot. In *International Joint Conference on Artificial Intelligence*, pages 1146–1153.

Bridgeman et al., 1975. Bridgeman, G., Hendry, D., and Stark, L. (1975). Failure to detect displacement of visual world during saccadic eye movements. *Vision Research*, 15:719–722.

Briscoe, 2008. Briscoe, R. (2008). Egocentric spatial representation in action and perception. *Philosophy and Phenomenological Research*, 79(2):423–460.

Brooks et al., 1999. Brooks, R.A., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson, M.M. (1999). The Cog project: Building a humanoid robot. In *Lecture Notes in Computer Science*, pages 52–87. Springer-Verlag.

Bruce and Tsotsos, 2006. Bruce, N.D.B. and Tsotsos, J.K. (2006). Saliency based on information maximization. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 155–162. MIT Press.

Bruce and Tsotsos, 2009. Bruce, N.D.B. and Tsotsos, J.K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24.

Carpenter, 1988. Carpenter, R.H.S. (1988). *Movements of the Eyes (second edition)*. Pion Ltd.

Cave, 1999. Cave, K.R. (1999). The FeatureGate model of visual selection. *Psychological Research*, 62(2-3):182–194.

Chang and Krumm, 1999. Chang, P. and Krumm, J. (1999). Object recognition with color cooccurrence histograms. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2498–2504.

Chen and Zelinsky, 2006. Chen, X. and Zelinsky, G.J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research*, 46(24):4118–4133.

Choi et al., 2004. Choi, S.-B., Ban, S.-W., and Lee, M. (2004). Biologically motivated visual attention system using bottom-up saliency map and topdown inhibition. In *Neural Information Processing-Letters and Review*.

Coltheart, 1980. Coltheart, M. (1980). The persistences of vision. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 290(1038):57–69.

Coombs, 1992. Coombs, D.J. (1992). *Real-time gaze holding in binocular robot vision*. PhD thesis, The Univ. of Rochester, Rochester, NY, USA. UMI Order No. GAX92-18529.

Crow, 1984. Crow, F.C. (1984). Summed-area tables for texture mapping. In *Conference on ComputerGraphics and Interactive Techniques*, pages 207–212, New York, NY, USA. ACM.

Dankers et al., 2009. Dankers, A., Barnes, N., Bischof, W., and Zelinsky, A. (2009). Humanoid vision resembles primate archetype. In Khatib, O., Kumar, V., and Pappas, G., editors, *Experimental Robotics*, volume 54 of *Springer Tracts in Advanced Robotics*, pages 495–504. Springer Berlin / Heidelberg.

Dankers et al., 2007. Dankers, A., Barnes, N., and Zelinsky, A. (2007). A reactive vision system: Active-dynamic saliency. University of Bielefeld.

Dodge, 1900. Dodge, R. (1900). Visual perception during eye movement. *Psychological Review*, 7:454–465.

Dornaika and Horaud, 1998. Dornaika, F. and Horaud, R. (1998). Simultaneous robot-world and hand-eye calibration. *IEEE Transactions on Robotics and Automation*, 14(4):617–622.

Doucet et al., 2000. Doucet, A., Freitas, N. de, Murphy, K.P., and Russell, S.J. (2000). Rao-blackwellised particle filtering for dynamic bayesian networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 176–183.

Ekstrom et al., 2008. Ekstrom, L.B., Roelfsema, P.R., Arsenault, J.T., Bonmassar, G., and Vanduffel, W. (2008). Bottom-up dependent gating of frontal signals in early visual cortex. *Science*, 321(5887):414–417.

Ekvall and Kragic, 2005. Ekvall, S. and Kragic, D. (2005). Receptive field cooccurrence histograms for object detection. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 84–89.

Eriksen and St. James, 1986. Eriksen, C.W. and St. James, J.D. (1986). Visual attention within and around the field of focal attention: a zoom lens model. *Perception & Psychophysics*, 40(4):225–240.

