Foundations and Trends® in
Robotics
Vol. 1, No. 1 (2010) 1–78

DOI: 10.1561/2300000001

Vision for Robotics

By Danica Kragic and Markus Vincze

Contents

Foundations and Trends® in
Robotics
Vol. 1, No. 1 (2010) 1–78
© 2009 D. Kragic and M. Vincze
DOI: 10.1561/2300000001

now
the essence of knowledge

Vision for Robotics

Danica Kragic[1] and Markus Vincze[2]

[1] *Centre for Autonomous Systems, Computational Vision and Active Perception Lab, School of Computer Science and Communication, KTH, Stockholm, 10044, Sweden, dani@kth.se*
[2] *Vision for Robotics Lab, Automation and Control Institute, Technische Universitat Wien, Vienna, Austria, vincze@acin.tuwien.ac.at*

Abstract

Robot vision refers to the capability of a robot to visually perceive the environment and use this information for execution of various tasks. Visual feedback has been used extensively for robot navigation and obstacle avoidance. In the recent years, there are also examples that include interaction with people and manipulation of objects. In this paper, we review some of the work that goes beyond of using artificial landmarks and fiducial markers for the purpose of implementing vision-based control in robots. We discuss different application areas, both from the systems perspective and individual problems such as object tracking and recognition.

1

Introduction

For many living species, not least in the case of humans, visual perception plays a key role in their behavior. *Hand–eye coordination* ability gives us flexibility, dexterity, and robustness of movement that no machine can match yet. To locate and identify static, as well as moving objects, to determine how to grasp and handle them, we often rely strongly on our visual sense. One of the important factors is our ability to *track* objects, that is, to maintain an object in the field of view for a period of time using our oculomotor system as well as head and body motions. Humans are able to do this quickly and reliably without much effort. It is therefore natural to expect that the artificial cognitive systems we aim at developing will, to a certain extent, be able to demonstrate similar capabilities.

Robot vision refers to the capability of a robot to visually perceive the environment and interact with it. Robot vision extends methods of computer vision to fulfill the tasks given to robots and robotic systems. Typical tasks are to navigate toward a given target location while avoiding obstacles, to find a person and react to the person‘s commands, or to detect, recognize, grasp and deliver objects.

Thus, the goal of robot vision is to exploit the power of visual sensing to observe and perceive the environment and react to it. This follows the example of humans. It has been found that more than half of the human sensory cortex is attributed to seeing. Computer vision attempts to achieve the function of understanding the scene and the objects of the environment. With the increasing speed of processing power and progress in computer vision methods, making robots see became a main trend in robotics.

There, however, remains a fundamental difference between computer vision and robot vision. Computer vision targets the understanding of a scene mostly from single images or from a fixed camera position. Methods are tailored for specific applications and research is focused on individual problems and algorithms. On the other hand, robot vision requires to look at the system level perspective, where vision is one of several sensory components that work together to fulfill specific tasks. This property of the robotic system is also referred to as embodiment, where similar to biological systems the properties of the body shape the tasks of perception. Vision is used as a mean for the robot to act in and interact with the world–a robot system perceives to act and acts to perceive. Hence, visual processing is not an isolated entity, but part of a more complex system.

The future expectation is that robots will become ubiquitous. To robustly and safely interact with the world, robots need to perceive and interpret the environment so as to achieve context awareness and act appropriately. In general, we want to equip robots with minimal information in advance and get them to gather and interpret the necessary information required for execution of new tasks through interaction and on-line learning. This has been a long-term goal and one of the main drives in the field of artificial cognitive systems development. As an example, for a service robot that is to perform tasks in a human environment, it has to be able to learn about objects and object categories. However, the robots will not be able to form useful categories or object representations by being a passive observer of the environment. They should, like humans, learn about objects and their representations through interaction.

Vision has been used in robotic applications for more than three decades. Examples include applications in industrial settings, service, medical, and underwater robotics, to name some. In this paper we review some of the aspects of robot vision from early beginnings to more recent works. We concentrate in particular on attempts of developing active vision systems and examples where visual processing is considered as a primary aspect of the work rather than just a necessary input to the control loop.

There are many characteristics in common in computer vision research and vision research in robotics. For example, the Structure-and-Motion problem in vision has its analog of SLAM (Simultaneous Localization and Mapping) in robotics, visual SLAM being one of the important topics. Tracking is another area seeing great interest in both communities, in its many variations, such as 2D and 3D tracking, single and multi-object tracking, rigid and deformable object tracking. Other topics of interest for both communities are object and action recognition. In the subsequent sections, we will discuss the differences in more detail.

1.1 Scope and Outline

Visual feedback enables robots to interact with the environment in various ways. In some cases, visual feedback is used for navigation and obstacle avoidance, while more complex examples include interaction with the user and manipulation of objects. The simplest interaction that can occur between a robot and an object may be to, for example, push an object in order to retrieve information about the size or weight of the object. Here, simple visual cues providing approximate 3D position of the object may be sufficient. A more complex interaction may be to grasp the object for the purpose of gaining the physical control over the object. Once the robot has the object in its hand, it can perform further actions on it, such as examining it from other views. Information obtained during interaction can be used to update the robots representations about objects and the world.

In cases where visual feedback is input for robot localization, mapping, or obstacle avoidance algorithms, extraction of low level visual features such as corners, interest features such as SIFT [132], or optical flow may be sufficient. Hence, visual feedback facilitates only state estimation step and no advanced reasoning is needed to explain what is really happening in a video sequence.

For the applications we envision in the future, this is not enough. We need vision systems that are able to provide adequate information no matter if the system is to manipulate an object or interact with a human. We need systems that understand what they "see" according to known or autonomously acquired models: these systems must perceive to act and act to perceive. An example may be a robot that enters a room, detects a table from a few meters distance, localizes a number of objects on it, and shifts it gaze toward each of them to obtain a more detailed foveal view of the whole or parts of an object. This information can then be used to either approach an object for picking it up or for storing the information about typical object positions in the environment. The processes that are necessary here are figure-ground segmentation and attention–these are commonly not considered in specific applications of object tracking or recognition.

Thus, the nature and level of detail of the extracted visual information depends on several factors: (i) the task a robot system is required to accomplish, (ii) number and position of visual sensors, (iii) required processing rate and (iv) indoor/outdoor environment, to name some. In this paper, we discuss different applications of visual input, both from the systems perspective and individual problems such as object tracking and recognition. This is structured as follows.

The discussion starts with Chapter 2, where we give an overview of methods from the early days and the use of vision in industrial applications (Section 2.1) to more recent trends in robot vision taking into account findings from biology, neuroscience, and cognitive science (Section 2.2). As last part of this section we stress the importance of considering not only individual functions in robot vision but also robot vision *systems*.

A tentative model of a robot vision system is shown in Figure 1.1. The overview aims at indicating that, at this rather abstract level of description, a robot vision system fulfills three major functions: navigation, grasping, and Human Robot Interaction (HRI). The interplay of these functions depends on the task. For example, navigation is today

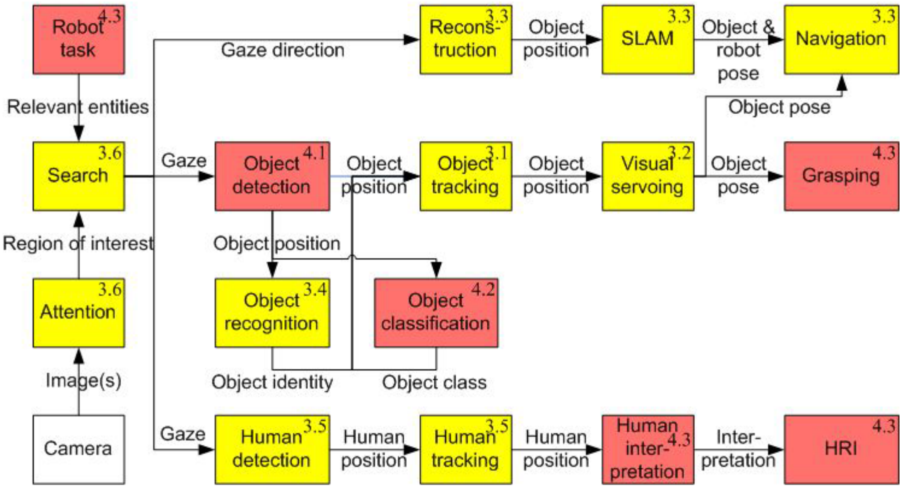

Fig. 1.1 Block diagram of the main tasks of a robot vision system: navigation, grasping and Human Robot Interaction. The numbers refer to Sections. Yellow indicates Chapter 3 “What works” and red indicates Chapter 4 “Open challenges”. Please see text for more details.

considered a largely solved problem with methods suitable for applications and advanced topics open to research. Thus, in Chapter 3 we present aspects of robot vision for which robust performance has been achieved. This is indicated by boxes colored in yellow in Figure 1.1. In Chapter 4 we review the open challenges that are still considered unsolved (indicated in red) and more related to formalizing the semantics of robot tasks and binding them to grasping and HRI. Finally, the review ends with a discussion and a short outlook in Chapter 5.

We note that the strict sequence of functions in Figure 1.1 is only for clarity. There are several approaches that combine functions and establish direct links that are not shown. Other functions, such as adaptation of functions to specific tasks or learning are also not explicitly given, may apply to several of the blocks, and will be mentioned when appropriate.

2

Historical Perspective

2.1 Early Start and Industrial Applications

In his seminal book, Horn [90] provides the first thorough analysis of computer vision topics related to the robot domain. Techniques, such as optical flow, inherent to a body moving in the environment, are developed rigorously for the first time. However, it turned out that the methods were not yet practical enough for real-time applications. Only recently processing power became sufficient to compute optical flow at reasonable rates. Already in the early days it was discovered that the robustness of the methods was not sufficient and this is a problem studied even today.

First laboratory experiments demonstrated already in 1973 the principal feasibility of using vision to correct the position of the robot to increase task accuracy [182]. Objects were dark and clearly visible on a bright background, greatly simplifying the visual processing. Throughout the years industrial applications of visually guided robots demonstrated robust performance while being practically blind–working with known objects and highly optimized visual processing streams.

After the early work in the 1970s, robot vision evolved with the advances in information and silicon technologies and put the old ideas into work, both in the research and industrial settings. Despite the hardware development, there is close to 90% of industrial robots that are still teach-in programmed. A number of machine vision vendors, however, supply robot controllers that consider "more intelligent" visual processing. Examples are overhead cameras that recognize and localize parts on a conveyor belt. The approaches are commonly two-dimensional, parts may overlap only if they are flat, and controlled illumination produces good contrast for segmentation. The robot is calibrated with respect to the area the vision system surveys and operates in an open-loop mode referred to as *look-then-move*. The known conveyor velocity is superimposed on the robot motion and the robot picks up the part without further external feedback. There exist commercial products that determine the object location, for example, from ADEPT or ABB. These packages require the objects to be separated, to have good contrast, and a unique circumference of the part. The location is determined in two dimensions based on a calibration to the ground plane constraint of the conveyor belt. Small displacements of large parts can be corrected in three dimensions using three or more 2D systems that rely on simple features such as dark holes (e.g., ISRA Vision Systems, Germany; Vitronik, Germany; and Volkswagen, Germany). The full 3D location of parts can be measured with range cameras that project a patterns onto the part (e.g., Integrated Vision Products, SE; EADS Lasercamera, Germany). These systems are still costly and relatively large. Technical advance renders it possible to execute several of the above steps in parallel. The result is a continuous control by interlacing sensing and motion. The objective is to obtain fast motion of the robot while approaching the desired final pose, a necessity of any commercial installation. First successes in autonomous car driving and air vehicle guidance indicate the advance of vision in systems technology [6, 48].

It has been recognized that, besides the price, accuracy, easy integration, modularity, and flexibility, there are two major requirements for commercialization in real-world scenarios [213]:

(1) There are several examples of using vision as an integral part of more complex systems. Vision and control must be coupled to assure good dynamic performance. Real-time

performance is needed to justify the use of vision-based control commercially.

(2) Vision system must be robust and reliable: it must be able to evaluate the state of the environment to enable a reaction to changes and to assure the safety of the robot and its environment.

Even if biological vision systems are not perfect in their performance, the above two seem not to be a big problem. Below, we shortly discuss some of the biological influences on the design of artificial vision systems.

2.2 Biological Influences and Affordances

Biological vision systems are active: we are able to control eye and neck movements to direct the attention on particular parts of scenes we observe. The ability to attend to parts of the scene takes away the need to process a large amount of visual information at all times and concentrate on what it is important for the task at hand. Biological systems can do this fast and with a high accuracy. In the light of perception–action coupling, research in humans and primates has provided inspiration for development of artificial perception–action systems. Recent neuroscientific findings show that tasks such as object grasping and manipulation are realized through distributed information flows between multiple regions within the nervous system [167, 169, 62, 165].

The first comprehensive summary of the biological vision system was published by Marr [137]. He introduced a theory based on a series of abstraction levels that guide the processing from the image over a 2D sketch to a structured and object-centered 3D model. The view put forward by Marr is that vision is seen as a grouping and reconstruction process of 3D shape models. In particular brain research has challenged several of the claims and changed general views. An example is that there exist different streams of processing that are tailored to specific tasks. An example is the processing in *what* and *where* streams [73] or recognition of objects from parts of the objects [196]. Nevertheless, it remains an attractive and clear formulation of the visual interpretation and even today many works are still influenced by these structured levels of processing.

Of particular interest to robot vision is the work started in the 1980s on active vision. It was put forward by Bajcsy who argued that the problem of perception was not of image processing nature but of control of data acquisition [7]. The work was influenced by the ecological approach to perception as formulated by Gibson [71]. According to Gibson's information pick-up theory, the environment consists of affordances, e.g., terrain, water, and vegetation, that provide the cues necessary for perception. Information is actively and continuously generated and updated: an active organism actively searches for invariants that are linked to the task. Already here the idea of embodied vision shows up, which is popular again today.

The active vision paradigm has been pushed forward by the works of Aloimonos et al. [4] and Ballard [8], who proposed approaches to combine the different visual observations with a priori information in such a way that the process achieves a common task. Given the robotic embodiment, this includes the active control of the gaze direction of the cameras. The active vision paradigm has several consequences on the level of a robot vision system [41]: (i) the system is always running, (ii) it filters relevant information, (iii) it works in real time within a fixed delay to be useful, and (iv) it processes a region of interest in order to meet the performance goals. Section 2.3 below will further discuss this aspect.

Following the paradigm of active vision, two lines of research emerged: work on visual attention and work that more closely integrates robot's action with the visual feedback. The area of visual attention can be summarized as a mechanism and methods that optimize the search and detection processes inherent in vision [206]. Visual attention seems necessary due to inherent limits in processing capacity in the brain. While visual attention is a rather accepted branch in the computer vision community, robot vision is still seen little in the major conferences of either discipline. Section 3.6 will highlight the work in this area.

A closer integration of the robot mechanism with the vision processing is manifested in a series of works on visual servoing, where the robot actively follows the object motion for tracking, navigation, or grasping. Continuous vision-based feedback control of the position of the robot is referred to as Visual Servoing (VS) [87, 63]. This term today encompasses also control of active vision heads, vehicles, or any other mechanisms that are vision-controlled. The control problem has received a lot of attention in the literature (e.g., [39, 92]), but robust visual tracking is just as critical and has received little attention [213]. Work in this direction is reviewed in detail in Section 3.2.

There is also a trend in the integration of findings from biological vision into computer vision approaches such as object recognition [177]. Section 3.4 reviews approaches to object recognition and categorization. With the strong emphasis on recognition in the computer vision community, several benchmarking data sets have been created and established (e.g., PASCAL network). However, these provide rather a playground for sophisticated machine learning techniques that mostly use appearance-based image descriptors. Recent work shows that these kinds of techniques do scale to a large number of different object classes and that they can even be learnt from quite cluttered and loosely labeled training data. Section 4.4 lists notable databases.

Another trend is to explore developmental approaches to build artificial cognitive systems. Developmental approaches focus on the autonomous self-organization of general-purpose, task non-specific control systems. The approaches are inspired by developmental psychology and developmental neuroscience. Developmental robotics is a move away from task-specific methodologies where a robot is designed to solve a particular predefined task. The idea is very much in line with enabling robots to adapt and learn the necessary capabilities instead of being completely pre-programmed. The spectrum of work is very wide and starts from work on motivation to learn from control, morphology, and to work on emotions.

A series of works developed this approach with respect to robot vision. The robot first learns to observe itself and then starts to investigate objects in the environment by poking them [141]. Poking moves the object and drastically simplifies the foreground/background separation due to the induced motion. Additionally, object characteristics such as pokiness or size can be inferred. A recent work manipulated a few objects and tried to infer affordances such as sliding or rolling of the detected objects [145]. Color is used to segment blob-like regions and these blob features such as maximal extensions are used to infer object shape in the image. Circle and square are used to infer sphere and cube like objects and to adhere the detected affordances to them. There is still a large body of work to be done toward handling realistic objects and extracting their 3D shape for enabling experiments with robotic systems in unstructured environments.

2.3 Vision Systems

In recent years an increasing tendency can be observed toward development of robot and computer vision *systems*. The conference series on computer vision Systems (ICVS) fosters this and assembles work not visible in the computer vision community and partially the robotics conferences. Besides robot vision, human-centered vision systems and vision for human robot interaction (HRI) is gaining a lot of attention. The system trend is also reflected in the cognitive vision area (e.g., www.ecvision.org), which has expanded into an inherently interdisciplinary approach on artificial cognitive systems. In this area the focus has shifted from specific vision techniques toward fundamental principles on how cognitive abilities, like vision, emerge and how specific knowledge can be acquired from the interaction with the environment. A versatile robot vision system remains to be developed that can demonstrate many different capabilities.

