International Journal of Humanoid Robotics

Towards Grasp-Oriented Visual Perception for Humanoid Robots

Jeannette Bohg, Carl Barck-Holst, Kai Huebner, Maria Ralph, Babak Rasolzadeh, Dan Song, Danica Kragic

Computational Vision and Active Perception Laboratory
Centre for Autonomous Systems
Royal Institute of Technology, Stockholm, Sweden
{*bohg,barck,khubner,mralph,babak2,dsong,danik*}*@csc.kth.se*

Received 30 November 2008
Accepted 4 May 2009

A distinct property of robot vision systems is that they are embodied. Visual information is extracted for the purpose of moving in and interacting with the environment. Thus, different types of perception-action cycles need to be implemented and evaluated.

In this paper, we study the problem of designing a vision system for the purpose of object grasping in everyday environments. This vision system is firstly targeted at the interaction with the world through recognition and grasping of objects and secondly at being an interface for the reasoning and planning module to the real world. The latter provides the vision system with a certain task that drives it and defines a specific context, i.e. search for or identify a certain object and analyze it for potential later manipulation. We deal with cases of: (i) known objects, (ii) objects *similar* to already known objects, and (iii) unknown objects. The perception-action cycle is connected to the reasoning system based on the idea of *affordances*. All three cases are also related to the state of the art and the terminology in the neuroscientific area.

Keywords: Perception, Attention, Reasoning, Neuroscience, Grasping, Affordances

1. Introduction

Within robotics, particularly humanoid robotics, there is significant interest in linking low-level sensory information such as visual, auditory, or haptic feedback to higher-level symbolic representations in order for an agent to reason how it should act in the world. This coupling of perception and action, often referred to as the perception-action cycle, requires that an agent first senses or perceives the environment in order to find places or things of interest to act on. For instance, an agent can attend to specific objects that it must grasp in a particular way in order to complete some task.

At present there has been increased interest in developing robots for both commercial and personal use. However, how these types of robots interact with the world (i.e. human users) depends significantly on how they perceive and act in their surroundings. Robots that are designed to interact with humans cannot assume these types of users to be technically inclined. As such, robotic systems must now go beyond the traditional perception-action cycle which has been presented over the past few decades, and introduce more advanced

cognitive reasoning and planning capabilities. However, in order for robots to learn from humans, or to learn through their own exploration of the world, they need to be able to share their understanding and representation of the world as their human counterparts.

Although there has been a significant amount of work presented which focuses on the perception-action cycle from a number of different perspectives, our motivation is to explore this approach from the perspective of human-robot and robot-object interaction. How robots engage in interactions with users and learn from those interactions provides valuable insights into designing robotic systems capable of perceiving, reasoning, planning, and acting effectively. The questions that remain however are: (i) how can a robot's perceptual system both attend to and recognize objects in the world, (ii) how can we use this information to select appropriate actions (i.e. grasp types), (iii) how can we represent low-level sensory information using high-level symbolic representations, and (iv) how can these symbols be used to reason about the appropriate actions a robot should perform. As a first step towards answering these questions we can examine on-going work within the human-robot interaction community that explores how robots can learn to perceive the world, and how they can act appropriately according to user expectations.

Firstly, Gray et al.[1] present a cognitive architecture for a robot named Leonardo, designed to interact with human users within a social context in order to learn how to detect objects of interest and what to do with those objects once found. Using a button pressing scenario, Leonardo is used in a series of experiments as an assistant to a human user. The user is asked to press one of three buttons located on a table. One button is occluded from the user. The second button is operational, and the final button is locked so that a pushing down action cannot be performed. The robot first attends to points of interest in its visual stream. In this case it is the user's hand near a button that is the focus of the perception system. The reasoning system developed uses reinforcement learning to teach the robot what events should take place and when, and to reason as to why certain actions did not execute as expected. From here, based on its prior experiences, the robot can infer (i.e. reason) as to why the user is having difficulty with the task and provide possible solutions to help the user. Leonardo's reasoning system reasons at the symbolic level, where low-level sensory information in the form of visual data is used to understand what the user is doing or trying to do.

Other work focused on linking perception and action has been presented by Kyriacou et al.[2]. In their work a mobile robot engages in instruction-based learning to learn routes between different locations in a miniature town. The perception system uses stereo vision to extract images acquired in its visual stream and assign features from those images to specific road landmarks found in the environment. From here the robot builds a map of the sequence of road landmarks in order to construct a route map from one location to another. When the robot is asked to traverse the town independently, it perceives these landmarks in order to trigger the appropriate set of actions to take (i.e. turn left, go straight, etc.).

However, we are particularly interested in the perception-action cycle as it applies to robotic grasping. Within this domain, McGuire et al.[3] for example have developed a system for attending to pointing hands in reference to objects of interest. First, the robot's vision system focuses on human hand gestures (i.e. pointing) to draw the robot's attention to

specific objects. From here control is then transferred to the robot hand for a grasp attempt while maintaining a steady stream of visual feedback. Once grasped a second location pointed to by the user is used as the goal or target endpoint for the object's final placement.