Feldman, 1982. Feldman, J. (1982). Dynamic connections in neural networks. *Biological Cybernetics*, 46:27–39.

Fidler et al., 2009. Fidler, S., Boben, M., and Leonardis, A. (2009). Learning hierarchical compositional representations of object structure. In Dickinson, Sven, Leonardis, Aleš, Schiele, Bernt, and Tarr, Michael J., editors, *Object Categorization: Computer and Human Vision Perspectives*, pages 196–215. Cambridge University Press.

Figueira et al., 2009a. Figueira, D., Lopes, M., Ventura, R., and Ruesch, J. (2009a). From pixels to objects: Enabling a spatial model for humanoid social robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3049–3054.

Figueira et al., 2009b. Figueira, D., Lopes, M., Ventura, R., and Ruesch, J. (2009b). Towards a spatial model for humanoid social robots. In Lopes, Luís, Lau, Nuno, Mariano, Pedro, and Rocha, Luís, editors, *Progress in Artificial Intelligence*, volume 5816 of *Lecture Notes in Computer Science*, pages 287–298. Springer Berlin / Heidelberg.

Findlay and Brown, 2006. Findlay, J.M. and Brown, V. (2006). Eye scanning of multi-element displays: II. Saccade planning. *Vision Research*, 46(1-2):216–227.

Finlayson et al., 1998. Finlayson, G.D., Schiele, B., and Crowley, J.L. (1998). Comprehensive colour image normalization. In Burkhardt, Hans and Neumann, Bernd, editors, *European Conference on Computer Vision (ECCV)*, volume 1406 of *Lecture Notes in Computer Science*, pages 475–490. Springer Berlin / Heidelberg.

Frintrop et al., 2005. Frintrop, S., Backer, G., and Rome, E. (2005). Goal-directed search with a top-down modulated computational attention system. In Kropatsch, Walter G., Sablatnig, Robert, and Hanbury, Allan, editors, *Pattern Recognition*, volume 3663 of *Lecture Notes in Computer Science*, pages 117–124. Springer Berlin / Heidelberg.

Frintrop and Jensfelt, 2008. Frintrop, S. and Jensfelt, P. (2008). Attentional landmarks and active gaze control for visual slam. *IEEE Transactions on Robotics*, 24(5):1054–1065.

Frintrop et al., 2007. Frintrop, S., Jensfelt, P., and Christensen, H.I. (2007). Attentional robot localization and mapping. *IEEE International Conference on Computer Vision (ICCV)*.

Frintrop et al., 2010. Frintrop, S., Rome, E., and Christensen, H.I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception*, 7(1):1–39.

Gilles, 1998. Gilles, S. (1998). *Robust description and matching of images*. PhD thesis, University of Oxford.

Gratal et al., 2010. Gratal, X., Bohg, J., Bjorkman, M., and Kragic, D. (2010). Scene representation and object grasping using active vision. In *Workshop at IEEE International Conference on Intelligent Robots and Systems*.

Grimson et al., 1994. Grimson, W.E.L., Klanderman, G., O'Donnell, P.A., and Ratan, A.L. (1994). An active visual attention system to 'play where's waldo'. In *Image Understanding Workshop*, pages II:1059–1065.

Haralick et al., 1973. Haralick, R.M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621.

Hare and Lewis, 2003. Hare, J.S. and Lewis, P.H. (2003). Scale saliency: Applications in visual matching, tracking and view-based object recognition. In *Distributed Multimedia Systems 2003 / Visual Information Systems 2003*, pages 436–440. Knowledge Systems Institute.

Harris and Stephens, 1988. Harris, C. and Stephens, M. (1988). A combined corner and edge detection. In *Fourth Alvey Vision Conference*, pages 147–151.

Harris and Wolpert, 2006. Harris, C.M. and Wolpert, D.M. (2006). The main sequence of saccades optimizes speed-accuracy trade-off. *Biological Cybernetics*, 95:21–29.

Hartley and Zisserman, 2004. Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition.