With limited resources in terms of memory storage and computational power, both biological and robotic systems need to find an acceptable balance between the size of the visual field and its resolution. Otherwise, the amount of visual data is too large for the system to be handled efficiently. This balance depends also on the tasks the systems have to perform. An animal that has to stay alert in order to detect an approaching predator would prefer a wide field of view. The opposite is true if the same animal acts as a predator itself. Similarly, a robotic system benefits from a wide field of view, in order not to collide with obstacles while navigating through a cluttered environment. A manipulation task, on the other hand, requires a high resolution in order to grasp and manipulate objects. That is, to find objects in the scene a wide field of view is preferable, but recognizing and manipulating the same objects require a high resolution.

Although there are systems that demonstrate the use of monocular vision for visual servoing, most of the robot systems that move about in the environment and interact with people and objects use binocular setups. Using two cameras simplifies the problem of reconstructing the 3D structure and obstacle detection [207, 146, 91]. A related example is presented in [19] that uses a combination of two pairs of cameras, a peripheral set for attention, and a foveated one for recognition and pose estimation. In this work, in order to facilitate transfers of object hypotheses from one pair to the other, and replicate the nature of the human visual system, the pairs were placed next to each other. With a binocular set of cameras, differences in position between projections of 3D points onto the left and right image planes (disparities) were used to perform figure-ground segmentation and retrieve the information about 3D structure of the scene. When the relative orientation and position between cameras is known, the disparities can be mapped to actual metric distances. One of the commonly used settings is where the cameras are rectified and their optical axes are mutually parallel. However, one of the problems arising is that the part of the scene contained in the field of view of both cameras simultaneously is quite limited.

Following the very generic description of a robot vision system in Figure 1.1, a concrete example of a vision system has been presented in [19]. It consists of the modules as shown in Figure 2.1:

- Visual Front-End: extracts visual information needed for figure-ground segmentation and other higher level processes.
- Hypotheses Generation: produces hypotheses about the objects in the scene relevant to the task at hand.
- Recognition: uses either image features or color histograms to determine the relevancy of observed objects.
- Action Generation: triggers actions, such as visual tracking and pose estimation, depending on the outcome of the recognition and current task specification.

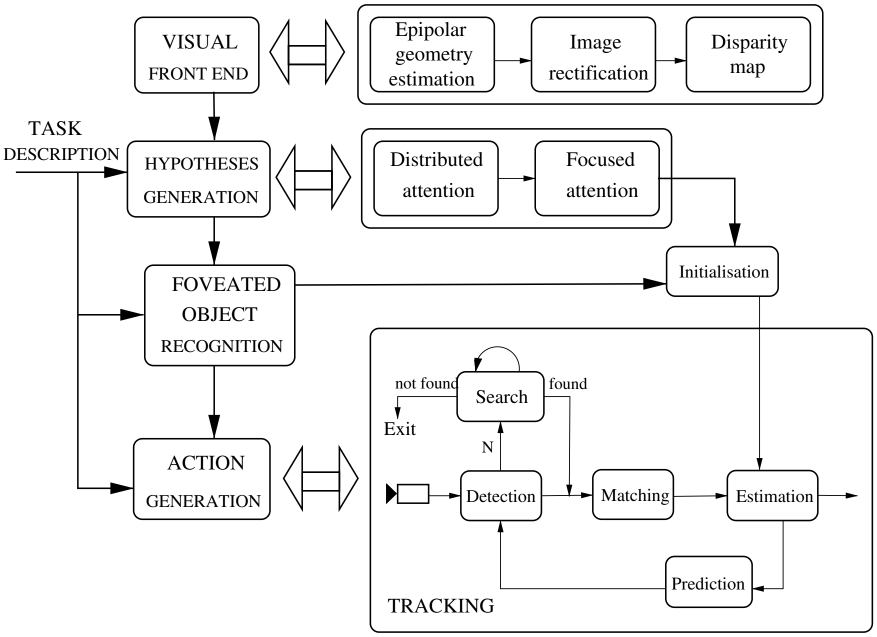

Fig. 2.1 Basic building blocks of the vision system presented in [19].

Vision as Process presented a vision-controlled robot system utilizing a binocular stereo head. The project developed a software system to integrate vision modules on an active head [40]. Integration united 2D image data with 3D structure data to control the head motion. Objects had different gray values on each surface or white markers and objects on black background. Following the idea of robot vision as an active process, the construction of vision systems further advanced to the ambition of developing cognitive Computer Vision Systems (CVS) [33]. This term is used to characterize systems that not only involve computer vision algorithms but also employ techniques of machine learning in order to acquire and extend prior knowledge. Furthermore, they aim at using automatic and contextual reasoning to verify the consistency of results obtained from several modules as well as to manage the coordination of these modules, where modules are typical vision functions such as detection, recognition, and tracking of relevant entities (humans, objects, environment).

The Intelligent Service Robot (ISR) System at KTH investigated methods for systems integration and perception in a domestic or an office setting. It was one of the first studies of service robotics for office and home applications. Cue integration was used for object tracking [114], while color and motion were used for gesture and person tracking [184]. Tracking was then coupled with a method for automated grasp planning [157]. The project was special as all the components were investigated in realistic scenarios to learn to cope with the environments common for service robot applications.

In order to ensure architectural soundness of integrating different modules, the use of frameworks is thus mandatory. These frameworks are specialized in the sense that they are tailored to certain project specific requirements and thus are of limited generality. However, there are common needs in traditional as well as in cognitive computer vision that can easily be identified, e.g., efficient handling of image data. These of course provide general criteria that can guide a comparison of frameworks.

Practical experience shows that building vision systems not only requires domain specific requirements but also faces problems of programming in the large. Examples encountered in practice are reusability, scalability, or transparency. As a consequence, techniques and approaches from software engineering are used when designing vision systems. Owing to the increasing importance of this topic, Wrede et al.'s study [218] is a first attempt on evaluating of integration frameworks. Noticeable frameworks targeted for the demands of cognitive vision systems are highlighted in several projects financed by the European Commission in the area of cognitive systems. We mention two of these below.

In the PACO-PLUS project (*www.paco-plus.org*), the aim is a design of a cognitive robot that is able to develop perceptual, behavioral, and cognitive categories in a measurable way and communicate and share these with humans and other artificial agents. The main paradigm of the project is that Objects and Actions are inseparably intertwined and that categories are therefore determined (and also limited) by the action an agent can perform and by the attributes of the world it can perceive; the resulting, so-called Object-Action Complexes (OACs) are the entities on which cognition develops (action-centered cognition) [111, 217]. Entities ("things") in the world of a robot (or human) will only become semantically useful "objects" through the action that the agent can/will perform on them.

The aim of a project in activity interpretation [214] was to develop a cognitive vision methodology that interprets and records the activities of people handling tools. Focus is on active observation and interpretation of activities, on parsing the sequences into constituent behavior elements, and on extracting the essential activities and their functional dependence. The expert activities are interpreted and stored using natural language expressions in an activity plan. The final outcome is an indexed manual in the form of 3D reconstructed scenes, which can be replayed at any time and location to many users using Augmented Reality equipment.

3

What Works

Today, vision-based control is demonstrated in various applications. Different variants of target tracking represent one of the most important building blocks of a robot vision system. In some cases, retrieving only the image position may be enough. 2D tracking approach may, for example, be used to maintain the target in the field of view (surveillance) or for applications where the accuracy is not the crucial parameter–when the object is relatively far from the camera and tracking is used to keep a robot “on the right” path while approaching the object. Some of the 2D approaches use very little *a priori* information about the object which is both an advantage and a difficulty. Since there is no knowledge of the different views of the object, significant changes in the object pose may result in a loss of tracking if not appropriately accounted for in the tracking method.

The visibility and appearance of the object in the image depends on the geometry of the object and its pose relative to the camera. One way to cope with the problems outlined above is to build and maintain a 3D model of the object which facilitates the estimation of its pose. Hence, a tracking system may be designed to continuously update the state of the object/model. The type of the model used will depend on the application of the tracking system, required accuracy, the geometry of the object, its appearance, etc.

The following sections will discuss the necessary ingredients of object tracking, pose estimation, and its use in the visual servoing loop.

3.1 Object Tracking and Pose Estimation

Depending on the application, there may be requirements from the tracking system to:

(1) handle temporal inconsistencies in appearance and occlusions of the target,
(2) reinitialize the tracking once the target has left the field of view,
(3) adapt to unpredictable object motion,
(4) be insensitive to lighting conditions and specular reflections,
(5) perform in "real-time", and
(6) use minimum *a priori* knowledge about the tracked object.

This list is far from being complete. So, what makes things difficult? Lack of robustness is primarily due to three problems: (i) figure-ground segmentation–detection of the target or initialization of the tracking sequence, (ii) matching across images, in particular in the presence of large and varying inter-frame motion, and (iii) inadequate modeling of motion to enable prediction of the target in subsequent images.

To obtain robustness, integration of visual cues has been proposed [3, 34]. Better tracking and initialization can be obtained by using several cues to more reliably locate a specific object characteristic. The work in [23] uses m-out-of-n voting on four cues (perspective distortion of texture, intensity, 5-point-invariant, disparity) plus an estimator to verify the existence of texture. Voting is also applied to figure-ground segmentation using typical cues of target objects such as motion, color, and intensity. Results indicate that plurality voting gives best results [162], while later work improves the results by using an unsupervised learning approach in a probabilistic framework [84].

Another approach is to exploit sequential modeling of the target object. On of the examples in this direction uses the idea of selecting different tracking techniques depending on the target object and to build an automatic initialization procedure. The technique of Incremental Focus of Attention (IFA) places the different tracking techniques in a state machine of search levels and tracking levels [201]. A predefined hierarchy of trackers based on more and more specific cue extraction methods is invoked to track the target. The idea is to fall back to lower resolution trackers for recovery if a higher resolution cue or tracker fails. Faces are tracked by first locating face color at low resolution and then following the face accurately with a template tracker [201]. To find door handles, the search first looks for vertical edges and then uses an image template to locate the handle [64]. This scheme allows to recover from failure and to automatically find a target that can be described with such a series of search and tracking levels. On the other hand, all cues must be salient to finally locate the object. In [201] the authors propose to use this scheme for recovery after failure. The argument is that robustness of systems will never be perfect and therefore a scheme to automatically recover can improve overall system performance. Another approach to utilize knowledge about the target object is to automatically use the model knowledge to obtain the tracking levels [215]. The idea is to select cues that can be found more easily in the image and to subsequently refine the search. The advantage is that some cues can fail or do not need to be salient.

3.1.1 2D Image Based Tracking

Here we review tracking methods that use single or several cues to estimate the image position of the target. Examples include tracking of objects and humans or parts of humans such as heads or hands. The methods can be classified into two main groups: (i) methods that track local features or image cues such as line segments, edges or color and (ii) methods that rely directly on the image intensity. The former are commonly sensitive to feature detection and cannot be applied to images that do not contain features that are distinctive.

The latter methods estimate the movement, the deformation or the illumination parameters of a reference template between two frames by minimizing an error measure based on image intensity. These methods are commonly based on tracking of region templates. A template is a 2D entity that represents a portion of an image [81]. During the tracking sequence, the object of interest can be represented by one 2D template or within a multi-template framework where the configuration of individual templates is constrained by some model-based information.

Regarding templates or region-based tracking, two approaches have been considered in the literature: optical flow-based tracking and correlation-based tracking.

Smith [188] developed a system for detection and tracking of independently moving objects against a non-stationary background. Motion was estimated through tracking of image features (corners and edges) and segmentation was based on an affine motion model. The system was tested on video streams taken from a moving platform–a vehicle traveling along the road.

Brandt et al. [22] developed a system using the sum of squared differences (SSD) optical flow measurements as input to the visual control loop. Hager and Toyoma [81] developed the XVision system that has been widely used for manipulation tasks [78]. The system gives a possibility for off-line model selection and performs well when there is good agreement between the model and the actual motion. However, for the case of unexpected object motions the result is usually a loss of tracking. Therefore, there is a need for a system that adaptively selects a motion model in response to current image changes. As pointed out in [180], translational (rigid) motion model gives more reliable results than an affine one when the inter-frame camera motion is small. However, affine changes are necessary to compare distant frames to allow determination of dissimilarity. One example of how to make the similarity metrics invariant to more complex distortion, such as affine changes or variations in local scene illumination, is presented in [79]. Here, the information from several templates is used and each template represents the same feature under different illumination conditions. A more recent approach of [11] proposes a homography-based approach tracking using an efficient second-order minimization method. The output of the visual tracking is a homography linking the current and the reference images of a planar target.

In [34], integration of multiple cues is studied and several problems in machine vision are addressed. The authors classify methods for integration of visual cues into *weak coupling* and *strong coupling*. Weak coupling combines the outputs of different cues while in strong coupling the output of one cue affects the output of another cue. In [112], the weak coupling approach is adopted where the redundancy of visual cues is exploited.

There are several notable contributions in terms of multi-cue integration for visual tracking. The *Incremental Focus of Attention* (IFA) architecture [201] mentioned above, uses a multi-layered framework where each layer in the framework is an algorithm, denoted either as a *selector* or a *tracker*. The IFA has a pyramid structure - at the top, there are high precision trackers and low-resolution trackers/selectors at the bottom.

Prokopowicz et al. [164] describe an approach to visual tracking in which one out of several tracking cues (color, motion, disparity, correlation) is chosen depending on the situation. The emphasis of the work is on how to obtain enough *a priori* information about the target and use that information to choose the suitable cue. This implies that the emphasis is on the higher level related more to cognitive aspects. The authors argue that, if there are no hardware limitations in terms of speed and storage, multiple cues should be used concurrently.

Shirai et al. [183] use a probabilistic framework to integrate optical flow, disparity, and regions of uniform brightness for people tracking. Uhlin [208] integrates optical flow-based motion segmentation (background and object) and depth from stereo to achieve dynamic binocular fixation. These cues are integrated on a binary basis. To achieve fast processing, a transputer network is used.

3.1.2 Model-Based 3D Object Tracking

A typical model-based tracking system usually involves the following steps: detection, matching, pose estimation, update, and prediction of

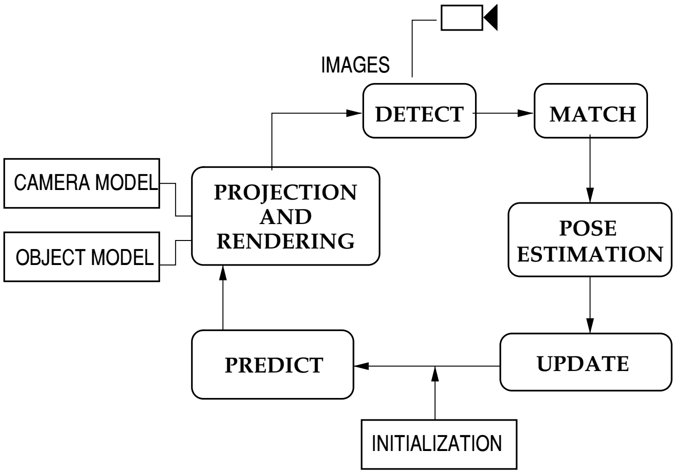

Fig. 3.1 Block diagram of a model-based tracking system (from [112]).

the state used to render the model of the object into the image, see Figure 3.1.

In other words, the input to the algorithm is usually a model of the object. This model is then used during the *initialization* step where the initial pose the object relative to the camera or some other coordinate system is estimated. The main loop starts with a *prediction* step where the state of the object is predicted using the current pose (velocity, acceleration) estimate and a motion model. The visible parts of the object are then projected into the image (*projection and rendering* step). After the *detection* step, where a number of features are extracted in the vicinity of the projected ones, these new features are *matched* to the projected ones and used to estimate the new pose of the object. Finally, the calculated pose is input to the *update* step.

Model-based 3D object tracking has earned significant importance in areas such as augmented reality, surveillance, visual servoing, robotic object manipulation, and grasping. Key problems to robust and precise model-based object tracking are the outliers caused by occlusion, self-occlusion, cluttered background, reflections, and complex appearance properties of the object. Two most common solutions to the above problems have been the use of robust estimators and the integration of visual cues.

It is interesting to notice that there are almost as many proposed tracking algorithms as there are applications. One reason for this is the multitude of camera-object configurations: moving camera/static object (visual servoing, visual navigation, AR), static camera/moving object (activity interpretation, surveillance), moving camera/moving object (visual servoing, AR). Another reason is the appearance of objects considered: some of the approaches have specifically been designed for tracking of textured objects [209, 197] while others, mainly based on gradient information, have mainly been evaluated on non-textured objects [35, 52, 82, 109, 212]. Despite their number, model-based tracking systems are still prone to drift and jitter, and can lose track if the geometrical model of the object is simple but the appearances of the object and background are complex. In applications such as robotic object manipulation, tracking is typically model-based, because the grasping can only be performed after aligning the manipulator and the object precisely if no additional sensory modalities are available. Thus, the absolute pose of the object with respect to the manipulator-mounted camera needs to be recovered. Some of the visual servoing applications solve this problem by using the teaching-by-showing approach that requires an image of the target in the desired pose [135].

In general, the use of only a wire-frame model for tracking is difficult when the background and the object appearance properties are complex, as it is difficult to distinguish between background and object edges, as well as multiple edges on the object itself. Tracking of textured objects can also be problematic since the “signal-to-noise ratio” is small, that is, only a fraction of detected edges really belong to the outer edges of the object.