It is thereby exemplified that symbolic representations, like buttons, road landmarks or gestures, as objects in general, play an important role. The cognitive robot's environment is built by such objects that are ought to be recognized, classified, interpreted or manipulated. Though also low-level sensory features, denoted here as *things*, may help for some tasks, semantic representations of *objects* are more valuable or even necessary in others. Also from the human point of view, and therefore from the perspective of a human-robot discourse, symbolic representations are highly reasonable. Nevertheless, questions arise of how these symbols do look like, i.e. what makes an object an object, what makes a cup being a cup, both for the human and for the cognitive robot?

In the field of robotics and computer vision, plenty of research has been concentrated on example-based recognition of objects by learned appearance models. In such systems, a cup can be recognized after it has been shown to the robot. In this process, the robot creates an internal model of this particular cup. Understandably, it will not be able to identify any arbitrary, unseen cup that differs from the learned model. In such a model-based philosophy two cups are as different from each other as a cup from a bottle. Over the last years, terms of affordances have therefore moved into focus of interest, i.e. what does a cup or a bottle afford an agent to do, e.g. filling it or carrying any fluid in it. From this viewpoint, cups and bottles are not anymore that different from each other, but have to be described according to their properties in close relation to the actions connected to them. Therefore, this approach requires that the perception of object properties and actions must be somehow intertwined.

In this paper we examine the perception-action cycle in more detail by discussing how we use attention and perception to reason about appropriate grasp choices. The first contribution of the paper is that the perception-action cycle is studied from a both a neuroscience and robotic systems perspective. Second, we study three cases of dealing with i) known objects, ii) objects *similar* to already known objects, and iii) unknown objects. Finally, the perception-action cycle is connected to the reasoning system based on the idea of *affordances*.

The paper is organized as follows. Section 2 will discuss the system design in more detail and provides an overview of the perception-action cycle from a both a neuroscience and robotic systems perspective. This section acts as the biological motivation behind the work we present in this paper. Section 3 presents the perception and attention systems developed. Section 4 focuses on both low and high-level data representation and symbolic reasoning. Section 5 outlines an example scenario which could be implemented on a humanoid robot to illustrate how each of the components in the perception-action cycle work together to achieve a target objective. Section 6 concludes our discussion and outlines future work.

2. From Neuroscience to System Design

Human and primates are able to manipulate every kind of object dexterously, and to adapt their motor programs flexibly and swiftly according to environmental and task require-

ments. This is a fundamental skill that has been pursued by roboticists for many years. Understanding how the brain controls grasp and manipulation tasks in humans and primates can provide great inspiration and motivation for developing efficient and effective artificial beings. In this section, we will firstly present a neuroscience overview focused on (i) visual-guided control in humans and other primates, (ii) the computational models of cortical mechanisms in grasp control, and (iii) the recent attempts in robotic implementation. Finally, based upon the advances along this path, the design of a system for vision-based grasping and grasp-oriented visual perception will be described.

2.1. *Neuroscience*

Unlike locomotion and reaching, grasping and manipulation tasks are highly interactive with the environment; it requires direct access of visual information to extract object properties for grasp planning and execution, and the gradual build-up of semantics from experiences for reasoning and improved movement planning. Recent neuroscientific findings show that such tasks are realized through distributed information flows between multiple regions within the nervous system[4,5,6,7] with specific attention paid to the role of frontal-parietal interaction and its relation to the visual cortex.

Spatial Attention Mechanism. Firstly, there is a visuospatial attention mechanism that directs our gaze towards the most interesting part of the visual field. Spatial attention is physiologically non-unitary: we shift our gazes to the 'pop-up' part of a scene, towards an identified object, or to the target of intended action. In other words, attention can be controlled through a bottom-up process (or stimulus-driven) and/or through a top-down process (or goal-driven). Both control mechanisms have their neural correlates.

An attention map existing in primary visual cortex, V1, ranks the saliency levels of various components in a visual space[8], thus substantiating the bottom-up mechanism. The projections of both the dorsal and ventral visual streams provide action and identification guided attention shifts[9]. Furthermore, neurons in F7 receive task-relevant information from the prefrontal cortex[10], and control eye movement towards task-intended targets in the frontal eye field[11]. Taken together, the attention mechanism plays an essential role in grasp-oriented visual perception in the sense that it integrates the desired goals or motor intentions of an individual with the intrinsic properties of a visual field.

Dorsal and Ventral Visual Pathways. Human visual processing is characterized by a two-mode dichotomy: locating and identifying systems[12]. Both systems, in the human or primate brains, originate in the basic visual areas V1 and V2, where simple but consistent visual features, such as edges, corners, textures, and basic orientation information are processed. The locating system extends dorsally towards the caudal intraparietal sulcus (CIP). CIP receives basic visual information such as object edges and surfaces and produces grasp-related object properties such as object dimensions, 3D orientation, object shape and curvature[13]. The information then reaches the anterior intraparietal sulcus (AIP) where visuomotor transformations occur and a set of grasp configurations are selected[14].

The identifying systems extends from V1 and V2 ventrally to the inferior temporal cortex (IT). Neurons in IT encode object-centered descriptions or object identities that are independent of viewing conditions, thus the ventral stream is dedicated to object recognition [15]. It also stores memories of previous interactions with the target objects [16].