Heidemann et al., 2003. Heidemann, G., Rae, R., Bekel, H., Bax, I., and Ritter, H. (2003). Integrating context-free and context-dependent attentional mechanisms for gestural object reference. In Crowley, James, Piater, Justus, Vincze, Markus, and Paletta, Lucas, editors, *Computer Vision Systems*, volume 2626 of *Lecture Notes in Computer Science*, pages 22–33. Springer Berlin / Heidelberg.

Hollingworth and Henderson, 2002. Hollingworth, A. and Henderson, J.M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, 28:113–136.

Horowitz, 2006. Horowitz, T.S. (2006). Revisiting the variable memory model of visual search. *Visual Cognition*, 14(4):668–684.

Hough, 1962. Hough, P. (1962). Method and means for recognizing complex patterns. U.S. Patent 3.069.654.

Hunt and Kingstone, 2003. Hunt, A.R. and Kingstone, A. (2003). Covert and overt voluntary attention: linked or independent? *Brain Research. Cognitive Brain Research*, 18(1):102–105.

Irwin, 1992. Irwin, D.E. (1992). Memory for position and identity across eye movements. *Journal of experimental psychology. Learning, memory, and cognition*, 18(2):307–317.

Irwin and Andrews, 1996. Irwin, D.E. and Andrews, R.V. (1996). *Attention and Performance XVI*, chapter Integration and accumulation of information across saccadic eye, pages 125–154. MIT Press.

Itti and Baldi, 2006. Itti, L. and Baldi, P. (2006). Bayesian surprise attracts human attention. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, pages 547–554. MIT Press, Cambridge, MA.

Itti and Baldi, 2009. Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306.

Itti et al., 1998. Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259.

Julier and Uhlmann, 1996. Julier, S. and Uhlmann, J.K. (1996). A general method for approximating nonlinear transformations of probability distributions. Technical report, University of Oxford.

Kadir and Brady, 2001. Kadir, T. and Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, 45:83–105.

Klatzky, 1998. Klatzky, R.L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial Cognition, An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*, pages 1–18, London, UK. Springer-Verlag.

Koch and Ullman, 1985. Koch, C. and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227.

Kragic et al., 2005. Kragic, D., Bjorkman, M., Christensen, H.I., and Eklundh, J.-O. (2005). Vision for robotic object manipulation in domestic settings. *Robotics and Autonomous Systems*, 52(1):85–100. Advances in Robot Vision.

Lee et al., 2005. Lee, K., Buxton, H., and Feng, J. (2005). Cue-guided search: a computational model of selective attention. *IEEE Transactions on Neural Networks*, 16(4):910–924.

Lee and Yu, 2000. Lee, T.S. and Yu, S. (2000). An information-theoretic framework for understanding saccadic behaviors. In Solla, S.A., Leen, T.K., and Muller, K-R., editors, *Advances in Neural Information Processing Systems*, volume 12, pages 834–840. MIT Press.

Levenberg, 1944. Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, 2:164–168.

Li et al., 1994. Li, M., , Betsis, D., and Lavest, J.-M. (1994). Kinematic calibration of the kth head-eye system. Technical report, Computational Vision and Active Perception Lab., Dept. of Numerical Anaysis and Computing Science, Royal Institute of Technology (KTH).

Li, 1998. Li, M. (1998). Kinematic calibration of an active head-eye system. *IEEE Transactions on Robotics and Automation*, 14(1):153–158.

Lindeberg, 1998. Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116.

Liu et al., 2008. Liu, H., Song, D., Rüger, S., Hu, R., and Uren, V. (2008). Comparing dissimilarity measures for content-based image retrieval. In Li, Hang, Liu, Ting, Ma, Wei-Ying, Sakai, Tetsuya, Wong, Kam-Fai, and Zhou, Guodong, editors, *Information Retrieval Technology*, volume 4993 of *Lecture Notes in Computer Science*, pages 44–50. Springer Berlin / Heidelberg.