Although there have been examples of appearance-based 3D pose tracking systems [100], most current systems for 3D pose tracking are based on tracking of object boundaries. One of the early systems called RAPID [82] uses the dynamic vision approach presented by Dickmanns and Graefe [49], which is based on the use of extended Kalman filtering to integrate image measurements through a non-linear measurement function to estimate the pose. However, this work does not consider the modeling of the motion in detail. The same applies to most of the other approaches presented, such as [130, 219]. Drummond and Cipolla presented an approach using Lie algebra formalism as the basis for representing the motion of a rigid body [52]. The approach has been shown to give good results in the case of non-textured objects.

Considering tracking based on texture, an approach for model-based tracking based on local bundle adjustment has been presented [209]. It relies on the use of a CAD model of the object and requires off-line matching of model points to their 2D projections in a set of reference key frames. Matching between the current frame and a key frame is based on homographies, which is suitable for locally planar (polyhedral) objects. An approach that also considers curved surfaces is presented in [120].

Masson et al. [138] propose a 3D tracking algorithm based on a fast patch registration that provides 3D–2D correspondences for pose estimation. An off-line stage using a textured model is used to learn a Jacobian for the patch registration. In [120], features are generated on-line and there is no need for an off-line registration process. Comport et al. [35] present a tracking system based on a non-linear pose computation formulated by means of a virtual visual servoing approach. Tracking of different features including lines, circles, cylinders, and spheres is demonstrated using the interaction matrix formulation. Robustness is obtained by integrating an M-estimator into the visual control law via an iteratively re-weighted least squares approach. Work presented in [105] demonstrates a tracking system based on integration of visual and inertial sensors. A good performance is achieved for fast camera movements due to the integration with an inertial sensor but it is argued that, in order to have a robust system, more stable visual features should be considered.

Integration of visual cues has been found to provide increased robustness and has been used successfully in application such as object tracking and scene segmentation [85, 114, 120, 166, 203]. In tracking, multiple cues have been applied mostly for image space tracking and recently they have been proposed for 3D tracking of polyhedral objects for which a textured model is available [197]. The use of a Kalman filter to integrate model-based and model-free cues was presented for objects composed of planar surfaces in [119], and different integration models for model-free cues were studied in [122].

These works have been developed further for tracking of curved surfaces where some degrees of freedom are not observable [120]. The virtual visual servoing approach has also extended to account for model-free cues [163], which is accomplished by using an image intensity-based part in the Jacobian of the virtual visual servoing. Thus, their model-free part closely resembles the 2D tracking approach of [80].

3.1.3 Object Detection, Initialization of Tracking

Initialization has the goal to select and possibly identify a target. It is in itself hardly treated as a research topic and commonly relies on object recognition methods or constraints on tracking. As an example, most approaches of visual servoing exploit an initial constraint. Such simplifying constraints are light objects on dark background, LEDs, black and white markers, colored objects, objects with surfaces of different grey values, given correspondences, restriction to a ground plane or manually selected features. Obviously, these constraints limit applicability and do not allow recovery after loss of tracking. Another example are static tasks, where initialization is relatively simple and uses techniques such as image subtraction or optical flow calculations. Both techniques highlight areas in the image where a motion has been detected. If a CAD-model of the target is available, an initial pose estimate can be used to project the features into the image, e.g., [49, 114, 209, 215].

A classical technique of initialization is object recognition. A common approach is to match image features in the initial image to features in a database made for single or multiple objects. The match reports either object hypotheses or hypothesis of a specific view, which are subsequently verified to report the most likely object [75, 216, 221]. Thus, as a byproduct of the recognition process, an estimate of the object pose may be provided. For a comparison of different feature detectors we refer to [144].

It should be noted that initialization differs from recognition in that one specific object has to be found and located, practically reversing the classical recognition process of identifying all objects in the scene. Impressive recognition results have been reported, which are discussed in detail in Section 3.4. Nevertheless, object recognition suffers from two common problems:

(1) Methods today rely mostly on sets of features. If these features cannot be detected reliably, recognition rates decrease rapidly. The difficulty is that recognition as part of robot vision requires the ability to detect objects under varying viewing angle, significant changes in scale and illumination conditions. Although great improvements have been achieved over recent years, robotic tasks remain difficult.
(2) Methods are mostly designed for databases, where objects are centered in the image. Hence a first selection of targets has been achieved by the person that took the image, while a robot would need to search and take these images first. This difficulty adds to the problem discussed above.

For example, methods require good features [99, 216] or a perfect segmentation [47]. The latter methods show sensitivity to changing background or lighting (reported are objects on dark or pasted background) [17]. Probabilistic handling of the image templates can improve the sensitivity but reduces the likelihood of successful recognition [110]. Grouping requires good feature extraction, which is usually assumed via manual selection or images with special objects [47]. The search tree can be reduced by using attributes of the object model such as surface characteristics but still requires high processing power [26]. Invariant features (invariant to specific perspective distortions [221] or to illumination [2]) claim robustness, however, perfect segmentation of the outline is assumed, an equally difficult problem. A promising approach is to use several cues and many local features that are statistically grouped to indicate object existence [132].

The integration of recognition methods into visual servoing systems has not been achieved, since methods are complex and not reliable. Two exceptions are [53] and [157]. Eberst et al. [53] regularly invoke a recognition scheme for re-initialization. The idea is to exploit model knowledge. In most approaches this is done purely for the tracking step.

Another typical approach to improve initialization over the methods using simplifying constraints is to enhance tracking or segmentation methods to enable the initialization.

In summary, the initialization of tracking is most of the time achieved by using simplifying constraints. Promising roads of work are fast object recognition approaches, methods to reliably extract features using cue integration and the goal directed use of modeling knowledge. Open problems in object detection relate to extraction of shape and structure of objects and relating them to known objects (see Section 4.1).

3.2 Visual Servoing–Arms and Platforms

The continuous control of a mechanism using visual input is referred to as Visual Servoing [92]. It means to control the pose of a mechanism in a closed loop (e.g., the gaze direction of the end-effector) using the input of a machine vision system. Sometimes the term vision-based control of motion is used. Thus, apart from image processing and computer vision, visual servoing also requires techniques from control theory. Hence, visual servoing consists of two intertwined processes: tracking and control. In addition, the system may also require an automatic initialization procedure which may include figure-ground segmentation and object recognition, as outlined above. In the robotics community, visual servoing has been used to control the movement of robotics arms as well as mobile robots. In terms of camera configurations and their number, there are examples of both single and multiple cameras, that are either fixed in the workspace or are attached to the robot. Cameras fixed in the workspace may additionally be attached to a pan-tilt unit or another robot.

The aim of visual servoing is twofold. On the one hand, visual servoing makes it possible to follow arbitrary object motions. On the other hand, it becomes possible to control the motion toward an arbitrary object location when seeing the direct relation between robot end-effector and the object. Hence, visual servoing eliminates the requirement to calibrate the camera-robot system. It can be shown that two cameras using standard calibration parameters are sufficient for accurate robot control [60, 78]. This is achieved by either seeing the robot and the object within the images or by mounting the camera(s) directly on the end-effector.

3.2.1 Visual Servoing Control Loop

Aspects relevant to all tracking and visual servoing approaches are the control theoretical considerations to obtain good dynamic performance. The goal is to consider the entire system including visual sensing, the controller(s), the mechanism, and all interfaces. Visual servoing is different from conventional robot control in the respect that visual sensing imprints specific properties on the control loop. The most significant property is the delay or latency of the feedback generated by the vision system, a problem encountered with any use of cameras and computer vision methods. The basic control loop is depicted in Figure 3.2. It contains three major blocks: the vision system, the controller and the Mechanism or robot. The vision system determines the error between the command location and the present location of the target. First, the result is expressed as an error in the image plane. The controller converts the signal to a pose or directly into command values for the axes of the mechanism and transfers the values to the robot. The robot or vehicle commonly uses a separate controller to control the motors at the axes level.

The structure of the loop in Figure 3.2 derives from the fact that the target motion is not directly measurable. Therefore, the target motion is treated as a non-measurable disturbance input [39].

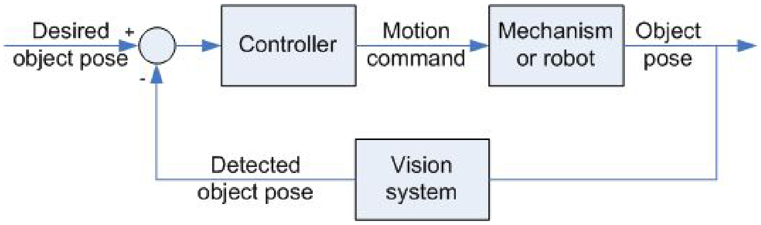

Fig. 3.2 A basic block diagram of visual servoing. The vision system may give object features as output or directly the object pose for use in image-based resp. position-based visual servoing. The same basic loop applies to mobile robots, manipulators, or other mechanisms.

The objective of a tracking system here is to keep the target in the field of view at all times. If the camera is fixed, the application needs to assure this is sufficient, e.g., surveillance of a fixed area. Mounting the camera on an active head or robot increases the viewing range. Certainly, even with very wide field of view the limiting factor is that the target moves out of this field. Hence it is useful to think how tracking of the highest possible target velocity (or acceleration) can be achieved. The analysis of the tracking cycle indicates how to build a tracking system [179, 211]. The two main factors to take care of are (1) the latency or delays in one cycle from obtaining the image and (2) the part or window of the image that is actually processed.

While it seems intuitive that latencies delay tracking, the second factor, window size, is often not discussed much in the literature. If the full image is processed, this may take much longer than the frame time of the camera. If a smaller window is used, for example, around the location where the target has been seen in the last image, it is possible to exploit every image. The optimum is reached when the window size is selected such that processing is as fast as acquiring images [211]. Further guidelines to enable tracking of fast moving targets are:

- Latency is the dominating factor. Hence, constant times (e.g., image transfer, control algorithm) linearly reduce the performance and should be minimized.
- The use of high-speed cameras can highly increase performance. The rationale is that high-speed cameras reduce the sampling time of the vision system and the sampling time of the overall system. As a consequence, the development of faster cameras highly improves tracking of fast-moving targets even when using the same computer hardware.
- A space-variant image tessellation [13] further increases tracking performance. This is gained at the loss of imaging resolution.

It is interesting to note that the human eye exhibits space-variant tessellation with the high-resolution fovea in the center and a wide field of view at logarithmically decreasingly lower resolution. It should be also noted that particle filter approaches subsample the image to obtain space-variance though at quasi-constant detection or recognition resolution, e.g., [93].

The next section discusses the main approaches used in visual servoing.

3.2.2 Visual Servoing Approaches

Two classical visual servoing approaches are known as *image*- and *position*-based visual servoing, Sanderson and Weiss [172] denoted IBVS and PBVS, respectively. The former approach bases the control on the estimation of 2D image measurements while the latter relies on the 3D reconstructed estimates of image measurements. For a tutorial on visual servoing see [92] and for a recent review [30]. Both image- and position-based servoing minimize an error between the current and desired positions of visual features. In IBVS features are represented by their 2D image coordinates and *image Jacobian*, also called the *interaction matrix*, [60] is used to relate the spatial velocity of the camera to the relative change in features' positions.

To cope with the problems inherit to position- and image-based visual servoing, several hybrid approaches have been proposed. The method known as 2.5D visual servoing, Malis et al. [135] decouple the translation and rotational degrees of freedom in order to provide better stability conditions. The control is based on image coordinates of a point and the logarithm of its depth which are computed from a partial pose estimation algorithm. Other examples adopt a partitioning approach where one uses image features each of which is related to a different degree of freedom to be controlled [38].

There are approaches that concentrate on problems related to mobile platforms, referred to also as *homing* [171, 205, 158]. Sometimes the problem of homing is solved by using the fundamental matrix, but this approach is ill conditioned in case of planar scenes, which occur frequently in natural environments. In addition, it is common to compare small baseline images with high disparity due to rotation, where the fundamental matrix also gives bad results. Sagüés and Guerrero [171] proposed to use a monocular vision system and compute motion through a homography obtained from matched lines. Finally, a 2D homography is proposed to correct motion directly.

Piazzi and Prattichizzo [158] present a visual servoing method for holonomic robots based on the auto-epipolar property which does not need calibration of the camera neither computation of the fundamental matrix. The presented algorithm consist of three steps which correct the rotation, lateral, and depth error, respectively. A method that extends the auto-epipolar visual servoing method for non-holonomic mobile robots has been presented in [136, 129]. Both of the approaches result in a three-step motion for regulating the rotational and translational errors. In the former case, the robot sometimes needs to move backward which is a drawback when no sensors on the rear part of the robot are available. The latter method builds upon that work and deals with that problem.

Another way to solve the problem of the non-holonomic constraints is presented in [205]. This work is also concerned with bringing the robot to a desired pose considering a non-holonomic mobile robot. The control loop uses visual data from a hand–eye system. In particular, a controller is designed by using the extra degrees of freedom provided by a hand–eye system.

Specific problems such as the effect of camera calibration errors have been studied in [59]. The convergence properties of the control part of the systems are known for most cases as discussed in [29, 134]. While the convergence of the system is an essential performance property, it does not reveal much about the generated robot trajectory and its uncertainty.

The error characteristics of visual servoing are usually investigated from the stability of the closed-loop system or the steady-state error [46]. It is known that the convergence of position-based visual servoing (PBVS) might be inhibited by the loss of stability in pose estimation [29]. 2.5D servoing does not seem to suffer from this problem [134], unless the partial pose estimation becomes unstable. Deng et al. [46] have proposed use of the steady-state error as a measure of sensitivity of visual servoing. However, if long trajectories are executed, it is important to estimate the sensitivity of the system along the trajectory to, for example, predict the set of adequate trajectories in the presence of errors. Another approach is to consider the outliers in the image data. Comport et al. [36] have proposed a scheme to increase the robustness by embedding the outlier processing into the control law. Kyrki et al. [121] address the issue of measurement errors in visual servoing. The error characteristics of the vision-based state estimation and the associated uncertainty of the control are investigated. The major contribution is the analysis of the propagation of image error through pose estimation and visual servoing control law. An example of high-speed visual servoing has been demonstrated in [148]. An important application of using vision-based control in medical application has been reported in [117].

3.3 Reconstruction, Localization, Navigation, and Visual SLAM

It is widely recognized that a mobile robot needs the ability to build maps of the environment using natural landmarks and to use them for localization [27, 51, 195, 199, 200]. Solving the SLAM problem with vision as the only external sensor is now the goal of much of the effort in the area [44, 68, 72, 97, 151, 185]. Monocular vision is especially interesting as it offers a highly affordable solution in terms of hardware. We adopt the term vSLAM [101] for visual SLAM. Currently, vSLAM solutions focus on accurate localization, mostly based on the estimation of geometric visual features and a partial reconstruction of the environment. Thus, the resulting map is useful for the localization of the robot, but its use for other purposes is often neglected. This section concentrates on the geometric mapping while in the next section, we take a look at approaches that apply visual means for higher-level understanding of the environment. An example of data-flow in a SLAM system is shown in Figure 3.3.

Single camera SLAM is an instance of bearing only SLAM. Each image in itself does not contain enough information to determine the location of a specific landmark. Solving for the location requires that images from multiple viewpoints are combined. This approach is similar to what in the computer vision society if referred to as the Structure-from-Motion problem (SfM). The essential problem of simultaneously

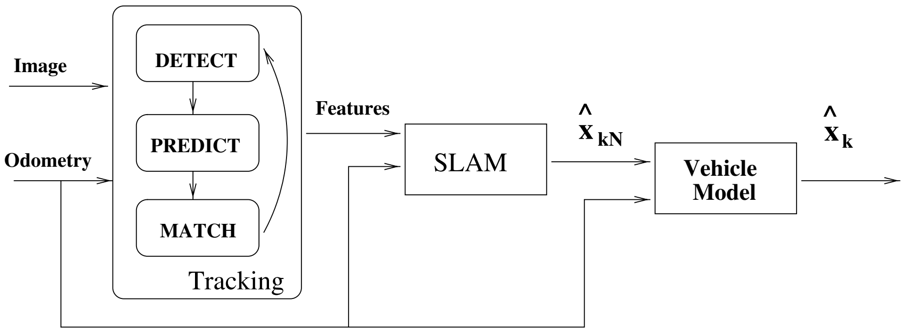

Fig. 3.3 The flow of data in the system. The image and odometry inputs are processed in the tracking module where matches are found between consecutive frames. The output is delayed N frames to the SLAM module. If an estimate of the current robot pose it desired one can be calculated by predicting forward the pose from the SLAM module using odometry or other dead-reckoning sensors [98].

estimating the structure of the environment and the motion of the observer is identical. In the computer vision community, SfM is nowadays considered mostly a solved problem, as commercial solutions for SfM-based camera motion estimation have become available from companies such as 2d3[1] The state-of-the-art SfM solutions are mostly based on using projective geometry as the geometrical model and bundle adjustment techniques (basically Levenberg–Marquardt minimization) for finding the maximum likelihood solution for the non-linear optimization problem.

The major difference is that the SfM methods are commonly run off-line and consider batch processing of all the images acquired in the sequence while SLAM requires incremental and computationally tractable approaches suitable for on-line and real-time processing. Furthermore, the SfM methods do not assume feedback from information sources such as odometry that are commonly used in SLAM. The fact that a landmark cannot be initialized from a single frame means that a solution to bearing only SLAM must explicitly address this problem. Different solutions for initial state estimation in bearing only SLAM have been proposed. A combination of bundle adjustment, commonly used in regular Structure-from-Motion approaches, and Kalman filter

[1] See `http://www.2d3.com`.

has been proposed in [45]. It has been shown that even if the method is less optimal than a regular Kalman filter approach, it gives better reconstruction results.