The nature of processed visual data along the two pathways suggests an action-perception dissociation along the two systems [9], with the object attributes processed in the dorsal stream subjected to a 'pragmatic' mode, whereas the one along the ventral stream to a 'semantic' mode. However, the division of the labor is not absolute [9]. Both behavioral data [9,17] and neural anatomic studies (rich projection from IT to AIP [18]) suggest that the semantics of the objects helps the pragmatic system in the action selection process, providing graspable properties and afforded actions from the knowledge learned during past events.

Reasoning and Planning. The grasp affordances produced in AIP are then passed to area F5 (or premotor cortex) which contains the movement primitives for composing grasping actions. The primary motor cortex, M1 or F1 is then responsible for sending muscle commands for the grasp execution. It should be noted that, the final selection of the grasp configuration in F5 is also constrained by the intention of an individual: what the agent of the action wants to do with the object. Thus a complete visuomotor transformation would also need information from the circuits where high-level task decisions are made (prefrontal lobe) [6,19]. This concept is strongly supported by the abundant neural connections between the AIP-F5 circuits and the prefrontal cortex [20,6,19] where reasoning and cognitive functions, complex task processing and working memory are hosted [10].

Movement Execution and Haptic Feedback. Motor plans such as specified grasp configurations in F5 are passed to the primary motor cortex (F1 or M1) for movement execution. Beside inputs from F5, F1 also receives connections from other premotor and somatosensory areas, and outputs muscle commands to control arm and hand motion [5]. As the hand interacts with the world, tactile sensory information is processed in the somatosensory areas (such as SII) of the cerebral cortex. Neurons in SII are believed to encode the higher-order tactile feedback [21] that is used for representing the shape and surface properties of the target objects. In addition, the rich connection between neurons in SII with AIP-F5 circuit [22] suggests that SII may generate a tactile expectation consonant with the grasp configurations selected in AIP-F5 circuit, so that any discrepancy between expectation and contact-based feedback will further refine the selection of grasp configurations or affordances.

2.2. *Computational Models of Visual-Guided Action*

FARS Model on Grasp Action-Execution. Many attempts to emulate the neural computation of the human perception-action cycle for grasping tasks have been carried out. In 1998, Fagg and Arbib [6] proposed the 'FARS' model, the most complete attempt at the time to simulate the neural cortical processes involved in the generation and execution processes of grasping tasks. It is centered on the AIP-F5 circuit, and also includes a variety

of supporting areas in the parietal and frontal cortices. AIP receives visual information of the objects from the two visual streams, and provides F5 with multiple affordances. F5 takes in the multiple affordances and selects the desired motor prototype on the basis of prefrontal inputs (through F6) that signals to F5 the object meaning and the current goals of the individual. The decision of F5 is fed back to AIP to reinforce the selected grasps upon the currently actuated grasp actions. FARS also includes the grasp execution cycle (F1-Hand-SII). This process provides both tactile expectations and tactile sensations so that any discrepancy between the two can trigger reprogramming of the grasp in F5.

This model provided accurate predictions on recorded neural firing patterns, and it advanced our understanding of information encoding in primate's cortices when controlling visual-guided grasp tasks. It is noted, however, that the FARS model is primarily focused on the *action-execution* process. It assumes that the model computation starts from AIP, and hence the detailed visual processing along the two visual pathways are not represented. Moreover, FARS is a neuro-computational model whose main purpose is to aid in further understanding of cortical mechanisms, as such, the link to the robotic application is lacking.

Vision-Based Grasping Model. To solve these problems, recently, Recatalá et al.[23] developed a model of vision-based grasping (VBG) following a *sense-plan-act* paradigm. The system is based on the early separation and late integration of visual analysis through the two visual streams. In addition, the visual reconstruction is driven through a top-down attention system that selectively chooses the regions of the object considered more interesting for grasping. The model is implemented using a 'filter-based architecture' (FBA) that is specifically used for adaptation of neuroscience models to the robotic setup.

The VBG model has been successfully implemented in a set of grasp synthesis tasks on a robotic platform. The limitation of the work, however, lies in three aspects. Firstly, the experiments presented are currently limited to the visuomotor processing in the dorsal stream. Secondly, the grasp execution in the model relies only on visual information, however the tactile inputs from finger-object contact are essential for stable grasps in the real world[24]. Finally, there is a lack of high-level task-related inputs from the prefrontal cortex, thus, although being relatively an autonomous system for single-object grasping tasks, the system is expected to be less effective in dealing with dynamic and complex environments.

2.3. *System Design of Grasp-Oriented Visual Perception*

The design of our proposed grasp-oriented visual perception system (GOVP) is motivated and inspired by neuroscience findings and computational models described above. On the one hand, vision plays an important role in extracting information from a scene in order to perform grasp actions in the world; on the other hand, the requirements of a grasping task direct visual processing towards the intended features on the object or locations in the world. This perception-action coupling forms the foundation of an *active vision* system, in our case, dedicated to the visual processing towards goal-directed grasping tasks.