Lowe, 1999. Lowe, D.G. (1999). Object recognition from scale invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1157.

Lowe, 2004. Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Maki et al., 2000. Maki, A., Nordlund, P., and Eklundh, J.-O. (2000). Attentional scene segmentation: Integrating depth and motion from phase. *Computer Vision and Image Understanding*, 78:351–373.

Mallon and Whelan, 2007. Mallon, J. and Whelan, P.F. (2007). Which pattern? biasing aspects of planar calibration patterns and detection methods. *Pattern Recognition Letters*, 28(8):921–930.

Marquardt, 1963. Marquardt, D.W. (1963). An algorithm for least squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441.

Marr, 1982. Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. W.H. Freeman, San Francisco.

Matthies and Shafer, 1990. Matthies, L. and Shafer, S.A. (1990). *Error modeling in stereo navigation*, pages 135–144. Springer-Verlag, New York, NY, USA.

McGowan et al., 1998. McGowan, J.W., Kowler, E., Sharma, A., and Chubb, C. (1998). Saccadic localization of random dot targets. *Vision Research*, 38(6):895–909.

Meilinger and Vosgerau, 2010. Meilinger, T. and Vosgerau, G. (2010). Putting egocentric and allocentric into perspective. In *Spatial Cognition VII*, volume 6222 of *Lecture Notes in Computer Science*, pages 207–221. Springer Berlin / Heidelberg.

Melcher, 2001. Melcher, D. (2001). Persistence of visual memory for scenes. *Nature*, 412(6845):401.

Metta et al., 2008. Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In *8th Workshop on Performance Metrics for Intelligent Systems*, PerMIS '08, pages 50–56, New York, NY, USA. ACM.

Mikolajczyk and Schmid, 2001. Mikolajczyk, K. and Schmid, C. (2001). Indexing based on scale invariant interest points. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 525–531.

Mikolajczyk and Schmid, 2004. Mikolajczyk, K. and Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86.

Miles, 1983. Miles, F.A. (1983). Plasticity in the transfer of gaze. *Trends in Neurosciences*, 6:57–60.

Mitroff et al., 2005. Mitroff, S.R., Scholl, B.J., and Wynn, K. (2005). The relationship between object files and conscious perception. *Cognition*, 96(1):67–92.

Montagnini and Chelazzi, 2005. Montagnini, A. and Chelazzi, L. (2005). The urgency to look: Prompt saccades to the benefit of perception. *Vision Research*, 45(27):3391–3401.

Montemerlo, 2003. Montemerlo, M. (2003). *FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem with Unknown Data Association*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Montemerlo et al., 2002. Montemerlo, M., Thrun, S., Koller, D., and Wegbreit, B. (2002). FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem. In *AAAI National Conference on Artificial Intelligence*, pages 593–598.

Moore and Fallah, 2004. Moore, T. and Fallah, M. (2004). Microstimulation of the frontal eye field and its effects on covert spatial attention. *Journal of Neurophysiology*, 91(1):152–162.

Moravec, 1980. Moravec, H. (1980). Obstacle avoidance and navigation in the real world by a seeing robot rover. In *tech. report CMU-RI-TR-80-03*. Robotics Institute, Carnegie Mellon University.

Moren et al., 2008. Moren, J., Ude, A., Koene, A., and Cheng, G. (2008). Biologically based top-down attention modulation for humanoid interactions. *International Journal of Humanoid Robotics*, 5(1):3–24.

Murphy and Russell, 2001. Murphy, K. and Russell, S. (2001). Rao-blackwellized particle filtering for dynamic bayesian networks. In *Sequential Monte Carlo Methods in Practice*.

Nakayama and Silverman, 1986. Nakayama, K. and Silverman, G.H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature*, 320(6059):264–265.

Natale et al., 2005. Natale, L., Orabona, F., Berton, F., Metta, G., and Sandini, G. (2005). From sensorimotor development to object perception. In *IEEE International Conference on Humanoid Robots (Humanoids)*, pages 226–231.