The most important problem that has to be addressed in bearing only SLAM is landmark initialization, because a single observation does not allow all degrees of freedom to be determined, as mentioned above. A particle filter used to represent the unknown initial depth of features has been proposed in [44]. The drawback of the approach is that the initial distribution of particles has to cover all possible depth values for a landmark which makes it difficult to use when the number of detected features is large. A similar approach has been presented in [118] where the initial state is approximated using a Gaussian Sum Filter for which the computational load grows exponentially with the number of landmarks. The work in [126] proposes an approximation with additive growth.

Several authors have demonstrated the use of multiple view approach in monocular SLAM [72, 97, 98]. These works demonstrate the difficulties related to landmark reconstruction when the robot performs only translational motion along the optical axis. To cope with the reconstruction problem, a stereo-based SLAM method was presented in [185] where Difference-of-Gaussians (DoG) is used to detect distinctive features which are then matched using SIFT descriptors. One of the important issues mentioned is that their particle filter-based approach is inappropriate for large-scale and textured environments.

One of the challenging problems in SLAM is loop closing. In [76] a portion of the map of laser scans near the current robot pose is correlated with older parts of the map every few scans to detect loops. In [151] visually salient so-called “maximally stable extremal regions” or MSERs are encoded using SIFT descriptors. Images are taken every few meters or seconds and compared to a database to detect loop closing events. As we will see later our framework also allows us to detected loop closing situations in an effective way. Another example of loop closing is demonstrated in [98]. Examples from this can be seen in Figure 3.4 that shows the situation as the robot is just closing the loop for the first time by re-observing one of the earliest detected landmarks. The two lines protruding from the robot show the bearing vectors defined

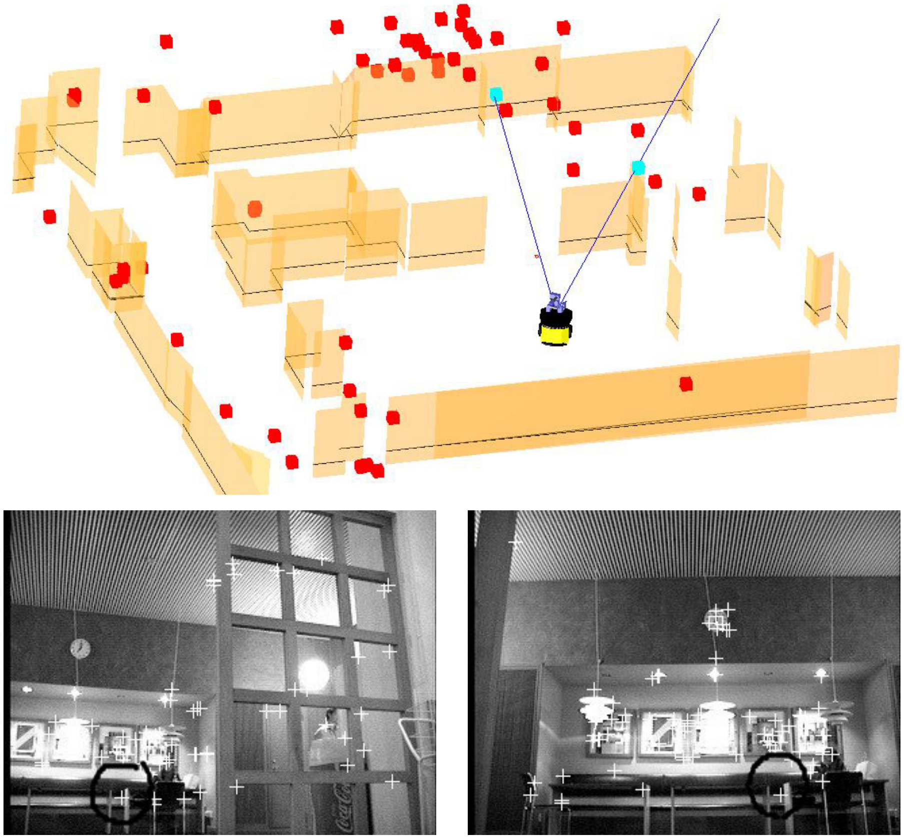

Fig. 3.4 The situation when the robot is closing the loop for the first time by re-observing a feature. The observed features are marked in cyan (light) in the upper part. The matched pair of features is circled in the lower two images.

by the observations. It is the landmark furthest away from the robot, toward the wall in the back, which is re-observed. The two images in Figure 3.4 show the image from the first time it was detected (right) and the image at which loop closing takes place (left). The landmark in question is marked with a circle in the images.

3.4 Object Recognition

For most of the tasks a robot needs to perform, it must be able to determine *what* things there are in the environment and *where* they are. Determining *what* requires object recognition which is far from trivial to solve. Object recognition is one of the main research topics in the field of computer vision. For a comprehensive review see [161].

There are several aspects of object recognition:

- Detection versus recognition
  Detection is different to recognition in that the target object is given and needs to be found in the image, while in recognition an image is given and the task is to identify the object(s). Object detection methods have been summarized in Section 3.1.3 above.
- Generic (categorization, see Section 4.2) versus specific object recognition
  Object recognition algorithms are typically designed to classify objects to one of several predefined classes assuming that the segmentation of the object has already been performed. In robotic applications, there is often a need for a system that can locate objects in the environment. This means that the distance to the object and thus its size in the image can vary significantly. Therefore, the robot has to be able to detect objects even when they occupy a very small part of the image. This requires a method that evaluates different parts of the image when searching for an object.
- Global versus local
  In general, object recognition systems can roughly be divided into two major groups: global and local methods. Global methods capture the appearance of an object and often represent the object with a histogram over certain features extracted during the training process, e.g., a color histogram represents the distribution of object colors. In contrast, the local methods capture specific local details of objects such as small texture patches or particular features. For the robot to recognize an object, the object must appear large enough in the camera image. If the object is too small, local features cannot be extracted from it. Global appearance-based methods also fail, since the size of the object is small in

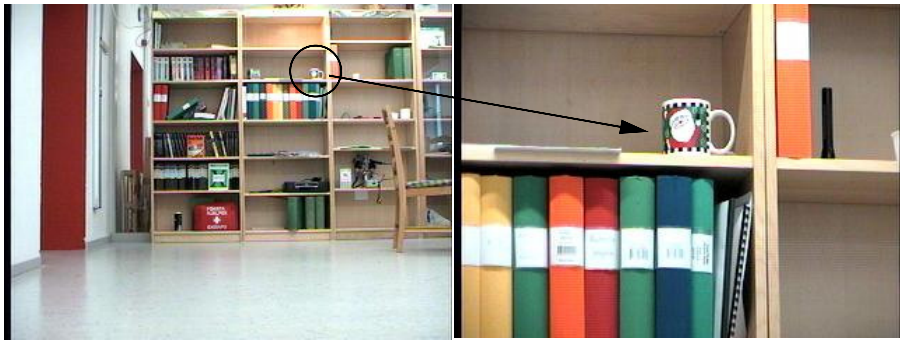

Fig. 3.5 Left: The robot cannot recognize the cup located in the bookshelf. Right: Minimum size of the cup required for robust recognition.

relation to the background which commonly results in high number of false positives. As shown in Figure 3.5, if the object is too far away from the camera (left), no adequate local information can be extracted.

- 2D versus 3D
  Despite the large body of work on vision-based object recognition, few have investigated strategies for object recognition when the distance to the object (scale) changes significantly. Similarly, there are very few object recognition systems that have been evaluated in a mobile robot setting. In [74], a mobile, self-localizing robot wanders in an office environment and can learn and recognize objects encountered. The idea of generating hypotheses and then zooming on them in the verification step to provide richer information has been used [103]. The authors use regular color histograms that only work for relatively simple objects and requires many training images. The problem studied in [58] is a mobile robot that autonomously navigates in a domestic environment, builds a map as it moves along and localizes its position in it. In addition, the robot detects predefined objects, estimates their position in the environment, and integrates this with the localization module to automatically put the objects in the generated map.

3.4.1 Specific Object Recognition

Visual recognition implies matching features or visual cues derived from an image to stored representations of objects or images. This is a classical problem that has been studied extensively in the area of image processing and pattern recognition. Techniques used rely either on the extraction of (i) distinctive features, (ii) distinctive regions, (iii) combination of features and regions, or (iv) patches and local histograms. Good examples are found in [21, 65, 66, 86, 127, 131, 186]. With the increased trend of large data sets, new approaches of how matching of these features is performed have been proposed. Some of the most prominent examples are based on different types of vocabularies [187].

As a relation, modeling only a specific brain area (V1) of the primate or human vision cortex seems not to be sufficient for object recognition under more than regular views [160]. This nicely shows why the present approaches are good on a database, where a human took images of objects. But this means the human solved most of the problem of bringing the object into the image and at reasonable size and viewpoint.

As a conclusion one might state that the above methods present good solutions to recognize a large number of objects. The computer vision community moved on to tackle object categorization, also see Section 4.2 below. However, in the context of robotics vision, the task of specific object recognition under changing illumination and in clutter is not solved yet, as for example the Semantic Robot Vision Challenge will demonstrate in Section 4.4.

3.4.2 Object Recognition from Range Images

One of the difficulties of 2D images is that depth information is not available directly but needs to be inferred from the appearance of objects. Because shape is not directly encoded, this problem is in general difficult or ill-posed [12]. However, recent progress in invariant feature extraction is the basis to obtain first good results in realistic settings [113, 132]. Although these approaches are rather fast, they do not work satisfactorily in cluttered scenes and inherit the major problem of intensity-based systems, that is, dependency on lighting conditions [102].

One way to surpass this problem is to obtain the 3D shape of objects from range images. In robot vision range images can be obtained through various methods ranging from laser scanning over structured light approaches to stereo. Stereo vision follows human vision and obtains depth from focusing both eyes on the target object. For a recent review see [25].

An important question in computer vision is how to model or represent the object for detection in depth data. One assumption is that humans represent shape by its parts [181]. Human vision parses shapes into component parts, and it organizes them using these parts and their spatial relationships. From a computational perspective, parts are useful for many reasons. First, many objects are articulated: A part-based description allows one to decouple the shapes of the parts from the spatial relationships of the parts–hence providing a natural way to present and recognize articulated objects. Second, one never sees the whole object in one view: the rear side of an object is not visible due to self-occlusions, and its front may be occluded by other objects. Representing shapes by parts allows the recognition process to proceed with those parts that are visible.

One theoretical approach defining parts is to postulate that human vision uses general computational rules, based on the intrinsic geometry of shapes, to parse visual objects [181]. A visual system decomposes a shape into a hierarchy of parts. Parts are not chosen arbitrarily. When two arbitrarily shaped surfaces are made to interpenetrate, they always meet at a contour of concave discontinuity of their tangent planes [14, 88]. This helps the segmentation task, which is also for range images still unsolved in general [89]. Multiple models have been introduced that are suited to describe parsed objects according to the rule of transversality.

One way to review the representation methods is with respect to the number of parameters they use to describe the 3D shape. In the past decade much work has been made describing range data with geometric primitives (sphere, cylinder, cone, torus) except the cube. This can be easily explained, because rotational symmetric primitives can be described with an implicit closed form, while the cube model has to be composed of six planes. More complex descriptions explained below provide this capability. Generalized cylinders are the dedicated part-level models and form a volume by sweeping a two-dimensional contour along an arbitrary space curve. The contour may vary along the curve (axis). Therefore, definitions of the axis and the sweeping set are required to define a generalized cylinder. An often cited early vision system which applied generalized cylinders is the ACRONYM system to detect air planes [24]. A difficulty is the complicated parameterization of generalized cylinders, and the lack of a fitting function that would provide a direct evaluation criteria on how well the model generalized cylinder fits the image data. In other early systems such as 3DPO a CAD model has been used to define a sequence of edge feature to locate objects [20].

Superquadrics are perhaps the most popular approach due to several reasons. The compact shape can be described with a small set of parameters ending up in a large variety of different basic shapes. Solina and Bajcsy [189] pioneered work in recovering single Superquadrics with global deformations in a single-viewpoint cloud and demonstrated that the recovery of a Superquadric from range data is sensitive to noise and outliers, in particular from single views. Jaklic et al. [96] summarize the recover and select paradigm for segmenting a scene with simple geometric objects without occlusions. This method aims at a full search with an open processing time incompatible to most applications such as robotics. Recently Krivic and Solina [115] show the recovery of a known complex object in a scene using the connectivity information of the Superquadrics, handling the scene occlusions by using the redundancy information of the part connections. Biegelbauer et al. [16] detect known objects using a probabilistic approach sampling small patches before fitting the full model.

In summary, the recovery of superquadrics has been largely investigated. They are also useful to describe geons [14], which are a set of 36 basis geometric shapes proposed to be sufficient to describe all object parts. An open problem is to handle sparse data due to one-view scans of the scene and occlusions in cluttered scenes. Using closed-form models has proven useful because they impose a part symmetry, which can then be used in robotic tasks such as grasping. Representations with more parameters suffer from the necessity to fit many parameters and can only be applied for range scans of objects from all sides. Katsoulas proposed a novel object detection approach searching for box-like objects using parabolically deformable Superquadrics for taking bags from a pallet [102]. He weakened the bottleneck of the scene segmentation using a 3D edge detector and achieved some improvement in processing time, but this method cannot handle non box-like objects and scene occlusions.

Robotics takes up these results and exploits them to find objects on tables, e.g., from scans over the table [198]. The range image is obtained from scanning the table with a laser and camera triangulation set-up. A more sophisticated sensor is the DLR sensor head, which combines laser scan, laser range sensor, and stereo cameras in one head-like configuration that can be mounted on a robot end-effector. However, the combination of sensors comes at high costs [193]. Using this principle of scanning the table, it is possible to obtain one view of the table scene, segment objects, obtain their 3D shape and estimate potential points for grasping. Figure 3.6 gives an example of such an approach.

In an industrial setting, the use of laser systems has also been demonstrated. Biegelbauer et al. [15] report how a part geometry can be obtained for the purpose of painting. The idea is that, given a specific

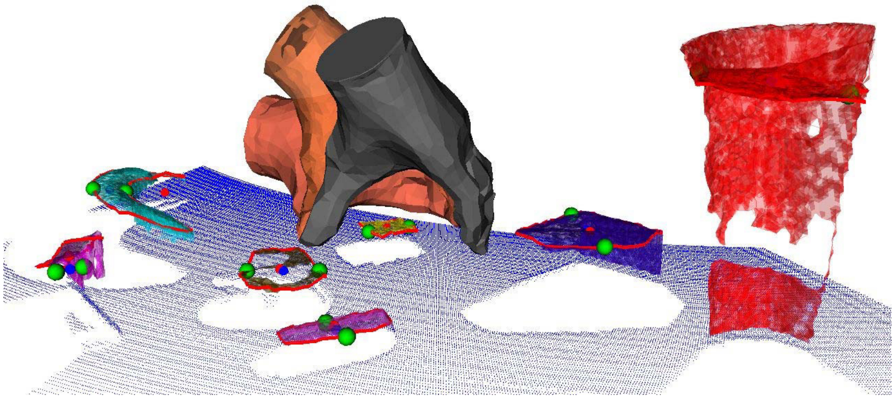

Fig. 3.6 Range image of table scene acquired by a laser scanner and the objects detected including potential grasp points.

geometric shape, a painting primitive can be selected and applied onto the part surface. Example shapes distinguished are free-form surface, ribs, or parallel structures where painting needs to be along the structure, and cavities which require a specific paint procedure into the hole. Hand held sensors are also feasible in industry, e.g., to scan parts for accurately locating the bore holes [16]. These examples indicate that laser-based approaches start to penetrate industrial settings.

3.5 Action Recognition, Detecting, and Tracking Humans

There is strong neurobiological evidence that human actions and activities are directly connected to the motor control of the human body [168]. When viewing other agents performing an action, the human visual system seems to relate the visual input to a sequence of motor primitives. The neurobiological representation for visually perceived, learned, and recognized actions appears to be the same as the one used to drive the motor control of the body. These findings have gained considerable attention from the robotics community [43, 174] where the goal of *imitation learning* is to develop robot systems that are able to relate perceived actions of another (human) agent to its own embodiment in order to learn and later to recognize and to perform the demonstrated actions. Here, action representations based on detailed human body models are usually applied.

In robotics as well as in vision, the neurobiological findings motivate research to identify a set of action primitives that allow (i) representation of the visually perceived action and (ii) motor control for imitation. In addition, this gives rise to the idea of interpreting and recognizing activities in a video scene through a hierarchy of primitives, simple actions, and activities. Many researchers in vision and robotics attempt to learn the action or motor primitives by defining a “suitable” representation and then learning the primitives from demonstrations. The representations used to describe the primitives vary a lot across the literature and are subject to ongoing research.

As an example, for imitation learning a teacher may attempt to show a robot how to setup or clean a dinner table. An important aspect is that the setting of the environment may change between the demonstration and the execution time. A robot that has to setup a dinner table may have to plan the order of handling plates, cutlery, and glasses in a different way than originally demonstrated by the human teacher. Hence, it is usually not sufficient to just replicate the human movements. Instead, the robot must have the ability to recognize what parts of the whole task can be segmented and considered as subtasks so that it can perform on-line planning for task execution given the current state of the environment.

The robotics community has recognized that the acquisition of new behaviors can be realized by observing and generalizing the behaviors of other agents and it is thus mainly concerned with generative models of actions. The combination of generative models and action recognition leads to robots that can imitate the behavior of other individuals [50, 57, 174].

Hence, the interest of roboticist is to enable robots with action recognition capabilities, both if these actions are performed by humans or other robots. In some cases, the action recognition is used for pure recognition purposes in context understanding or interaction. Consequently, different discriminative approaches are commonly adopted here. However, recent developments in the field of humanoid robots have motivated the use and investigation of generative approaches with the particular application of making robots move and execute their action in a *human-like* way, thus raising interest in integrated action recognition and action generation approaches.