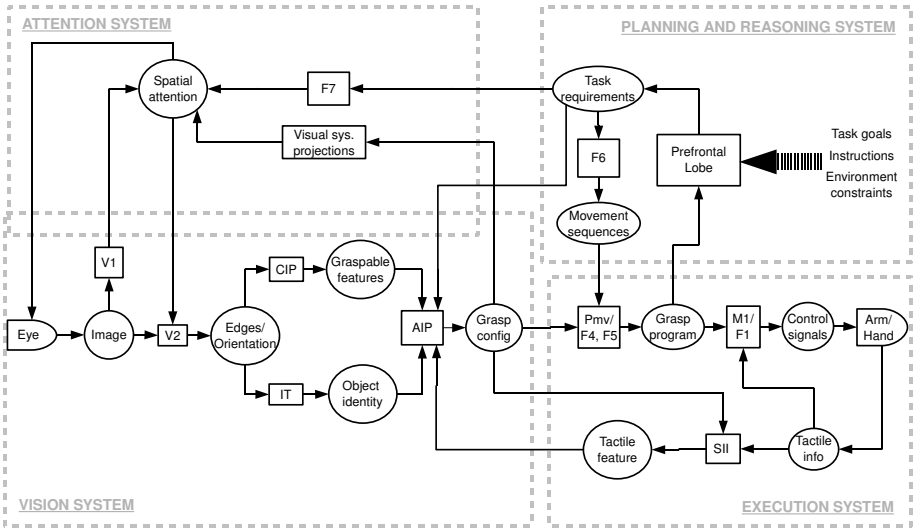


Fig. 1. The proposed grasp-oriented visual perception (GOVP) system architecture.

Filter-Based Architecture. The GOVP system, as shown in Fig. 1, is constructed using a filter based architecture (FBA) initially proposed by Recatalá et al.[23]. It consists of three types of basic components: (i) hardware components (bullet-shape) including sensors and actuators, (ii) virtual filters (rectangular) that handle operations such as feature extraction or a control law, and (iii) data sets (ellipsoid) that store groups of data produced and processed by the above modules. A task is realized through a set of connected model components that are simultaneously active, where the data sets constitute an internal, non-centralized memory spreading along the chain of processors. The formation specifies the set of interfaces between the model components, and thus signifies a clear input-output information flow through the entire system. It also allows grouping a set of components serving a more abstract functionality into a larger module or subsystem, thus aiding the translation from individual brain functions into robotic implementations.

System Structure. The proposed system follows a *sense-reasoning-plan-act* paradigm which combines the previous FARS[6] (*action-perception*) and VBG[23] (*sense-plan-act*) models. On top of it, we add subsystems that both models are lacking, namely, a prefrontal mental model that performs cognitive reasoning and planning of complex motor tasks (reasoning system), and a high-level attention system that actively controls the focus of visual processing to the intended locations of the world. As a result, GOVP consists of four subsystems: attention, vision, reasoning, and execution. Each of the four modules can function independently, but they also integrate coherently in a visual-guided grasping cycle to continuously enrich and improve the robot's knowledge of the world.

Firstly, the attention system plays an important role in closing the loop of the entire GOVP system. In an 'active' visual perception system, the fixation of the eye (or a camera) and the subsequent visual analysis are usually directed towards the most interesting

component in a visual field. The ranking of such 'interestingness' is found, as reviewed in Section 2.1, to be done through two control mechanisms: bottom-up (or stimulus-driven) and top-down (or goal-driven). The attention system in GOVP includes both control mechanisms, being the bottom-up control provided by a saliency map (V1) of the visual field, and top-down control driven by the object identities (vision system) and task requirements (prefrontal cortex). The detailed robotic implementation is presented in Section 3.2.

Secondly, in the vision system, we model the early separation and later integration of dorsal and ventral streams similar to the VBG model[23]. Along the dorsal stream, grasp-relevant visual features of an object are extracted and processed (CIP) in order to select and evaluate (AIP) a set of appropriate grasp configurations (or affordances). At the same time, the object identification process that is performed along the ventral pathway can further bias the grasp selection through its memory of previous experiences. These processes are implemented later in this paper by two robotic vision methods (see Section 3.3 and 3.4).

Thirdly, the planning and reasoning module in the Prefrontal-F6 circuit introduces a high-level cognitive function of the human brain into the system. The system receives the task goal, instructions and environmental constraints as the overall system inputs. In addition, it takes in the symbolic visual and motor representations abstracted through the perception-action cycle. Then, through a series of mental inferencing based internal logic learned by experiences (prefrontal lobe), the reasoning system outputs task requirements which in turn determines the appropriate motor sequences (F6). The task requirements can also aid the selection of grasp configurations or affordances for a specific object (AIP). At the same time, the task information shifts the visuospatial attention towards the goal-related areas of the world. The reasoning system thus provides the system with high-level autonomy, i.e. it is able to handle, on its own, the plan and execution of specified tasks with a high degree of complexity. A Bayesian network is used to implement the reasoning system which is described later in Section 4.

Finally, the selected movement primitives, in this case the grasp types, are then programmed (F5) and executed (M1/F1) on the robotic actuators (arm and hand). The actual contact with the object provides tactile as well as visual sensory information for the system to learn whether the selected grasp configuration is successful or not. Such feedback can in turn reinforce the selected affordances provided that the actual grasp is successful, or eliminate the ones when failure occurs. These experiences are stored in the robot's working memory (in prefrontal lobe) for future reasoning and planning. They also update the semantic attributes of the objects processed in the ventral pathway of the vision system. In the current paper, the execution system is simply implemented as an open-loop robot controller that does not involve the haptic based feedback control and learning. But one of our ongoing developments is to extract graspable features of the objects through haptic exploration, which will be later integrated to complete the GOVP system.