Navalpakkam and Itti, 2005. Navalpakkam, V. and Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, 45(2):205–231.

Navalpakkam and Itti, 2006. Navalpakkam, V. and Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2049–2056.

Neisser, 1967. Neisser, U. (1967). *Cognitive Psychology*. Prentice Hall.

Neubert and Ferrier, 2002. Neubert, J. and Ferrier, N.J. (2002). Robust active stereo calibration. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2525–2531.

Noble, 1988. Noble, J.A. (1988). Finding corners. *Image and Vision Computing*, 6(2):121–128.

Noles et al., 2005. Noles, N.S., Scholl, B.J., and Mitroff, S.R. (2005). The persistence of object file representations. *Perception & Psychophysics*, 67(2):324–334.

Obdrzálek and Matas, 2002. Obdrzálek, S. and Matas, J. (2002). Local affine frames for image retrieval. In *International Conference on Image and Video Retrieval*, pages 318–327, London, UK. Springer-Verlag.

Oliva et al., 2003. Oliva, A., Torralba, A., Castelhano, M.S., and Henderson, J.M. (2003). Top-down control of visual attention in object detection. In *International Conference on Image Processing*, pages 253–256.

Olshausen et al., 1993. Olshausen, B.A., Anderson, C.H., and Van Essen, D.C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700–4719.

Olson et al., 2003. Olson, C.F., Matthies, L.H., Schoppers, M., and Maimone, M.W. (2003). Rover navigation using stereo ego-motion. *Robotics and Autonomous Systems*, 43(4):215–229.

Orabona et al., 2005. Orabona, F., Metta, G., and Sandini, G. (2005). Object-based visual attention: a model for a behaving robot. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 89–96.

Orabona et al., 2007. Orabona, F., Metta, G., and Sandini, G. (2007). A proto-object based visual attention model. In Paletta, L. and Rome, E., editors, *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, volume 4840 of *Lecture Notes in Computer Science*, pages 198–215. Springer Berlin / Heidelberg.

Orin and Schrader, 1984. Orin, D.E. and Schrader, W.W. (1984). Efficient Computation of the Jacobian for Robot Manipulators. *International Journal of Robotics Research*, 3(4):66–75.

Ouerhani and Hugli, 2005. Ouerhani, N. and Hugli, H. (2005). Robot self-localization using visual attention. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 309–314.

Park and Martin, 1994. Park, F.C. and Martin, B.J. (1994). Robot sensor calibration solving AX = XB on the Euclidean group. *IEEE Transactions on Robotics and Automation*, 10(5):717–721.

Peters II et al., 2001. Peters II, R.A, Hambuchen, K.A., Kawamura, K., and Wilkes, D.M (2001). The sensory ego-sphere as a short-term memory for humanoids. In *IEEE International Conference on Humanoid Robots (Humanoids)*, pages 451–460.

Posner, 1980. Posner, M.I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, 32(1):3–25.

Posner and Dehaene, 1994. Posner, M.I. and Dehaene, S. (1994). Attentional networks. *Trends in Neuroscience*, 17(2):75–79.

Rao and Ballard, 1995. Rao, R.P.N. and Ballard, D.H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78:461–505.

Rao et al., 1996. Rao, R.P.N., Zelinsky, G.J., Hayhoe, M.M., and Ballard, D.H. (1996). Modeling saccadic targeting in visual search. In Touretzky, D.S., Mozer, M.C., and Hasselmo, M.E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 830–836. The MIT Press.

Rasolzadeh et al., 2010. Rasolzadeh, B., Björkman, M., Huebner, K., and Kragic, D. (2010). An active vision system for detecting, fixating and manipulating objects in the real world. *International Journal of Robotics Research*, 29:133–154.

Rasolzadeh and Eklundh, 2006. Rasolzadeh, B and Eklundh, J.-O. (2006). An attentional system combining top-down and bottom-up influences. International Cognitive Vision Workshop at ECCV2006.