For a robot that has to perform tasks in a human environment, it is also necessary to be able to learn about objects and object categories. It has been recognized recently that grounding in the embodiment of a robot, as-well as continuous learning, is required to facilitate learning of objects and object categories [67, 192]. The idea is that robots will not be able to form useful categories or object representations by only being a passive observer of its environment. Rather a robot should, like a human infant, learn about objects by interacting with them, forming representations of the objects and their categories that are grounded in its embodiment. Most of the work on robotic grasping has been dealing with analytical methods where the shape of the objects being grasped is known *a priori*. This problem is important and difficult mainly because of the high number of DOFs involved in grasping arbitrary objects with complex hands.

One of the most basic interactions that can occur between a robot and an object is for the robot to push the object, i.e., to simply make a physical contact. Already at this stage, the robot should be able to form two categories: physical and non-physical objects, where a physical object is categorized by the fact that interaction forces occur. A higher level interaction between the robot and an object would exist if the robot was able to *grasp* the object. In this case, the robot would gain actual physical control over the object and having the possibility to perform controlled actions on it, such as examining it from other angles, weighing it, and placing it. Information obtained during this interaction can then be used to update the robots representations about objects and the world. Furthermore, the successfully performed grasps can be used as ground truth for future grasp refinement [67].

Vision-based recognition of a hand grasping or manipulating an object is a difficult problem, due to the self-occlusion of the fingers as well as the occlusion of the hand by the grasped object. To simplify the problem, some approaches use optical markers, but markers make the system less usable when service robot applications are considered.

Approaches to grasp recognition [28, 152] first reconstruct the hand in 3D, from infrared images [152] or from an optical motion capture system which gives 3D marker positions [28]. Features from the 3D pose are then used for classification. The work of Ogawara et al. [152] views the grasp recognition problem as a problem of shape reconstruction. The more recent work of Chang et al. [28] learns a non-redundant representation of pose from all 3D marker positions — a subset of features — using linear regression and supervised selection combined. Some of the most recent approaches strive to develop a markerless grasp recognition system [104], also depicted in Figure 3.7.

3.6 Search and Attention

In all the above there is the inherent assumption that the robot or vision system has its view at relevant things to start with. In cases of navigating through an environment it is a fair assumption that looking

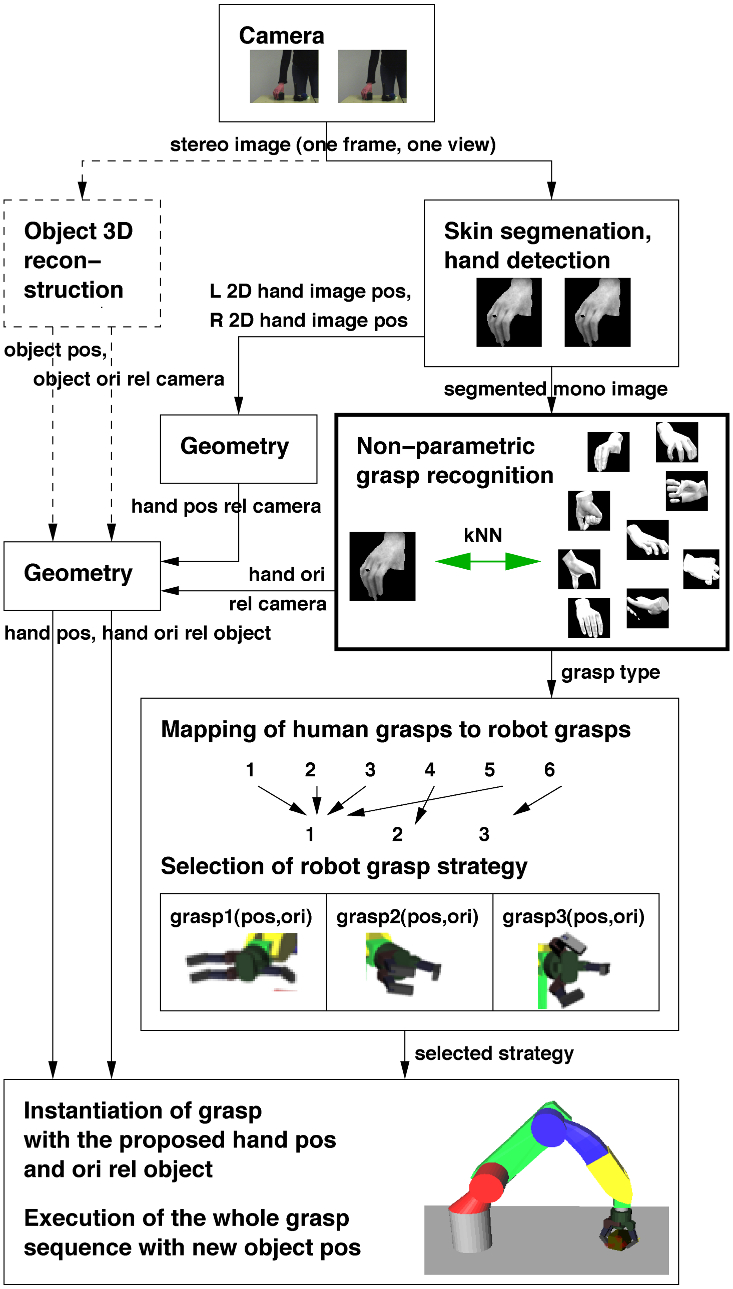
Camera
stereo image (one frame, one view)
Object 3D recon–struction
Skin segmenation, hand detection
L 2D hand image pos,
R 2D hand image pos
object pos,
object ori rel camera
segmented mono image
Geometry
Non–parametric grasp recognition
kNN
hand pos rel camera
Geometry
hand ori
rel camera
hand pos, hand ori rel object
grasp type
Mapping of human grasps to robot grasps
1 2 3 4 5 6
1 2 3
Selection of robot grasp strategy
grasp1(pos,ori)
grasp2(pos,ori)
grasp3(pos,ori)
selected strategy
Instantiation of grasp with the proposed hand pos and ori rel object
Execution of the whole grasp sequence with new object pos

Fig. 3.7 An example of the grasp recognition system from [104].

in the direction of motion is sufficient. When using vision in tracking applications, the detection of the target is assumed to resolve the problem. However, this also means that the target first needs to move into the field of view of the camera. Note, compared to the more than 180 degrees of the human visual field the typical cameras have a field of view of 60 to sometimes 100 degrees. Hence there is the need to first search or look around when investigating an environment, and second to detect and attend to the relevant objects related to the task of the robot system.

For robotic applications, attention can be seen as a selection mechanism serving the higher level tasks such as object recognition or map building. Human studies may provide an insight into the process of attention. Some of the studies show that humans tend to do a subconscious ranking of the "interestingness" of the different components of a visual scene. This ranking depends on the observers goals as well as the components of the scene, how the components in the scene relate to their surroundings (bottom-up) and to the task (top-down) [94, 128]. In humans, the attended region is selected through dynamic modifications of cortical connectivity or through the establishment of specific temporal patterns of activity, under both top-down (task dependent) and bottom-up (scene dependent) controls [154].

Current models of how the attentional mechanism is incorporated in the human visual system generally assume a bottom-up, fast, and primitive mechanism that biases the observer toward selecting stimuli based on their saliency which is encoded in terms of center-surround mechanisms. Then, there is a slower, top-down mechanism with variable selection criteria, which directs the "spotlight of attention" under cognitive, volitional control [202]. In computer vision, attentive processing for scene analysis initially dealt mostly with salience-based models, following [202] and the influential model of [107]. However, several computational approaches to selective attentive processing that combine top-down and bottom-up influences have also been presented in recent years.

In the example of [108], attention and search processes are intertwined where the saliency-based search model is emerges from the stochastic shifts in attention. Choi et al. [32] suggest learning the desired modulations of the saliency map, based on the Itti et al.'s model [95], for top-down tuning of attention, with the aid of an ART-network. Navalpakkam and Itti [149] enhance the bottom-up salience model to yield a simple, yet powerful architecture to learn target objects from training images containing targets in diverse, complex backgrounds. Lee et al. [124] showed that an Interactive Spiking Neural Network can be used to bias the bottom-up processing in a face detection task. In the VOCUS-model [69] there are two versions of the saliency map: a top-down map and a bottom-up one. The bottom-up map is similar to that of [95], while the top-down map is a tuned version of the bottom-up one. The total saliency map is a linear combination of the two maps using a fixed user provided weight. This makes the combination rigid and non-flexible, which may result in loss of important bottom-up information. Oliva et al. [153] show that top-down information from visual context can modulate the saliency of image regions during the task of object detection. Their model learns the relationship between context features and the location of the target during past experience in order to select interesting regions of the image.

One shortcoming of most of these computational models is that they are usually limited to the study of attention itself, and besides some works on the use of attention for object recognition, it has never been studied in an "active vision" perspective such as a service robotic context. One of the few recent works that does in fact incorporate a computational mechanism for attention into a humanoid or mobile platform is the work presented in [113, 146].

4

Open Challenges

For the future of robotics and artificial cognitive systems, representations in general play a major role. A robot's local world is built by objects that are thought to be recognized, classified, interpreted, or manipulated. Though also *things*, as untreated basic sensory features, might help for some of these tasks, the semantic representation of an *object* seems to be more intuitive. Nevertheless, the question arises what makes an object an object, what makes John's cup being John's cup? There has been plenty of research on this issue, most of which concentrates on example-based recognition of objects by learned features, may they be visual or shape-based. In such systems, John's cup has been shown to the robot and can thus be recognized again. However, this does not make the robot identify arbitrary cups it has never seen before. Section 4.1 discusses work toward detecting shape and structure to model objects and their function and Section 4.2 will review work toward object categorization.

One of the major requirements of a cognitive robot is to continuously acquire perceptual information to successfully execute mobility and manipulation tasks [54, 150, 194]. The most effective way of performing this is if it occurs in the context of a specific task. This was, for a long time, and still is the major way of thinking in the field of robotics as outlined in Section 4.3. Focus is usually put on the on task-specific aspects when processing sensor data which may reduce the overall computational cost as well as add to the system robustness. However, in most cases this leads to the development of special-purpose systems that are neither scalable nor flexible. Thus, even if significant progress has been achieved, from the view of developing general system able to perform various tasks in domestic environments, research on autonomous manipulation is still in its embryonic stage. Evaluating parts of a robot vision system (as in Figure 1.1) will be discussed in Section 4.4.

4.1 Shape and Structure for Object Detection

Humans are astoundingly apt at detecting objects even in cases of objects that they have never seen before. First, detection is closely linked to figure-ground segmentation [55, 159]. Second, we can rapidly classify the object based on typical properties that are often linked to the shape and structure of the object. Furthermore, robots need the shape of objects to grasp them or the shape of the environment to navigate in it. In computer vision this is summarized under "shape from X" methods. Obviously the easiest way to obtain shape is to use range images (see also Section 3.4.2). Another option to use the reconstructions from SLAM (Section 3.3). We summarize the results regarding shape perception below and then proceed to discuss other options such as stereo or grouping of features. One should keep in mind that the goal is to relate shape to affordances relevant for the robotics task.

Recognition methods using range images have been discussed in Section 3.4.2. The review shows that given good laser scans, a relatively large class of objects can be recognized. The difficulty for robotics is to obtain the good data in real-time. Recognition approaches in computer vision exploit rather excellent data from expensive scanners, objects being placed on rotating tables to acquire several views, e.g., [142]. The laser sensors used in navigation either only work in one specific plane or need to scan by using an additional pan-tilt unit. This is mechanically expensive and allows only sequential scanning. For robotics, to link the structure with object functions as required in robotics, full scans of scenes are needed [194]. Today, detection is possible only for full 3D scans in little clutter [156]. To scan over surfaces to obtain object shape such as in [193, 102] and Figure 3.6 is more appropriate for robotic grasping, but sensor systems are still bulky and expensive.

Stereo is cheap and can also produce a range image. However, it requires objects to have texture. The shape of objects and a simple 3D model of the scene can then be obtained from the stereo point cloud, e.g., [125]. Accuracy of data is still, however, not sufficient for a more detailed analysis of scenes or the detection of smaller objects.

The approaches discussed in Section 3.3 also build up a 3D point cloud representation of the space for a mobile robot from monocular or stereo cameras. The very approach can also be used to build up a 3D representation of a table top and objects placed on it. There are examples that rely on matching different types of image features across views. For example, [106] track short edge segments and correct for motion blur, which helps to reconstruct not only points on the object surface but also the outer contours of objects. Another example is to combine the point data of SLAM approaches into higher level features. Gee et al. [70] estimate on-line dominant planes from the points used in SLAM. This seems a useful way to go to obtain larger shape elements such as horizontal or vertical planar surfaces.

The task of the future in general is to extend the work to the 3D perception of objects and their characteristics relevant for robotic tasks. While 3D object tracking is successful when using an object model (refer to Section 3.1), the detection of objects and object shape modeling is much more difficult. As outlined by Fagg and Arbib [62], studies on humans indicate that the affordance of grasping includes information about object orientation, size, shape, and grasping points. This needs to be extracted from visual data. A good approach to detect using vision so far provides appearance-based features to indicate grasping positions [173]. The features require clear separation from the background and pick up object shape that consists basically of parallel edges. Another good example of extracting edge features and linking them to an affordance is [9], where grasping points are extracted from opposing rim features. These works for the first time showed a way to link visual features with grasp points. Shape of objects is not reconstructed, rather points with a specific characteristic found suitable to grasping.

Finally, there is a plethora of work on shape in monocular images. Computer vision started from grouping of features [133] to build up object representations. The idea is to use Gestalt laws to obtain the outline and main shape of an object. However, a large number of grouping parameters and the quadratic complexity of grouping are the main problems. Guy and Medioni [77] avoid thresholds and use an infinitely large grouping search area. Edgels and segment endpoints vote with directional vectors weighted by the length of segments and decreasing with distance. Using incremental operation the complexity can be reduced to search linear in the image space and using a parameter-free approach and ranking good groupings are found first [220]. This allows to obtain a hierarchy of feature from edgels to lines and arcs, ellipses, polygons, to basic shapes such as cones or cubes. This gives the 3D shape of objects directly for robotic experiments [191] and not only in 2D features to indicate 3D properties [145]. Nevertheless, there remains a large gap between robotics scenes with a few objects and realistic settings. To come closer to realistic scene it is useful to exploit further constraints. One such approach to obtain higher level shapes is to use vanishing points as indication for the main room structures and to obtain rectangular shape [143] and from this extract doors [147].

In summary, several approaches for obtaining shape have been demonstrated. However, there is not yet a method that can reliably extract shape properties for realistic settings at different scales. Regarding navigation, SLAM-like methods are furthest developed but are not easily applicable for extracting shapes of individual objects. In summary, the extraction of object shape lags behind the recognition of specific objects. There are indications that human vision works from several cues and range information is reconstructed only as one of these cues. It might well be that approaches are needed that integrate cues to obtain shape information of objects.

4.2 Object Categorization

At present, the task of object categorization is one of the major research topics in the computer vision community. An excellent summary has been presented by Pinz [161]. Object categorization consists of two major steps: a learning phase, where from many images categories of objects are learned, and the detection phase, where a new image is classified to belong to or to contain an object of one of these categories. Similarly to specific object recognition, learning may use different forms of supervision and constraints regarding batch processing or incremental learning. Regarding supervision in learning, we distinguish supervised learning, where the images and the category labels are given, and unsupervised learning, where only images are given and sets of labels need to be inferred for example from feature clusters. Batch processing requires a set of samples to be given to obtain the categories, while incremental learning refers to the ability of extending and possibly altering the tree of categories. At present the tree is shallow and contains categories at the same level without any grouping into higher categories (such as dog and cat into animals). A very good tutorial on the ideas of object categorization is given in Fei-Fei, Fergus and Torralba at ICCV 2005 (http://people.csail.mit.edu/torralba/iccv2005). The main stream of work today follows the good results from specific object recognition, namely exploiting local appearance-based features rather than global object appearance or shape. Work largely follows three steps [161]:

- modeling the object appearance locally,
- grouping simple geometric primitives (also referred to as codebooks), and
- using learning algorithms to find common patterns that can be shared over many individuals of a category.

These approaches are commonly tested on databases such as the Caltech[1,2] and Graz databases.[3] While the first presents samples of the categories centered in the image and in similar poses, the latter has objects

[1] http://www.robots.ox.ac.uk/˜ vgg/data3.html

[2] http://www.vision.caltech.edu/html-files/archive.html

[3] http://www.emt.tugraz.at/˜ pinz/data/

with large viewpoint variation and at different sizes and locations in the image. These two databases present different challenges and require to solve different tasks. A problem is that the specific task is often not made clear or explicit, it is all subsumed in the categorization task. A task not specified in the description of learning above is the localization of the object in the image. We will see that this makes the task much more difficult when scrutinizing the results of the Pascal challenges.

In the last years the Pascal challenges on visual object categorization (EU PASCAL project[4]) attempted to better formalize the procedure and degrees of difficulty of the data. Objects are annotated with bounding boxes, main view direction, and a truncated, occluded or difficult flag. In 2008 there have been the object class competition, the detection challenge, and the segmentation taster challenge. To highlight the present state-of-the-art, some of the results of the Pascal Challenge are reported subsequently.

An observation we can make is that approaches at present seem to converge on using SIFT [132] and histogram features (e.g., spatial pyramid [175] or HoG [42]), weighting features of the created codebook, and using a classification scheme, where Support Vector Machine (SVM) is most popular. This all aims at learning statistics of image features over the object classes. It is interesting to note that in some databases, e.g., Caltech, background is treated as one category although containing images with something visible such as boxes, fields, or buildings.