In this section, we reviewed the recent neuroscience findings and computational models regarding the cortical mechanisms in controlling visual-guided grasping, and designed a grasp-oriented visual perception (GOVP) system. We believe that this implementation on a robotic system can not only help validate the neuroscience hypotheses, but also that robots can be endowed with advanced perception and grasping capabilities typical of primates.

3. Visual Perception

Similar to the human vision system, but unlike many systems from the computer vision community, robotic vision systems are embodied. Recent works exhibiting this are presented by Ude et al.[25] and Björkman and Eklundh[26]. Here, vision is embodied in a robotic system capable of visual search as well as simple object manipulation.

In this section, we will present such a system developed not as an isolated entity but as part of a larger system comprising hardware and its controllers and also reasoning and planning modules. It is firstly targeted at the interaction with the world through recognition and manipulation of objects and secondly at being an interface for the reasoning and planning module to the real world. The latter will provide the vision system with a certain task that drives it and defines a specific context, i.e. search for or identify a certain object, maybe analyze it for potential later manipulation.

Manipulable objects can either be previously known or completely new to the system. Even if confusion does occur frequently, a human being is able to immediately divide the perceived world into different physical objects, seemingly without effort. The task is performed with such ease that the complexity of the operation is easily underestimated. There are two possibilities for a robotic system to carry out this task, i.e. to form hypotheses about entities from a visual percept. Either they are defined based on common properties such as proximity and appearance, or they are similar to previously known objects. The resulting perceptual entities might or might not correspond to unique physical objects in 3D space. It is not until the robot acts upon an entity, that the hypothesis about a physical object can be verified. Without interaction the entity has no real meaning to the robot. We call these entities *things* to differentiate them from *objects* that are known to be physical objects, through interaction or other means. As soon as such an object emerges from a thing, we can provide the reasoning and planning modules with a *symbol* as a base for further actions.

Motivation and Structure. In this section, we will present a vision system that provides the means for realizing such a cycle starting from reasoning/planning, going through perception to action and back. The flow of information through it is summarized in Fig. 2. The single components are outlined in the following.

Firstly, we will introduce the attention component in Section 3.2 that enables the whole visual system to deal with the overwhelming amount of perceptual data.[27,28,29] Since resources will always be limited in one way or the other, there is a need for a mechanism that highlights the most relevant information and suppresses stimuli that is of no use to the system. Instead of performing the same operations for all parts of the scene, resources should be spent where they are needed. Relevancy is not a static measure, but depends on context, on the scene in which the robot acts and the tasks the robot is performing. Consequently, there is a need for the attentional system to adapt to contextual changes. This component presented here is able to attend to and fixate on *things* in the scene. To facilitate object manipulation and provide an understanding of the world, there is support for figure-ground segmentation, recognition and pose estimation.

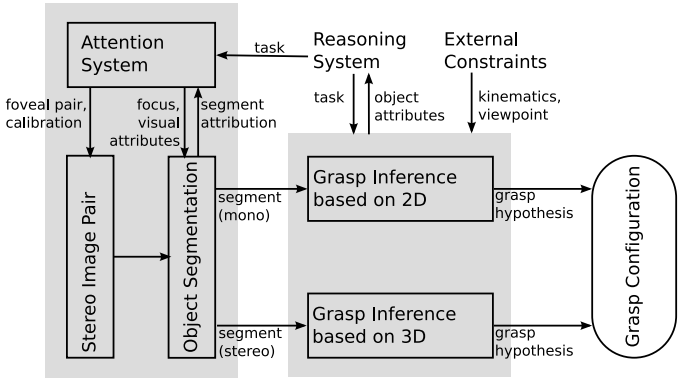


Fig. 2. Structure of the visual perception block of the system and the information flow between its components.

As a subsequent step to fixation on things we want to enable their manipulation in order to confirm hypotheses about physical objects. As discussed in Section 2, there are two main visual pathways that are commonly associated with a 'pragmatic' mode (dorsal stream) and a 'semantic' mode (ventral stream). While the former is claimed to be involved in action selection, the latter is usually seen to be related to object identification.

We see three different ways to employ these two visual streams for the purpose of grasping. Firstly, we have the case in which the task is to manipulate a known object. After this object has been identified, experience gathered from previous interaction can be directly applied. This process is purely based on information processed in the ventral stream. An example for a robotic application that applied this methodology was introduced in our previous work[30] and also by Morales et al.[31].

The second case to employ the two visual pathways is given when the task is to manipulate a novel object that is similar to an object previously encountered. Known graspable properties and features that are recognized on the novel object help the pragmatic system to choose appropriate actions. A robotic example for such a system will be described in Section 3.3. It employs offline learned models of graspable features to detect prehensile points in monocular images of objects.

The task of grasping completely unknown objects without the possibility to exploit previous experience is the last case of visually guided manipulation discussed in this paper. Here, purely the dorsal stream is involved in extracting grasp relevant object features that facilitate the inference of grasp hypotheses. A robotic system that is related to this case is introduced in Section 3.4.

3.1. *Related Work*

Attention Systems. Seen either from a robotics or a biological perspective, attention can be thought of a selection mechanism that precedes the higher level tasks such as object recognition and manipulation. The biological studies on visual attention show that in biological

systems there is a subconscious, automatic ranking of the 'interestingness' of the different components of a visual scene. How this ranking is done depends on the objectives of the observer (top-down), as well as the relationships between the different components of the scene (bottom-up) [32,8]. It has been shown that in humans the finally attended region is selected by dynamic (temporal) alternations of synaptic connectivity, under both top-down (task dependent) and bottom-up (scene dependent) control[33].