Renninger et al., 2005. Renninger, L.W., Coughlan, J., Verghese, P., and Malik, J. (2005). An information maximization model of eye movements. *Advances in Neural Information Processing Systems*, 17:1121–1128.

Rensink, 2000. Rensink, R.A. (2000). The dynamic representation of scenes. *Visual Cognition*, 7(1):17–42.

Rensink, 2002. Rensink, R.A. (2002). Change detection. *Annual Review of Psychology*, 53:245–277.

Rensink et al., 1997. Rensink, R.A., O'Regan, K.J., and Clark, J.J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8:368–373.

Rizzolatti et al., 1987. Rizzolatti, G., Riggio, L., Dascola, I., and Umilta, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1A):31–40.

Rizzolatti et al., 1994. Rizzolatti, G., Riggio, L., and Sheliga, B. (1994). Space and selective attention. *Attention and Performance XV*, pages 231–265.

Ruesch et al., 2008. Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., and Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 962–967.

Salembier and Sikora, 2002. Salembier, P. and Sikora, T. (2002). *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Inc., New York, NY, USA.

Schmid et al., 2000. Schmid, C., Mohr, R., and Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172.

Schwartz, 1980. Schwartz, E.L. (1980). Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision Research*, 20(8):645–669.

Sharkey et al., 1993. Sharkey, P.M., Murray, D.W., Vandevelde, S., Reid, I.D., and McLauchlan, P.F. (1993). A modular head/eye platform for real-time reactive vision. *Mechatronics*, 3(4):517–535.

Shi and Tomasi, 1994. Shi, J. and Tomasi, C. (1994). Good features to track. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600.

Shiu and Ahmad, 1989. Shiu, Y.C. and Ahmad, S. (1989). Calibration of wrist-mounted robotic sensors by solving homogeneous transform equations of the form AX = XB. *IEEE Transactions on Robotics and Automation*, 5(1):16–29.

Shubina and Tsotsos, 2010. Shubina, K. and Tsotsos, J.K. (2010). Visual search for an object in a 3d environment using a mobile robot. *Computer Vision and Image Understanding*, 114(5):535–547. Special issue on Intelligent Vision Systems.

Swain and Ballard, 1991. Swain, M.J. and Ballard, D.H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

Theeuwes, 2004. Theeuwes, J. (2004). Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin and Review*, 11(1):65–70.

Thrun et al., 2005. Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic robotics*. Intelligent Robotics and Autonomous Agents. MIT Press.

Torralba, 2003. Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America*, 20(7):1407–1418.

Torralba et al., 2006. Torralba, A., Oliva, A., Castelhano, M.S., and Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113(4):766–786.

Treisman, 1986. Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, 255(5):114–125.

Treisman and Gelade, 1980. Treisman, A. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Treisman and Gormican, 1988. Treisman, A. and Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological Review*, 95(1):15–48.

Treisman et al., 1977. Treisman, A., Sykes, M., and Gelade, G. (1977). Selective attention and stimulus integration. In *Attention and Performance VI*, pages 333–361.

Trifa et al., 2007. Trifa, V.M., Koene, A., Moren, J., and Cheng, G. (2007). Real-time acoustic source localization in noisy environments for human-robot multimodal interaction. In *IEEE International Symposium on Robot and Human interactive Communication*, pages 393–398.

Truong et al., 1999. Truong, S., Kieffer, J., and Zelinsky, A. (1999). A cable-driven pan-tilt mechanism for active vision. In *Australian Conference on Robotics and Automation*, pages 172–177.

Tsai and Lenz, 1989. Tsai, R.Y. and Lenz, R.K. (1989). A new technique for fully autonomous and efficient 3d robotics hand/eye calibration. *IEEE Transactions on Robotics and Automation*, 5(3):345–358.

Tsotsos, 1989. Tsotsos, J.K. (1989). The complexity of perceptual search tasks. In *International Joint Conference on Artificial Intelligence*, pages 1571–1577, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tsotsos, 1990. Tsotsos, J.K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13:423–469.