In the object class competition the averaged precision reached 57% using methods from the Universities of Amsterdam and Surrey [210]. The approach builds a codebook from a circular color descriptor and classifies objects using Kernel Discriminant Analysis (KDA) with spectral regression, which performed better than SVM. Overall, one of the classes with worst results is the bottle. Dining table, potted plant, cow and sofa, are considered rather difficult too. Person, air plane, train, and horse are easy to distinguish. Chairs are often many, close to each other, and this “texture” seemed to be picked up. Indoor scenes are often false positives for chairs because these statistics pick up the image structure

[4] http://pascallin.ecs.soton.ac.uk/challenges/VOC

rather than the real object class. In general learning more (including the VOC 2007 data set) seems to slightly improve results [61].

In the detection challenge the objects also need to be localized, that is, the bounding box area needs to overlap ground truth at least 50%. The top detection rate was 23% by work using SIFT and HOG features with SVM [83]. In the segmentation taster challenge the task was to specify the class for each pixel allowing a 5-pixel-wide void region at the border. Pascal VOC presents lot of work with detailed annotation, e.g., tables and chairs. Average precision rate was 25% and the best entry by XEROX (using RGB and gradient histogram features and a mean shift segmentation) [61]. It is interesting to note that the second best approach did very good on tables using shape clustering of texton, color, and HOP features [123].

The differences in the results obtained for the object class and the detection challenges indicates that the image context plays a major role in the statistical approaches. This is confirmed when looking at false positives, where birds in the sky are mostly taken as planes or indoor structure as a chair. At present the methods rather learn scene class rather than object class. While infants train vision from looking at a few objects repeatedly and from many sides, databases contain a few images of many object categories. From the point of view of a robot the task is again different, because the search for the object contains yet another dimension of difficulty (also see the semantic robot vision challenge at CVPR in Section 4.4). Finally, it seems yet a long way to object classes that are based on affordances and functions of objects, that are required for the robot to fulfill a given task.

4.3 Semantics and Symbol Grounding: From Robot Task to Grasping and HRI

Robots of the future should be able to easily navigate in dynamic and crowded environments, detect as well as avoid obstacles, have a dialog with a user, and manipulate objects. It has been widely recognized that, for such a system, different processes have to work in synergy: high-level cognitive processes for abstract reasoning and planning, low-level sensory–motor processes for data extraction and action execution, and mid-level processes mediating these two levels.

A successful coordination between these levels requires a well-defined representation that facilitates anchoring of different processes. One of the proposed modeling approaches has been the use of *cognitive maps* [18]. The cognitive map is the body of knowledge a human or a robot has about the environment. In [18], it is argued that topological, semantic, and geometrical aspects are important for representation of spatial knowledge. This approach is closely related to Human-Augmented mapping (HAM) where a human and a robot interact so to establish a correspondence between the human spatial representation of the environment and robot's autonomously learned one [116].

In addition, both during the mapping phase and during robot task execution, object detection can be used to augment the map of the environment with objects' locations [56]. There are several scenarios here: while the robot is building the map it adds information to the map about the location of objects. Later, the robot is able to assist the user when she/he wants to know where an object X is. As object detection might be time consuming, another scenario is that the robot builds a map of the environment first and then when no tasks are scheduled for execution, it moves around in the environment and searches for objects.

Early work recognized that a robot has the potential to examine its world using causality, by performing probing actions and learning from the response [140]. Visual cues were used to determine what parts of the environment were physically coherent through interplay of objects, actions, and imitations. In relation to representation of object properties, there is a close connection to *anchoring* [37] that connects, inside an artificial system, the symbol-level and signal-level representations of the same physical object. Although some nice ideas about the representations are proposed, there is no attempt of developing the underlying vision system necessary for extraction of symbols.

Southey and Little [190] examine the problem of object discovery defined as autonomous acquisition of object models, using a combination of motion, appearance, and shape. The authors discuss that object discovery is complicated due to the lack of a clear definition of what constitutes an object. They state that rather than trying for an all-encompassing definition of an object that would be difficult or impossible to apply, a robot should use a definition that identifies objects useful for it. From the perspective of the object-fetching robot, useful objects would be structures that can be picked up and carried. Similar line of thinking is pursued in [91] that also extracts a set of object attributes that can be used for manipulation purposes or further learning of object properties.

4.4 Competitions and Benchmarking

An important common characteristic of both robotics and computer vision is that both are highly hardware and application dependent, and therefore many similar problems exist even though “pure” computer vision has still somewhat less variation. In both fields the tasks to be achieved are complex, making analytical performance prediction impossible in many cases thus leaving empirical study as the only available approach. For this reason, the test cases of the empirical studies, as well as the analysis of the results of experiments, are the most important considerations in benchmarking.

The increased interest to benchmarking in computer vision during the last decade can be easily seen in a number of projects concentrating on benchmarking. For example, the EU-funded Performance Characterization in Computer Vision (PCCV) project produced tutorials and case-studies for benchmarking vision algorithms [155].

The area of robotics is very wide and includes a large range of research fields. That this is the case is evident when studying the list of sessions or the topics of interest mentioned in the CFP for one of the major robotics conferences such as IROS and ICRA. An incomplete list of these areas include manipulation, obstacle avoidance, planning, humanoids, hardware design, SLAM, vision, sensors, teleoperation, and learning. Vision is thus just one of many topics in robotics. Many of the areas also have subdomains just like vision has (object recognition, tracking, etc.) and some of the areas are intimately connected.

Active control of sensors, which is the core part of robotic applications, does not allow for easy performance evaluation on data sets. In some cases, but not all, simulation provides a way to at least repeat experiments with exactly the same and well-known environmental conditions. The problem with simulation is that it is only as good as the simulation model and typically never fully captures the complexity of the real world.

There have been a number of successful contests within robotics. Some like the one at the AAAI conference has been running for a long time. Recently the DARPA Grand Challenge and Urban Challenge generated a lot of media attention and the RoboCup is also something that many outside of the community have heard about.

- The 12th annual AAAI Mobile Robot Competition
  http://robots.net/rcfaq.html
- Semantic Vision Challenge
  http://www.semantic-robot-vision-challenge.org
- RoboCup that has as a long-term goal to develop a robot soccer team that will beat human world champions
  http://www.robocup.org
- DARPA Grand Challenge and Urban Challenge-
  http://www.darpa.mil/grandchallenge/index.asp
- ELROB - 1st European Land-Robot Trial
  http://www.elrob2006.org

As was already discussed in the context of vision, there is a need for available baseline methods when evaluating new robot applications. It is quite common that a researcher only provides a comparison of the new results with his last result and that of his group. Having a set of available methods to compare to would advance the field. There is some code available but comparisons are made difficult because the hardware is typically different as well.

Semantic Robot Vision Challenge [176] is a research competition that is designed to push the state-of-the-art in image understanding and automatic acquisition of knowledge from large unstructured databases of images such as those generally found on the Web. In this competition, fully autonomous robots receive a text list of objects that they are to find. They use the Web to automatically find image examples of those objects in order to learn visual models. These visual models are then used to identify the objects in the robot's cameras. The results clearly demonstrate the difficulty. In 2008 the winning team of the University of British Columbia detected none of the ten category objects and three of the ten specific objects using a detection approach based on SIFT features [139]. The authors conclude that present methods have great difficulties to incorporate efficient object search methods with reliable object recognition while object categorization is not within reach yet.

In terms of performance evaluation in the area of SLAM, execution time and computational complexity have been the two most quantitative measures so far. The size of the environment that a certain algorithm could handle and still remain consistent has also been used but it is not until the same data sets have been used that this has been a really useful measure. A problem with using the common data sets is that the same set is typically used for parameter tuning and evaluation, that is, the parameters of the algorithm are adapted to make the results as good as possible on a certain data set as opposed to tuning for one and then running on another as would be the proper experimental evaluation procedure. What is still missing is a generally accepted metric for evaluating the quality the generated map on any of the many available data sets. To use an extreme example, how does one compare the result of two SLAM algorithms if one builds a metric map and the other a topological map? Evaluating the quality would require having some kind of ground truth which in anything but a toy environment or simulation is a staggering task. One thing that would be possible to measure and compare with ground truth would be the position estimate that a SLAM algorithm produces. There are plenty of accurate localization systems reported in the literature which could be used to gather the ground truth position data.

5

Discussion and Conclusion

Vision advanced to serve a large series of applications. The review attempted to give an overview of vision methods suitable for robotic systems. However, robotic applications are very diverse and also this article had to take focus. Vision begins to provide more and more functionalities and hence allows an increasing number of applications in the wide range of robotics.

For example, space applications use vision to surface estimation for landing or to navigate on Mars [31]. Space stations use vision to control remote robots (e.g., [1]) and to assist coping with the time delay in teleoperation (e.g., [178]). Teleoperation is also used in medical applications [170], where vision methods are used to scan and model body parts or to navigate the tool [5]. Navigation has also found its way to cars. While the DARPA Grand Challenge (Section 4.4) attempts to make vehicles autonomous, driver assistance systems are becoming standard technology in automobiles. Sensors around the car measure distances to other vehicles. Vision is used to observe the driver's eyes and to observe the situations on the road ahead. For a recent overview see [204]. The list of applications can be extended toward a review on applications of vision to the different fields of robotics. The main focus of this review was to give a first insight on what vision can do for an assistive robot, which is the main driving force for the future commercial applications [10].

An assistive robot must be able to perform actions in its environment and therefore needs the ability to interact with it. It should be able to fetch, pickup, and place objects, open and close doors and drawers, and support humans while walking. So, what are the prerequisites for a robot to successfully carry out such tasks? Initially, there is a need for flexible interaction with the user, both through user input to the robot and feedback to the user. The user should instruct the robot *what* to do and sometimes also *how* to do it.

Given the instruction, the robot should safely navigate to the place where it expects to find the object. To be able to plan the path, the robot must know its whereabouts, or ideally, the robot should be able to determine its position in the environment. If, for example, the task is to fetch a package of raisins from the kitchen table, the robot should navigate to the kitchen and position itself in front of the table. Then, a “fetch” or a “pick–up” task should be initiated.

All the above tasks can be solved using visual feedback. However, as outlined in the review, more has to be done in order to achieve systems with robust performance. The performance of a vision system depends also on a number of other factors and needs to be assessed in terms of a whole robot system. Using the classical [*sense–plan–act*] framework, below are the problems that affect and define robustness and flexibility of the complete robotic system:

(1) *Perceptual Robustness*: How do sensor design and choice of image processing techniques affect the performance of a visual servoing system?

(2) *Robustness Issues in Planning*: How to plan trajectories and grasps with respect to obstacle avoidance, robot singularities?

(3) *Robustness Issues in Control Generation*: What are the main issues and requirements with respect to the system design and how should the sensor measurements be used?

The review has attempted to make clear that there are only a few blocks in Figure 1.1 that work in realistic environments. Object tracking, human tracking, or specific object recognition are on good ways. However, one cannot expect that any of these work out-of-the-box. There are good methods but none works all times. Robustness remains as the biggest road block to widespread applications. In particular, problems still unresolved are robustness to varying lighting conditions and environments with more clutter than the typical orderly laboratory. Additionally, the databases of images used in computer vision do not adequately capture the environment as seen from a robot.

Besides *robustness* there are however a few other principle items that remain to be solved. One topic is *scalability*, which refers to the quality of a method to scale out to a large number of objects, tasks, environments and applications. Specific object recognition methods have been shown to scale well. Tracking of multiple objects will scale with the number of objects. In SLAM methods start to cope with large environments. There is still the issue of the kidnapped robot problem, which means to put the robot into an unknown context. In this case landmark detection needs to scale well while in navigation the present position enables scaling to the given context or location information. Yet open are tasks such as object categorization or the natural interaction with humans. Object categorization and binding objects to shapes is another open problem, where solutions might develop hand-in-hand. Interaction with a robot not only depends on vision but also on other sensor modalities, in particular speech. The data fusion into a common, amodal representation is a present research topic, e.g., the EU Project CogX, http://cogx.eu.

A constraint that should not be forgotten is that robot vision systems are a means to interact with the environment or humans. Hence performance is important. Examples of how the *real-time* constraint influences robot vision are optical flow and stereo. Both are rather intensive to process but excellent cues. In particular motion is known to be a dominant cue in object segmentation in humans. However, the Middlebury data sets list methods first that take rather excessive time to process image pairs (as required in both techniques). Recently GPU implementations highly improved the situation (e.g., www.gpu4vision.org). Nevertheless, even on today's computers robot vision requires a well-selected mix of fast methods to perform in reasonable time.

Looking at the above review one could summarize, if tuned to a specific setting, robot vision results look acceptable. However, there is little *flexibility* to change the task, the objects, or the context. This lack of *generality* presents great opportunities for future work. Which also brings us to one of the major constraints of robot vision, if not robotics at large. The evaluation of methods is often executed in specific settings, on specific robots and objects. Hence comparison is difficult to impossible. However, scientific advance requests comparison. Benchmarks and competitions are improving the situation (also refer to Section 4.4). Additionally, computer vision challenges, such as the Pascal challenge, have become popular, however, these do not solve the robot vision problem. Possibly we need similar robotics vision challenges to make the advances more transparent and methods more widely usable by others.

Which brings us to the final problem of robotics vision – *integration*. As stated, vision is part of the robot system and serves a certain task. This also means that there are many components of the robotic system that need to be integrated. Section 2.3 on vision systems already stressed this fact and a recent conference series (ICVS – International Conference on Computer Vision Systems) tries to make the community more aware. The importance of this aspect is that the capabilities of visual perception in robotics can only be assessed if operating in a system. Vision alone will always be different, solve a different task, and it will fail with certainty when added to the robot system. A typical example is object recognition (Section 3.4). While considered solved in the computer vision research community, the Semantic Robot Vision Challenge (Section 4.4) gave a completely different picture for a robot searching the object. Aspects of where to look, the viewing angle in relation to the object, and the different lighting and clutter conditions pose challenges that yet need to be solved.

Given the difficulties and open problems in robotics vision, in particular robustness, one has to think about the general approach to robot vision. Possibly a consequence is to rather move toward a [*predict–act–sense*] approach following the active vision paradigm. While already stated decades ago, the strict consequences are little followed, as the review here indicates. Work on actively using the arm to segment objects or linking affordances to visual features are first starts [141, 9]. A reason is certainly, that in this context vision needs to be treated in a completely different way. It is not a stand-alone component that delivers valuable input. Vision is one possible sensor modality to achieve a certain robot task. The task of vision is to provide specific information about the environment and its purpose is directly linked with the intended action.

Acknowledgments

This work has been supported by EU through the projects GRASP, FP7-2008-IST-4-215821, CogX, FP7-2008-IST-4-215181, and the Austrian Science Fondation, Project S9101-N13.

References

[1] A. Albu-Schaffer, W. Bertleff, B. Rebele, B. Schafer, K. Landzettel, and G. Hirzinger, "Rokviss–robotics component verification on ISS current experimental results on parameter identification," in *ICRA*, p. 38793885, 2006.

[2] R. Alferez and Y. Wang, "Geometric and illumination invariants for object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 6, pp. 505–536, 1999.

[3] Y. Aloimonos and D. Shulman, *Integration of Visual Modules.* Academic Press, Inc., 1989.

[4] Y. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," in *Proceedings of the DARPA Image Understanding Workshop*, pp. 552–573, 1987.

[5] R. T. amd D. Stoianovici, "Medical robotics in computer-integrated surgery," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 765–781, 2003.

[6] O. Amidi, T. Kanade, and R. Miller, *Vision-Based Autonomous Helicopter Research at CMU.* in [6], 2000.

[7] R. Bajcsy, "Active perception," *in Proceedings of the IEEE*, vol. 76, no. 8, pp. 996–1005, 1988.

[8] D. H. Ballard, "Animate vision," *Artificial Intelligence*, vol. 48, no. 1, pp. 57–86, 1991.

[9] E. Baseski, N. Pugeault, S. Kalkan, D. Kraft, F. Wrgtter, and N. Krger, "A scene representation based on multi-modal 2d and 3d features," in *ICCV*, pp. 1–7, 2007.

[10] G. Bekey and J. Yuh, "The status of robotics," *IEEE Robotics & Automation Magazine*, vol. 15, no. 1, pp. 80–86, 2008.

[11] S. Benhimane and E. Malis, “Homography-based 2d visual tracking and servoing,” *International Journal of Robotic Research (Special Issue on Vision and Robotics Joint with the International Journal of Computer Vision)*, vol. 26, no. 7, pp. 661–676, 2007.

[12] M. Bertero, T. Poggio, and V. Torre, “Ill-posed problems in early vision,” in *Proceedings of the IEEE*, pp. 869–889, 1988.

[13] F. Berton, G. Sandini, and G. Metta, “Anthropomorphic visual sensors,” in *In Encyclopedia of Sensors*, American Scientific Publishers, 2005.

[14] I. Biederman, “Recognition-by-components: A theory of human image understanding,” *APA Journal; Psychological Review*, vol. 94, no. 2, pp. 115–147, 1987.

[15] G. Biegelbauer, A. Pichler, M. Vincze, C. Nielsen, H. Andersen, and K. Haeusler, “The inverse approach of flexpaint [robotic spray painting],” *IEEE Robotics & Automation Magazine*, vol. 12, no. 3, pp. 24–34, 2005.

[16] G. Biegelbauer, M. Vincze, and W. Wohlkinger, “Model-based 3d object detection–Efficient approach using superquadrics,” *Machine Vision Applications,* vol. accepted, 2008.

[17] H. Bischof, H. Wildenauer, and A. Leonardis, “Illumination insensitive eigenspaces,” in *IEEE International Conference Computer Vision ICCV*, pp. 233–238, 2001.

[18] B.J.Kuipers, “The cognitive map: Could it have been any other way?,” in *Spatial Orientation: Theory, Research, and Application*, (H. L. Pick, Jr., and L. P. Acredolo, eds.), pp. 345–359, New York: Plenum Press, 1983.