Todays computational models of this intricate process thus assume the bottom-up mechanism as a fast process that biases the observer toward stimuli based on their saliency (encoded in terms of center-surround mechanisms) and the top-down mechanism as a slow process with variable selection criteria, which direct works under cognitive, volitional control[34]. The first uses of these computational modes in computer vision, were attentive processing for scene analysis. A majority of them still today are salience based models, following Treisman and Gelade[34] and the influential model of Koch and Ullman[35]. The extensions in recent years have mainly focused on the fusion of top-down and bottom-up computational processes.

There are approaches[36] using a stochastic Winner-Take-All (WTA) network to create a variable saliency based search model that enabled them looking for particular saliency structures. Yet another WTA approach[37] uses game theory on the statistics of foreground and background to search for objects. In an interesting and novel approach, Choi et al.[38] suggest learning the desired modulations of the saliency map, based on the Itti and Koch model[39], for top-down tuning of attention, with the aid of an ART-network. Navalpakkam and Itti[40] take the idea further by enhancing the bottom-up salience model to learn target objects from training images containing targets in diverse, complex backgrounds.

In a similar 'tuning approach', using an Interactive Spiking Neural Network, Lee et al.[41] bias the bottom-up processing towards a task (in their case in face detection). One major drawback of their model, compared to that of Navalpakkam and Itti[40], was that it could not learn the influence of context, i.e. how changing the background might alter the tuning that is needed for a particular search task. Many works since, such as that of Oliva et al.[42], have shown that information from visual context can indeed modulate the saliency of image regions during the task of object detection. This is usually done by learning the relationship between the context features and the top-down (task-based) tuning of the saliency. However, this has yet not been used together with a bottom-up, top-down fusing model. Frintrop[43] propose the VOCUS-model, that contains two versions of the saliency map: a bottom-up (similar to that of Itti et al.[39]) and a top-down (tuned through learning). The saliency maps are then linearly combined using a fixed weight. An obvious drawback is that this makes the combination rigid and non flexible, which may result in loss of important bottom-up information.

There are today still very few computational models of attention designed for being used in a 'active vision' scenario, e.g. for a service robot. With the recent exception of the work of Moren et al.[44], where they use a top-down attention model on a humanoid platform, most of the models were designed in order to aid the study of (biological) visual attention itself. In this work our attention model aims to fill this obvious need.

Grasp Inference Systems. For grasping, numerous approaches and concepts have been developed over the last decades. Designing grasping systems and planning grasps is difficult due to the large search space resulting from all possible hand configurations, grasp types, and object properties that occur in regular environments.

A grounded theory on stable contact-level grasps has been developed in the literature, [45,46] In this theory of grasp planning, finger contact locations, forces and grasp wrench spaces can be simulated. Different criteria can be defined to rate grasp configurations, e.g. force closure, dexterity, equilibrium, stability and dynamic behavior[46].

Based on this theory a number of approaches were developed that try to limit the amount of candidate grasps and thus prune the search tree for finding the most stable grasp. Ciorcarlie et al.[47] exploit results from neuroscience that show that human hand control takes place in a much lower dimensionality than the actual number of its degrees of freedom. This finding is applied to directly reduce the configuration space of a robotic hand to find pre-grasp postures from which the system searches for stable grasps. In the work by Borst et al.[48] the number of candidate grasps is reduced by random generation dependent on the object surface. It is shown that this approach works well if the goal is not to find an optimal grasp but instead a fairly good grasp that works well for 'everyday tasks'.

The above mentioned approaches[47,48] are all developed and evaluated in simulation. In our previous work Ekvall and Kragic[30] and in the work by Morales et al.[31] real and simulated data are combined for the purpose of grasping known objects, i.e. a 3D model is available. In a monocular image a known object is recognized and its pose within the scene is estimated. Given that information, an appropriate grasp configuration can be selected from a grasp experience database. This database was acquired offline beforehand through simulation of grasps on 3D models of a set of such known objects. While Ekvall and Kragic[30] still apply the selected grasp in simulation, Morales et al.[31] ported this approach to the platform as described in Asfour et al.[49].

However, the dependency on a-priori known or dense and detailed object models is apparent. This assumption is arguable since in practice it is very difficult to infer this structure fully and accurately from measurements of sensor devices such as cameras or laser range finders. The approaches introduced in this paper are not explicitly handling contact-level grasp planning. Instead, we classify them as *pre-grip* components that are both dependent on selected extrinsic (*orientation*, *location*) and intrinsic (*size*, *shape*) properties. We see precise shape, weight or surface texture properties as being handled by an adjacent fine-controller based on tactile feedback and corrective movements, like included in Tegin et al.[50] The transport component (also dependent on *position* and *orientation*) is seen as a predecessor. It would demand grasp planning and collision detection in terms of successful robot hand transport, being a research topic for itself. However, the final location of a grasp is also clearly dependent on the task at hand, making the *task* another extrinsic property.