Tsotsos et al., 1995. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545. Special Volume on Computer Vision.

Ude et al., 2003. Ude, A., Atkeson, C.G., and Cheng, G. (2003). Combining peripheral and foveal humanoid vision to detect, pursue, recognize and act. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 2173–2178.

Ude and Cheng, 2004. Ude, A. and Cheng, G. (2004). Object recognition on humanoids with foveated vision. In *IEEE International Conference on Humanoid Robots (Humanoids)*, pages 885–898.

Ude et al., 2004. Ude, A., Gaskett, C., and Cheng, G. (2004). Support vector machines and gabor kernels for object recognition on a humanoid with active foveated vision. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 668–673.

Ude et al., 2006. Ude, A., Gaskett, C., and Cheng, G. (2006). Foveated vision systems with two cameras per eye. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3457–3462.

Ude and Oztop, 2009. Ude, A. and Oztop, E. (2009). Active 3-d vision on a humanoid head. In *International Conference on Advanced Robotics*, pages 1–6.

Ude et al., 2005. Ude, A., Wyart, V., Lin, L.-H., and Cheng, G. (2005). Distributed visual attention on a humanoid robot. In *IEEE International Conference on Humanoid Robots (Humanoids)*, pages 381–386.

Ullman, 1984. Ullman, S. (1984). Visual routines. *Cognition*, 18:97–159.

Van der Stigchel and Theeuwes, 2007. Van der Stigchel, S. and Theeuwes, J. (2007). The relationship between covert and overt attention in endogenous cuing. *Perception & Psychophysics*, 69(5):719–731.

Vijayakumar et al., 2001. Vijayakumar, S., Conradt, J., Shibata, T., and Schaal, S. (2001). Overt visual attention for a humanoid robot. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 2332–2337.

Welke et al., 2008a. Welke, K., Asfour, T., and Dillmann, R. (2008a). Object separation using active methods and multi-view representations. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 949–955.

Welke et al., 2009. Welke, K., Asfour, T., and Dillmann, R. (2009). Bayesian visual feature integration with saccadic eye movements. In *IEEE International Conference on Humanoid Robots (Humanoids)*, pages 256–262.

Welke et al., 2011. Welke, K., Asfour, T., and Dillmann, R. (2011). Inhibition of return in the bayesian strategy to active visual search. In *IAPR Conference on Machine Vision Applications*.

Welke et al., 2010. Welke, K., Issac, J., Schiebener, D., Asfour, T., and Dillmann, R. (2010). Autonomous acquisition of visual multi-view object representations for object recognition on a humanoid robot. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Welke et al., 2007. Welke, K., Oztop, E., Cheng, G., and Dillmann, R. (2007). Exploiting similarities for robot perception. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 3237–3242.

Welke et al., 2008b. Welke, K., Przybylski, M., Asfour, T., and Dillmann, R. (2008b). Kinematic calibration for saccadic eye movements. Technical report, Institute for Anthropomatics, Universität Karlsruhe.

Williams and Reingold, 2001. Williams, D.E. and Reingold, E.M. (2001). Preattentive guidance of eye movements during triple conjunction search tasks: the effects of feature discriminability and saccadic amplitude. *Psychonomic Bulletin Review*, 8(3):476–488.

Wolfe, 1996. Wolfe, J.M. (1996). Visual search. In Pashler, H., editor, *Attention*. University College London Press, London, UK.

Wolfe et al., 1992. Wolfe, J.M., Friedman-Hill, S.R., Stewart, M.I., and O'Connell, K.M. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):34–49.

Yantis, 2000. Yantis, S. (2000). *Attention and Performance*, chapter Goal-directed and stimulus-driven determinants of attentional control, pages 73–103. Cambridge, MA: MIT Press.

Young et al., 1992. Young, G.-S., Hong, T.-H., Herman, M., and Yang, J.C.S. (1992). Kinematic calibration of an active camera system. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 748–751.

Zhang et al., 2008. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., and Cottrell, G.W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7).