[19] M. Björkman and D. Kragic, “Combination of foveal and peripheral vision for object recognition and pose estimation,” *in Proceedings of the IEEE International Conference on Robotics and Automation, ICRA’04,* vol. 5, pp. 5135–5140, April 2004.

[20] R. Bolles and P. Horaud, “3dpo: A three-dimensional part orientation system,” *The International Journal of Robotics Research*, vol. 5, no. 3, pp. 3–26, 1986.

[21] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, “Appearance based active object recognition,” *International Journal of Image and Vision Computing*, vol. 18, no. 9, pp. 715–728, 2000.

[22] S. Brandt, C. Smith, and N. Papanikolopoulos, “The Minnesota robotic visual tracker: A flexible testbed for vision-guided robotic research,” in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, “Humans, Information and Technology”*, vol. 2, pp. 1363–1368, 1994.

[23] C. Brautigam, J. Eklundh, and H. Christensen, “A model-free voting approach for integrating multiple cues,” in *ECCV*, pp. 734–750, 1998.

[24] R. Brooks, “Model-based 3d interpretation of 2d images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, pp. 140–150, 1983.

[25] M. Brown, D. Burschka, and G. Hager, “Advances in computational stereo,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 993–1008, p. 8, 2003.

[26] J. Byne and J. Anderson, “A CAD-based computer vision system,” *Image and Vision Computing*, vol. 16, pp. 533–539, 1998.

[27] J. A. Castellanos and J. D. Tardós, *Mobile Robot Localization and Map Building: A Multisensor Fusion Approach.* Kluwer Academic Publishers, 1999.

[28] L. Y. Chang, N. S. Pollard, T. M. Mitchell, and E. P. Xing, "Feature selection for grasp recognition from optical markers," in *IEEE International Conference on Intelligent Robots and Systems*, 2007.

[29] F. Chaumette, "Potential problems of stability and convergence in image-based and position-based visual servoing," in *The Confluence of Vision and Control,* no. 237 in Lecture Notes in Control and Information Sciences, pp. 66–78, Springer-Verlag, 1998.

[30] F. Chaumette and S. Hutchinson, "Visual servo control I: Basic approaches," *IEEE Robotics and Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[31] Y. Cheng, M. Maimone, and L. Matthies, "Visual odometry on the mars exploration rovers–A tool to ensure accurate driving and science imaging," *IEEE Robotics & Automation Magazine*, vol. 13, no. 54-62, p. 2, 2006.

[32] S. Choi, S. Ban, and M. Lee, "Biologically motivated visual attention system using bottom-up saliency map and top-down inhibition," *Neural Information Processing–Letters and Review*, vol. 2, pp. 19–25, 2004.

[33] H. I. Christensen and H.-H. Nagel, eds., *Cognitive Vision Systems: Sampling the Spectrum of Approach.* Springer Verlag, Lecture Notes in Computer Science, pp. 3948, 2006.

[34] J. Clark and A. Yuille, *Data Fusion for Sensory Information Processing Systems.* Kluwer Academic Publisher, 1990.

[35] A. Comport, E. Marchand, and F. Chaumette, "A real-time tracker for markerless augmented reality," in *IEEE International Symposium on Mixed and Augmented Reality*, pp. 36–45, 2003.

[36] A. Comport, M. Pressigout, E. Marchand, and F. Chaumette, "A visual servoing control law that is robust to image outliers," in *Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 492–497, Las Vegas, Nevada, October 2003.

[37] S. Coradeschi and A. Saffiotti, "An introduction to the anchoring problem," *Robotics and Autonomous Systems, Special Issue on Perceptual Anchoring*, vol. 43, no. 2–3, pp. 85–96, 2003.

[38] P. Corke and S. Hutchinson, "A new partitioned approach to image-based visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 17, pp. 507–515, 2001.

[39] P. I. Corke, *Visual Control of Robots: High Performance Visual Servoing.* Research Studies Press, John Wiley, 1996.

[40] J. Crowley and H. Christensen, *Vision as Process.* Springer Verlag, 1995.

[41] L. Crowley, J. Coutaz, and F. Bérard, "Things that see: Machine perception for human computer interaction," *Communications of the A.C.M.*, vol. 43, pp. 54–64, March 2000.

[42] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, pp. 886–893, 2005.

[43] B. Dariush, "Human motion analysis for biomechanics and biomedicine," *Machine Vision and Applications*, vol. 14, pp. 202–205, 2003.

[44] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *ICCV*, 2003.

[45] M. Deans and M. Hebert, “Experimental comparison of techniques for localization and mapping using a bearings only sensor,” in *ISER’00, Seventh International Symposium on Experimental Robotics*, December 2000.

[46] L. Deng, W. J. Wilson, and F. Janabi-Sharifi, “Characteristics of robot visual servoing methods and target model estimation,” in *Proceedings of the 2002 IEEE International Symposium on Intelligent Control*, pp. 684–689, Vancouver, Canada, October 27–30 2002.

[47] S. Dickinson, D. Wilkes, and J. Tsotsos, “A computational model of view degeneracy,” *IEEE Transactions on PAMI*, vol. 21, no. 8, pp. 673–689, 1999.

[48] E. Dickmanns, *Dynamic Vision for Perception and Control of Motion*. Springer, 2007.

[49] E. D. Dickmanns and V. Graefe, “Dynamic monocular machine vision,” *Machine Vision and Applications*, vol. 1, pp. 223–240, 1988.

[50] R. Dillmann, “Teaching and learning of robot tasks via observation of human performance,” *Robotics and Autonomous Systems*, vol. 47, pp. 109–116, 2004.

[51] G. Dissanayake, P. Newman, S. Clark, H. Durrant-Whyte, and M. Corba, “A solution to the slam building problem,” *IEEE TRA*, vol. 17, no. 3, pp. 229–241, 2001.

[52] T. Drummond and R. Cipolla, “Real-time visual tracking of complex structures,” *IEEE Transactions on PAMI*, vol. 24, no. 7, pp. 932–946, 2002.

[53] C. Eberst, M. Barth, K. Lutz, A. Mair, S. Schmidt, and G. Farber, “Robust vision-based object recognition integrating highly redundant cues for indexing and verification,” in *IEEE ICRA*, pp. 3757–3764, 2000.

[54] A. Edsinger and C. C. Kemp, “Manipulation in human environments,” in *IEEE/RSJ International Conference on Humanoid Robotics*, pp. 102–109, 2006.

[55] Eklundh, Bjorkman, and Hayman, “Object appearance from integration of 3d and 2d cues in real scenes,” *Journal of Vis.*, vol. 3, pp. 646–646, October 2003.

[56] S. Ekvall, P. Jensfelt, and D. Kragic, “Object detection and mapping for service robot tasks,” *Robotics*, vol. 25, no. 2, pp. 175–188, 2007.

[57] S. Ekvall and D. Kragic, “Interactive grasp learning based on human demonstration,” in *IEEE International Conference on Robotics and Automation, ICRA’04*, 2004.

[58] S. Ekvall, D. Kragic, and P. Jensfelt, “Object detection and mapping for service robot tasks,” *Robotica*, vol. 25, pp. 175–187, 2007.

[59] B. Espiau, “Effect of camera calibration errors on visual servoing in robotics,” in *3rd International Symposium on Experimental Robotics*, Kyoto, Japan, October 1993.

[60] B. Espiau, F. Chaumette, and P. Rives, “A new approach to visual servoing in robotics,” *IEEE Transactions on Robotics and Automation*, vol. 8, pp. 313–326, June 1992.

[61] M. Everingham, “Overview and results of the classification challenge,” *http: pascallin.ecs.soton.ac.uk challenges VOC voc2008 workshop everingham_cls.pdf.*

[62] A. H. Fagg and M. A. Arbib, “Modeling parietal–premotor interactions in primate control of grasping,” *Neural Networks*, vol. 11, no. 7–8, pp. 1277–1303, 1998.

[63] J. Feddema and C. Lee, "Adaptive image feature prediction and control for visual tracking with a hand–eye coordinated camera," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 20, no. 5, pp. 1172–1183, 1990.

[64] W. Feiten, B. Magnussen, J. Bauer, G. Hager, and K. Toyama, "Modeling and control for mobile manipulation in everyday environments," in *8th International Symposium on Robotics Research*, 1998.

[65] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *CVPR*, pp. 264–271, 2003.

[66] V. Ferrari, T. Tuytelaars, and L. Gool, "Simultaneous object recognition and segmentation from single or multiple model views," *International Journal of Computer Vision*, vol. 67, no. 2, pp. 159–188, 2006.

[67] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini, "Learning about objects through action–Initial steps towards artificial cognition," in *IEEE International Conference on Robotics and Automation*, pp. 3140–3145, 2003.

[68] J. Folkesson, J. P, and H. I. Christensen, "Vision slam in the measurement subspace," in *IEEE ICRA05*, 2005.

[69] S. Frintrop, "VOCUS: A visual attention system for object detection and goal-directed search," *Lecture Notes in Computer Science, Springer*, vol. 3899, 2006.

[70] A. Gee, D. Chekhlov, A. Calway, and W. Mayol-Cuevas, "Discovering higher level structure in visual slam," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 980–990, 2008.

[71] J. J. Gibson, *The Ecological Approach to Visual Perception.* Lawrence Erlbaum Associates, 1987.

[72] L. Goncavles, E. di Bernardo, D. Benson, M. Svedman, J. Ostrovski, N. Karlsson, and P. Pirjanian, "A visual front-end for simultaneous localization and mapping," in *IEEE ICRA*, pp. 44–49, 2005.

[73] M. Goodale and A. Milner, "Separate visual pathways for perception and action," *Trends Neuroscience*, vol. 15, no. 1, p. 205, 1992.

[74] A. Gopalakrishnan and A. Sekmen, "Vision-based mobile robot learning an navigation," in *IEEE International Workshop on Robot and Human Interactive Communication, RO-MAN'05*, pp. 48–53, 2005.

[75] W. Grimson, *Object Recognition by Computer: The Role of Geometric Constraints.* MIT Press, 1990.

[76] J. Gutmann and K. Konolige, "Incremental mapping of large cyclic environments," in *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp. 318–325, 1999.

[77] G. Guy and G. Medioni, "Inferring global perceptual contours from local features," *International Journal of Computer Vision*, vol. 20, no. 1–2, pp. 113–133, 1996.

[78] G. Hager, "A modular system for robust positioning using feedback from stereo vision," *IEEE Transactions on Robotics and Automation*, vol. 13, no. 4, pp. 582–595, 1997.

[79] G. Hager and P. Belhumeur, "Real-time tracking of image regions with changes in geometry and illumination," in *Proceedings of the Computer*

Society Conference on Computer Vision and Pattern Recognition, CVPR'96, pp. 403–410, 1996.

[80] G. Hager and P. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, 1998.

[81] G. Hager and K. Toyama, "The XVision system: A general-purpose substrate for portable real-time vision applications," *Computer Vision and Image Understanding*, vol. 69, no. 1, pp. 23–37, 1996.

[82] C. Harris, "Tracking with rigid models," in *Active Vision*, (A. Blake and A. Yuille, eds.), pp. 59–73, MIT Press, 1992. ch. 4.

[83] H. Harzallah, C. Schmid, F. Jurie, and A. Gaidon, "Classification aided two stage localization," *http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/workshop/harzallah.pdf*.

[84] E. Hayman and J. Eklundh, "Probabilistic and voting approaches to cue integration for figure-ground segmentation," in *ECCV, Springer LNCS 2352*, pp. 469–486, 2002.

[85] E. Hayman and J.-O. Eklundh, "Statistical background subtraction for a mobile observer," in *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pp. 67–74, 2003.

[86] S. Helmer and D. G. Lowe, "Object class recognition with many local features," in *CVPR GMBV Workshop on Generative-Model Based Vision*, 2004.

[87] J. Hill and W. Park, "Real time control of a robot with a mobile camera," in *Proceedings of the 9th ISIR*, pp. 233–246, 1979.

[88] D. Hoffman and W. Richards, "Parts of recognition," *Cognition*, vol. 18, pp. 65–96, 1996.

[89] A. Hoover, G. Jean-Baptiste, X. Jiang, and et al., "An experimental comparison of range image segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 673–689, 1996.

[90] B. Horn, *Robot Vision*. MIT Press, 1986.

[91] K. Huebner, M. Bjorkman, B. Rasolzadeh, M. Schmidt, and D. Kragic, "Integration of visual and shape attributes for object action complexes," in *6th International Conference on Computer Vision Systems (ICVS'08), Lecture Notes in Artificial Intelligence*, vol. 5008, pp. 13–22, D-69121 Heidelberg, Germany: Springer-Verlag, 2008.

[92] S. Hutchinson, G. D. Hager, and P. I. Corke, "A tutorial on visual servo control," *IEEE Transactions on Robotics and Automation*, vol. 12, pp. 651–670, October 1996.

[93] M. Isard and A. Blake, "CONDENSATION-Conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[94] L. Itti, "Models of bottom-up and top-down visual attention," PhD thesis, California Institute of Technology, 2000.

[95] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[96] A. Jaklic, A. Leonardis, and F. Solina, *Segmentation and Recovery of Superquadrics*. Kluwer Academic Publishers, 2000.

[97] P. Jensfelt, J. Folkesson, D. Kragic, and H. I. Christensen, "Exploiting distinguishable image features in robotic mapping and localization," in *1st European Robotics Symposium (EUROS-06)*, (H. I. Christensen, ed.), Palermo, Italy, March 2006.

[98] P. Jensfelt, D. Kragic, J. Folkesson, and M. Björkman, "A framework for vision based bearing only 3D SLAM," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'06)*, Orlando, FL, May 2006.

[99] F. Jurie, "Robust hypothesis verification: Application to model-based object recognition," *Pattern Recognition*, vol. 32, no. 6, 1999.

[100] F. Jurie and M. Dhome, "Real time tracking of 3D objects: An efficient and robust approach," *Pattern Recognition*, vol. 35, pp. 317–328, 2002.

[101] N. Karlsson, E. di Bernardo, J. Ostrowski, L. Goncalves, P. Pirjanian, and M. Munich, "The vSLAM algorithm for robust localization and mapping," in *International Conference on Robotics and Automation*, pp. 24–29, Barcelona, Spain, April 18–22 2005.

[102] D. Katsoulas, C. Bastidas, and D. Kosmopoulos, "Superquadric segmentation in range images via fusion of region and boundary infromation," in *IEEE Transactions on PAMI*, vol. 30, no. 5, pp. 781–795, 2008.

[103] T. Kawanishi, H. Murase, and S. Takagi, "Quick 3D object detection and localization by dynamic active search with multiple active cameras," in *IEEE International Conference on Pattern Recognition, ICPR'02*, pp. 605–608, 2002.

[104] H. Kjellstrom, J. Romero, and D. Kragic, "Visual recognition of grasps for human-to-robot mapping," in *IROS*, 2008.

[105] G. Klein and T. Drummond, "Robust visual tracking for non-instrumented augmented reality," in *Proceedings of the 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 113–122, 2003.

[106] G. Klein and D. Murray, "Improving the agility of keyframe-based slam," in *ECCV*, 2008.

[107] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[108] T. Koike and J. Saiki, "Stochastic guided search model for search asymmetries in visual search tasks," *Biologically Motivated Computer Vision*, pp. 408–417, 2002.

[109] D. Koller, K. Daniilidis, and H. Nagel, "Model-based object tracking in monocular image sequences of road traffic scenes," *International Journal of Computer Vision*, vol. 10, no. 3, pp. 257–281, 1993.

[110] S. Kovacic, A. Leonardis, and F. Pernus, "Planning sequences of views for 3-d object recognition and pose determination," *Pattern Recognition*, vol. 31, no. 10, pp. 1407–1417, 1998.

[111] D. Kraft, E. Baseski, M. Popovic, N. Krüger, N. Pugeault, D. Kragic, S. Kalkan, and F. Wörgötter, "Birth of the object: Detection of objectness and extraction of object shape through object action complexes," *International Journal of Humanoid Robotics*, vol. 5, pp. 247–265, 2008.

[112] D. Kragic, "Visual servoing for manipulation: Robustness and integration issues," PhD thesis, CVAP, Royal Institute of Technology, Stockholm, Sweden, 2001.

[113] D. Kragic, M. Bjorkman, H. Christensen, and J.-O. Eklundh, “Vision for robotic object manipulation in domestic settings,” *Elsevier; Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 85–100, 2005.

[114] D. Kragic and H. Christensen, “Cue integration for visual servoing,” *IEEE Transactions on Robotics and Automation*, vol. 17, no. 1, pp. 18–27, 2001.

[115] J. Krivic and F. Solina, “Art-level object recognition using superquadrics,” *Elsevier; Computer Vision and Image Understanding*, vol. 95, no. 1, pp. 105–126, 2004.

[116] G.-J. M. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen, “Clarification dialogues in human-augmented mapping,” in *Proceedings of the 1st Annual Conference on Human-Robot Interaction, HRI’06*, Salt Lake City, UT, March 2006.

[117] A. Krupa, J. Gangloff, C. Doignon, M. de Mathelin, G. Morel, J. Leroy, L. Soler, and J. Marescaux, “Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing,” *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 842–853, 2003.

[118] N. M. Kwok, G. Dissanayake, and Q. Ha, “Bearing only SLAM using a SPRT based Gaussian sum filter,” in *IEEE ICRA05*, 2005.

[119] V. Kyrki and D. Kragic, “Integration of model-based and model-free cues for visual object tracking in 3d,” in *IEEE International Conference on Robotics and Automation, ICRA’05*, pp. 1566–1572, 2005.