As previously mentioned, we make our grasp inference dependent on rough shape and size of the object as well as on its pose and the task at hand. There are two philosophies in this area that differ in the dimensionality of the visual data that is processed:

Inferring Grasps from 2D Images. Saxena et al.[51], Morales et al.[52] and Stark et al.[53] apply monocular images to derive suitable grasps. Such 2D approaches avoid the difficult problem of 3D reconstruction, as also their applicability is supported by a number of articles in the field of neurophysiology. As an example, Grezes and Decety[54], Tucker and Ellis[55] presented evidence for the theory that 2D visual perception of objects automatically activates relevant actions. Borghi[56], Creem and Proffitt[57] analysed what exactly influences the choice of the grasp in humans: knowledge of the object and its function or affordances as introduced by Gibson[58]. In both papers, it is claimed that in the case of novel objects, our actions are purely guided by Gibsonian affordances. In case of known objects, semantic information (e.g., through grasp experience) is needed to grasp them appropriately according to their function.

The work by Stark et al.[53] runs along the lines of the latter. Prehensile parts of objects are represented by $k-$Adjacent Segments (originally proposed for shape matching) that encode the relative geometric layout of distinct edge segments in an image. These so called *affordance cues* are obtained by observing the interaction of a person with a specific object. Grasp hypotheses for new stimuli are inferred by matching features of that object against a codebook of learned *affordance cues* that are stored along with relative object position and scale. However, how exactly to grasp these detected prehensile parts is not yet solved since hand orientation and finger configuration are not inferred from the affordance cues.

Saxena et al.[51] presented a system that infers a point at where to grasp an object directly as a function of its image. A learning algorithm is trained with labeled synthetic images of a number of different objects. The classification is based on a feature vector containing local appearance cues regarding color, texture and edges of an image patch in several scales and of its 24 neighboring patches in the lowest scale. The authors used their system specifically trained for a dishwasher scenario to pick up yet unseen objects from it and achieved impressive results.

In both these two papers[53,51] only local features are considered instead of the whole object. In contrast to that Goodale et al.[59], Cuijpers et al.[60], Gentilucci[61] emphasize the importance of global object shape for the purpose of reaching and grasping in humans. The work by Morales et al.[52] applies these findings on the relevance of global object shape to robotic grasping. Here, also the hand kinematics are considered to infer a number of planar grasp configurations directly from 2D object contours obtained through vision. To predict which of these grasps is the most stable one, a knn-approach is applied in connection with a grasp experience database. However, the approach is restricted to planar objects.

The related approach presented in this paper offers a base from which objects *similar* in shape can be grasped in a *similar* way. The approach is different from the one taken by Saxena et al.[51], Stark et al.[53] where only local appearance or affordance cues are used. In contrast to Morales et al.[52] where only planar objects are considered, we are considering arbitrarily shaped novel objects.

Reconstructing 3D Structure for the Purpose of Grasping. Though there is such interesting work on producing grasp hypotheses by visual features from 2D images, most

techniques rely on 3D data. 3D data, which in its simplest form may be a set of 3D points belonging to an object's surface, can be produced by several kinds of sensors and techniques, e.g. distance imaging cameras, laser scanners or stereo camera systems. These point clouds are usually afflicted with sensor noise and uncertainties.

A higher-level representation of these points as a set of shape primitives (e.g. planes, spheres or cylinders) obviously gives more valuable clues for object recognition and grasping by compressing the immanent shape information to its core. Miller et al.[62] therefore proposed grasp planning on simple shape primitives, like spheres, cylinders and cones, clearly demanding a pre-classification of object shape. Dependent on the primitive shape, one can test several grasp configurations on this shape. This work was continued by Goldfeder et al.[63] using more sophisticated shape primitives, known as superquadrics (SQs), which are parameterizable models offering a large variety of different shapes. Most approaches that consider this problem start from point-clouds and synthesizing object shapes by using superquadrics in a bottom-up manner. Considering the problem of 3D volume approximation, only superellipsoids are used out of the group of SQs, as only these represent closed shapes. There is a multitude of state-of-the-art approaches based on parameterized superellipsoids for modeling 3D range data with shape primitives.[64,65,63,66]

Assuming that an arbitrary point cloud has to be approximated, one SQ is not enough for most objects. The more complex the shape is, the more SQs have to be used to conveniently represent its different parts. However, good generality is not possible with few parameters for such cases.[64] Besides the advantages of immense parametrization capabilities with at least 11 parameters, intensive research on SQs has also yielded disadvantages in two common strategies for shape approximation. The first strategy is region-growing, starting with a set of hypotheses, the *seeds*, and let these adapt to the point set. However, this approach has not proved to be effective and suffers from the refinement problem of the seeds.[65,66] The second strategy uses a split-and-merge technique, which is more adapted to unorganized and irregular data.[65]

Independent of the strategy used, the models and seeds, respectively, have to be fitted to the 3D data. This is usually done by least square minimization of an inside-outside fitting function, as there is no analytical method to compute the distance between a point and a superquadric.[63] Thus, SQs are though a good trade-off between flexibility and computational simplicity, but sensitive to noise and outliers that will cause imperfect approximations. This is an important issue, as our work is oriented towards the use of dense stereo accompanied by highly distorted and incomplete data.