[120] V. Kyrki and D. Kragic, “Tracking unobservable rotations by cue integration,” in *IEEE International Conference on Robotics and Automation 2006. ICRA’06*, pp. 2744–2750, Orlando, Florida, 2006.

[121] V. Kyrki, D. Kragic, and H. Christensen, “Measurement errors in visual servoing,” *Robotics and Autonomous Systems*, vol. 54, no. 10, pp. 815–827, 2006.

[122] V. Kyrki and K. Schmock, “Integation methods of model-free features for 3d tracking,” in *Scandinavian Conference on Image Analysis*, pp. 557–566, 2005.

[123] L. Ladicky, P. Torr, and P. Kohli, “Object-class segmentation using higher order CRF,” *http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/workshop/ladicky.pdf*.

[124] K. Lee, H. Buxton, and J. Feng, “Selective attention for cueguided search using a spiking neural network,” in *Proceedings of the International Workshop on Attention and Performance in Computer Vision*, pp. 55–62, Graz, Austria, July 2003.

[125] S. Lee, D. Jang, E. Kim, S. Hong, and J. Han, “A real-time 3d workspace modeling with stereo camera,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2140–2147, August 2005.

[126] T. Lemaire, S. Lacroix, and J. Solà, “A practical 3D bearing-only SLAM algorithm,” in *IEEE/RSJ IROS*, pp. 2757–2762, 2005.

[127] A. Leonardis and H. Bischof, “Robust recognition using eigenimages,” *Computer Vision and Image Understanding: CVIU*, vol. 78, no. 1, pp. 99–118, 2000.

[128] Z. Li, “A saliency map in primary visual cortex,” *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 9–16, 2002.

[129] G. Lopez-Nicolas, C. Sagues, J. Guerrero, D. Kragic, and P. Jensfelt, "Nonholonomic epipolar visual servoing," in *In IEEE International Conference on Robotics and Automation 2006. ICRA'06*, pp. 2378–2384, Orlando, Florida, 2006.

[130] D. G. Lowe, "Robust model-based motion tracking through the integration of search and estimation," *International Journal of Computer Vision*, vol. 8, no. 2, pp. 113–122, 1992.

[131] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision*, pp. 1150–1157, 1999.

[132] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[133] D. G. Lowe, "Three-dimensional object recognition from single two-dimensional images," *Artificial Intelligence*, vol. 31, pp. 355–395, March 1987.

[134] E. Malis and F. Chaumette, "Theoretical improvements in the stability analysis of a new class of model-free visual servoing methods," *IEEE Transactions on Robotics and Automation*, 2002.

[135] E. Malis, F. Chaumette, and S. Boudet, "2-1/2-d visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 15, pp. 238–250, April 1999.

[136] G. Mariottini, D. Prattichizzo, and G. Oriolo, "Epipole-based visual servoing for nonholonomic mobile robots," in *ICRA*, 2004.

[137] D. Marr, *Vision*. San Francisco: W. H. Freeman and Company, 1982.

[138] L. Masson, F. Jurie, and M. Dhome, "Robust real time tracking of 3d objects," *International Confrence on Pattern Recognition*, pp. 23–26, 2004.

[139] D. Meger, P.-E. Forssen, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, and D. G. Lowe, "Curious george: An attentive semantic robot," *Robotics and Autonomous Systems* 2008 (in submission).

[140] G. Metta and P. Fitzpatrick, "Better vision through experimental manipulation," in *2nd International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, vol. 11, pp. 109–128, 2002.

[141] G. Metta and P. Fitzpatrick, "Early integration of vision and manipulation," *Adaptive Behavior*, vol. 11, no. 2, pp. 109–128, 2003.

[142] A. Mian, M. Bennamoun, and R. Owens, "Three-dimensional model-based object recognition and segmentation in cluttered scenes," *IEEE Transactions on PAMI*, vol. 28, no. 10, 2006.

[143] B. Micusik, H. Wildenauer, and J. Kosecka, "Detection and matching of rectangular structures," in *IEEE CVPR*, 2008.

[144] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[145] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensorymotor coordination to imitation," *IEEE Transactions on Robotics*, vol. 24, no. 1, pp. 15–26, 2008.

[146] J. Moren, A. Ude, A. Koene, and G. Cheng, "Biologically-based top-down attention modulation for humanoid interactions," *International Journal of Humanoid Robotics*, pp. 3–24, 2008.

[147] A. C. Murilo, J. Kosecka, J. J. Guerrero, and C. Sagues, "Visual door detection integrating appearance and shape cues," *Robotics and Autonomous Systems*, 2008.

[148] A. Namiki, K. Hashimoto, and M. Ishikawa, "A hierarchical control architecture for high-speed visual servoing," *International Journal of Robotic Research*, vol. 22, no. 10–11, pp. 873–888, 2003.

[149] V. Navalpakkam and L. Itti, "Sharing resources: Buy attention, get recognition," in *International Workshop on Attention and Performance in Computer Vision*, 2003.

[150] E. S. Neo, T. Sakaguchi, K. Yokoi, Y. Kawai, and K. Maruyama, "Operating humanoid robots in human environments," in *Workshop on Manipulation for Human Environments, Robotics: Science and Systems*, 2006.

[151] P. Newman and K. Ho, "SLAM-loop closing with visually salient features," in *IEEE ICRA*, pp. 644–651, 2005.

[152] K. Ogawara, J. Takamatsu, K. Hashimoto, and K. Ikeuchi, "Grasp recognition using a 3D articulated model and infrared images," in *IEEE International Conference on Intelligent Robots and Systems*, vol. 2, pp. 1590–1595, 2003.

[153] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," in *International Conference on Image Processing*, pp. 253–256, 2003.

[154] B. Olshausen, C. Anderson, and D. van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *Journal of Neuroscience*, vol. 13, pp. 4700–4719, 1993.

[155] PCCV. Performance Characterization in Computer Vision website, http://peipa.essex.ac.uk/benchmark/ index.html.

[156] M. Pechuk, O. Soldea, and E. Rivlin, "Learning function-based object classification from 3d imagery," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 173–191, 2008.

[157] L. Petersson, P. Jensfelt, D. Tell, M. Strandberg, D. Kragic, and H. I. Christensen, "Systems integration for real-world manipulation tasks," in *Proceedings of the IEEE International Conference on Robotics and Automation, ICRA 2002*, vol. 3, pp. 2500–2505, 2002.

[158] J. Piazzi and D. Prattichizzo, "An auto-epipolar strategy for mobile robot visual servoing," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 2, pp. 1802–1807, 2003.

[159] M. Piccardi, "Background subtraction techniques: A review," in *IEEE International Conference on Systems, Man and Cybernetics, vol. 4*, 2004.

[160] N. Pinto, D. D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard?," *PLoS Computational Biology*, vol. 4, p. e27, January 2008.

[161] A. Pinz, "Object categorization," *Foundations and Trends in Computer Graphics and Vision*, vol. 1, no. 4, pp. 255–353, 2006.

[162] P. Pirjanian, H. Christensen, and J. Fayman, "Application of voting to fusion of purposive modules: An experimental investigation," *Robotics and Autonomous Systems*, vol. 23, no. 4, pp. 253–266, 1998.

[163] M. Pressigout and E. Marchand, "Real-time hybrid tracking using edge and texture information," *International Journal of Robotics Research*, vol. 26, no. 7, pp. 689–713, 2007.

[164] P. Prokopowicz, R. Swain, and R. Kahn, "Task and environment-sensitive tracking," in *Proceedings of 1994 IEEE Symposium on Visual Languages*, pp. 73–78, 1994.

[165] V. Raos, M. Umilta, A. Murata, L. Fogassi, and V. Gallese, "Functional properties of grasping-related neurons in the ventral premotor area F5 of the macaque monkey," *Journal of Neurophysiology*, vol. 95, no. 2, pp. 709–729, 2006.

[166] C. Rasmussen and G. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Transactions on PAMI*, vol. 23, no. 6, pp. 560–576, 2001.

[167] G. Rizzolatti, L. Fadiga, M. Matelli, V. Bettinardi, E. Paulesu, D. Perani, and F. Fazio, "Localization of grasp representations in humans by PET: 1. Observation versus execution," *Experimental Brain Research*, vol. 111, no. 2, pp. 246–252, 1996.

[168] G. Rizzolatti, L. Fogassi, and V. Gallese, "Parietal cortex: From sight to action," *Current Opinion in Neurobiology*, vol. 7, pp. 562–567, 1997.

[169] G. Rizzolatti, G. Luppino, and M. Matelli, "The organization of the cortical motor system: New concepts," *Electroencephalography and Clinical Neurophysiology*, vol. 106, no. 4, pp. 283–296, 1998.

[170] J. Rosen and B. Hannaford, "Doc at a distance," *IEEE Spectrum*, vol. 43, no. 10, pp. 34–39, 2006.

[171] C. Sagüés and J. Guerrero, "Visual correction for mobile robot homing," *Robotics and Autonomous Systems*, To appear, 2005.

[172] A. Sanderson and L. Weiss, "Image-based visual servo control using relational graph error signals," in *Proceedings of the IEEE*, pp. 1074–1077, 1980.

[173] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.

[174] S. Schaal, "Is imitation learning the route to humanoid robots?," *Trends in Cognitive Sciences*, vol. 3, no. 6, pp. 233–242, 1999.

[175] C. Schmid, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, pp. 2169–2178, 2006.

[176] Semantic Robot Vision Challenge, http://www.semantic-robot-vision-challenge.org/.

[177] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Recognition with cortex-like mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, 2007.

[178] A. Shahdi and S. Sirouspour, "Adaptive/robust control for time-delay teleoperation," *IEEE Transactions on Robotics*, vol. 25, no. 196-205, p. 1, 2009.

[179] P. Sharkey and D. Murray, "Delays versus performance of visually guided systems," in *IEE Proceedings of Control Theory & Applications*, vol. 143, no. 5, pp. 436–447, 1996.

[180] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the IEEE Computer Vision and Pattern Recognition, CVPR'94*, pp. 593–600, 1994.

[181] T. Shipley and P. Kellman, "Advances in psychology: Form fragments to objects," *Elsevier Science B.V.*, vol. 130, 2001.

[182] Y. Shirai and H. Inoue, “Guiding a robot by visual feedback in assembling tasks,” *Pattern Recognition*, vol. 5, pp. 99–108, 1973.

[183] Y. Shirai, R. Okada, and T. Yamane, “Robust visual tracking by integrating various cues,” in *Robust Vision for Manipulation*, (M. Vincze and G. Hager, eds.), pp. 53–66, Spie/IEEE Series, 2000.

[184] H. Sidenbladh, D. Kragic, and H. I. Christensen, “A person following behaviour for a mobile robot,” in *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 670–675, 1999.

[185] R. Sim, P. Elinas, M. Griffin, and J. J. Little, “Vision-based slam using the rao-blackwellised particle filter,” in *IJCAI Workshop on Reasoning with Uncertainty in Robotics*, July 2005.

[186] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, “Discovering objects and their localization in images,” in *International Conference on Computer Vision*, pp. 370–377, 2005.

[187] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *International Conference on Computer Vision*, pp. 1470–1477, 2003.

[188] S. Smith, “ASSET-2 - Real-Time Motion Segmentation and Object Tracking,” Tech. Rep. TR95SMS2b, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB), Department of Clinical Vision and Image Processing Group, DRA Chertsey, DERA, UK, 1995.

[189] F. Solina and R. Bajcsy, “Recovery of parametric models from range images: The case for superquadrics with global deformations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 131–147, 1990.

[190] T. Southey and J. J. Little, “Object discovery using motion, Appearance and shape,” in *AAAI Cognitive Robotics Workshop*, 2006.

[191] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele, “Functional object class detection based on learned affordance cues,” in *ICVS–International Conference on Computer Vision Systems*, pp. 435–444, 2008.

[192] A. Stoytchev, “Behavior-grounded representation of tool affordances,” in *IEEE International Conference on Robotics and Automation*, pp. 3060–3065, 2005.

[193] M. Suppa, S. Kielhoefer, J. Langwald, F. Hacker, K. H. Strobl, and G. Hirzinger, “The 3d-modeller: A multi-purpose vision platform,” in *Proceedings of International Conference on Robotics and Automation*, 2007.

[194] M. Sutton, L. Stark, and K. Bowyer, “Function from visual analysis and physical interaction: A methodology for recognition of generic classes of objects,” *Image and Vision Computing*, vol. 16, pp. 746–763, 1998.

[195] J. Tardós, J. Neira, P. Newman, and J. Leonard, “Robust mapping and localization in indoor environments using sonar data,” *IJRR*, vol. 4, 2002.

[196] M. J. Tarr and H. H. Blthoff, “Image-based recognition in man, monkey, and machine,” *Cognition*, vol. 67, pp. 1–20, 1998.

[197] G. Taylor and L. Kleeman, “Fusion of multimodal visual cues for model-based object tracking,” in *Australiasian Conference on Robotics and Automation*, Brisbane, Australia, 2003.

[198] G. Taylor and L. Kleeman, “Robust range data segmentation using geometric primitives for robotic applications,” in *Proceedings of the 9th International Conference on Signal and Image Processing*, pp. 467–472, 2003.

[199] S. Thrun, D. Fox, and W. Burgard, “A probabilistic approach to concurrent mapping and localization for mobile robots,” *Autonomous Robots*, vol. 5, pp. 253–271, 1998.

[200] S. Thrun, Y. Liu, D.Koller, A. Ng, Z. Ghahramani, and H. Durrant-White, “SLAM with sparse extended information filters,” *IJRR*, vol. 23, no. 8, pp. 690–717, 2004.

[201] K. Toyama and G. Hager, “Incremental focus of attention for robust vision-based tracking,” *International Journal of Computer Vision*, vol. 35, no. 1, pp. 45–63, 1999.

[202] A. Treisman and G. Gelade, “A feature integration theory of attention,” *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[203] J. Triesch and C. V. der Malsburg, “Self-organized integration of adaptive visual cues for face tracking,” in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 102–107, 2000.

[204] M. Trivedi and T. G. J. McCall, “Looking-in and looking-out of a vehicle: Selected investigations in computer vision based enhanced vehicle safety,” in *IEEE International Conference on Vehicular Electronics and Safety*, pp. 29–64, 2005.

[205] D. Tsakiris, C. Samson, and P. Rives, “Extending visual servoing techniques to nonholonomic mobile robots,” in *The Confluence of Vision and Control*, vol. 1, (G. Hager, D. Kriegman, and S. Morse, eds.), Lecture Notes in Control and Information Systems, Springer-Verlag, 1999.

[206] J. Tsotsos, “On the relative complexity of passive vs active visual search,” *International Journal of Computer Vision*, vol. 7, no. 2, pp. 127–141, 1992.

[207] A. Ude, C. Gaskett, and G. Cheng, “Foveated vision systems with two cameras per eye,” in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3457–3462, 2006.

[208] T. Uhlin, *Fixation and Seeing Systems.* PhD thesis, NADA, Royal Institute of Technology, KTH. 1996.

[209] L. Vacchetti, V. Lepetit, and P. Fua, “Stable real-time 3D tracking using online and offline information,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 1385–1391, 2004.

[210] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, “Evaluation of color descriptors for object and scene recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, June 2008.

[211] M. Vincze, “On the design and structure of artificial eyes for tracking tasks,” *Journal of Advanced Computational Intelligence and Intelligent Informatics JACIII*, vol. 9, no. 4, pp. 353–360, 2005.

[212] M. Vincze, M. Ayromlou, and W. Kubinger, “An integrating framework for robust real-time 3D object tracking,” in *International Conference on Computer Vision Systems, ICVS’99*, pp. 135–150, 1999.

[213] M. Vincze and G. Hager, *Robust Vision for Vision-Based Control of Motion.* IEEE Press, 2000.

[214] M. Vincze, M. Zillich, W. Ponweiser, V. Hlavac, J. Matas, S. Obdrzalek, H. Buxton, J. Howell, K. Sage, A. Argyros, C. Eberst, and G. Umgeher, "Integrated vision system for the semantic interpretation of activities where a person handles objects," *CVIU*, vol. 113, no. 6, pp. 682–692, June 2009.

[215] M. Vinzce, M. Ayromlou, M. Ponweiser, and M. Zillich, "Edge projected integration of image and model cues for robust model-based object tracking," *International Journal of Robotics Research*, vol. 20, no. 7, pp. 533–552, 2001.

[216] I. Weiss and M. Ray, "Model-based recognition of 3d objects from single images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 116–128, 2001.

[217] F. Wörgötter, A. Agostini, N. Krüger, N. Shylo, and B. Porr, "Cognitive agents — A procedural perspective relying on the predictability of object–action–complexes," *Robotics and Autonomous Systems*, vol. 57, no. 4, pp. 420–432, 2009.

[218] S. Wrede, C. Bauckhage, G. Sagerer, W. Ponweiser, and M. Vincze, "Integration frameworks for large scale cognitive vision systems–an evaluative study," in *Proceedings of the 17th International Conference on Pattern Recognition ICPR*, pp. 761–764, 2004.

[219] P. Wunsch and G. Hirzinger, "Real-time visual tracking of 3-d objects with dynamic handling of occlusion," in *IEEE International Conference on Robotics and Automation, ICRA'97*, pp. 2868–2873, Albuquerque, New Mexico, USA, April 1997.

[220] M. Zillich and M. Vincze, "Anytimeness avoids parameters in detecting closed convex polygons," in *IEEE Computer Society Workshop on Perceptual Organization in Computer Vision at CVPR*, 2008.

[221] A. Zisserman, D. Forsyth, J. Mundy, C. Rothwell, J. Liu, and N. Pillow, "3D object recognition using invariance," *Artificial Intelligence*, vol. 78, no. 1–2, pp. 239–288, 1995.