An issue that is immanent in shape approximation is that of object shape decomposition for grasping. The work by Lopez-Damian[67], Lopez-Damian et al.[68] proposes a grasp planner to find a stable grasp in addition to such a decomposition technique. However, their concept, as also the one by El-Khoury and Sahbani[69], uses polygonal structures instead of 3D points. Though one could produce polygonal surfaces from 3D point data, for example by the Power Crust algorithm,[70] this introduces another step causing additional effort both in processing time and noise handling. The already mentioned approach by Goldfeder et al.[63] combines superquadric representation and decomposition on regularly spaced range data.

In our work, we will work with simpler shape primitives for the purpose of grasping 3D shape. We chose the box shape as one of the most simple ones and integrate an efficient bounding box algorithm for pure 3D point data.[71] Section 3.4 will revisit this approach and its capabilities in terms of the framework presented in this paper.

3.2. *Attention System*

First, we will describe our visual attention system developed for the robotic system presented in this paper. The attention system uses top-down information in terms of a task dependent and volitional influence, and a second scene dependent and contextual influence. Given these sources of information, an Artificial Neural Network (ANN) learns the optimal bias of the top-down saliency map[35]. An unbiased version of the saliency map acts as a bottom-up map. These are then combined dynamically depending on past involvement and entropy measures. An inhibition-of-return (IOR) mechanism and a stochastic winner-take-all (WTA) network prevent the system from getting 'stuck' on previously attended regions.

Saliency Maps. We start by defining a top-down saliency map, SM_{TD} as a saliency map biased by some kind of learner trained to find objects of interest in arbitrary scenes. The bottom-up saliency map, SM_{BU}, is simply an unbiased version. The model, illustrated in Fig. 3, consists of SM_{BU} and SM_{TD} computed in parallel. The top-down bias is achieved by weight association by the ANN. The system combines $SM_{BU}(t)$ and $SM_{TD}(t)$ with a linear combination that evolves over time t.

Although some limitations of the Koch-Ullman saliency model[35] used here, have been demonstrated 1by Draper and Lionelle[72], we choose it for it computational speed and trade off time against precision. Similarly to Itti's original model[39], we use color, orientation and intensity features, and have complemented these with a texture cue[73,74]. The 'giraffe' example in Fig. 4 clearly exploits this need. None of the three original cues are able to make the giraffe 'pop-out' in the way the texture cue does.

Next, we want to be able to alter the top-down map by changing (optimizing) these weights for a certain task given a context (scene). In other words, we also have to examine what kind of context information would be important. This is simple because the optimal weight parameters for the same task typically differ from one context to the other. Thus, besides providing a saliency map, there are aditionally three importnat steps in the system:

- **Weight Optimization and Contextual Learning**: In particular, the SM_{TD} is obtained by weighting the different feature cues. For more on the details see[75].
- **Learning Context with a Neural Network** In order to include the correlation between the optimal weights for a given task and the context (scene information) as mentioned above, we have to define context. There are a large number of different definitions of context in the computer vision literature[76,77,78]. The definition that best serves our purposes of visual search is one based on global appearance statistics. Here, we rely on the ANN based learning as presented in our previous work[75].

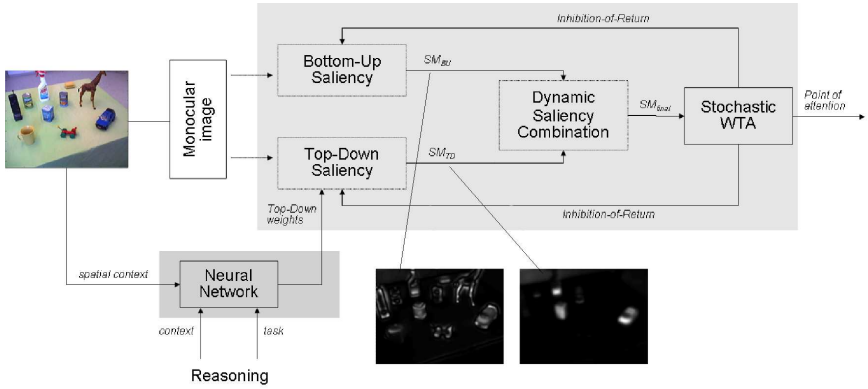


Fig. 3. Our attentional model combines bottom-up and top-down saliency in a dynamic manner. The top-down influence is tuned by an ANN.

(a) Original image (b) Texture cue (c) Color cue (d) Intensity+orientation

Fig. 4. The four feature cues for saliency maps.

- **Top-Down and Bottom-Up Integration** The mechanism for visual attention is obtained by combining SM_{BU} and SM_{TD} into a single saliency map that helps us to determine where to 'look' next. Using a ranking measure for each individual saliency maps, we can tell how much 'information' there is in attending that particular map. We use an energy measure (E-measure) similar to the Composite Saliency Indicator (CSI) of Hu et al.[79]. Accordingly, the top-down and bottom-up energies, E_{TD} and E_{BU}, are defined as the saliency density divided by the area of the convex hull of all salient points[80]. Thus, for a map with many salient points concentrated in a small area, the E-value is higher than for a map with the same number of salient points spread over a larger area. In other words, this measure favors saliency maps that contain a small number of very salient regions[75].

Implementation. The above-mentioned attentional system has been implemented on the four-camera Armar-III[a] stereo head shown in Fig. 5. The head consists of two foveal cam-

[a]More information about the whole Armar-III robot, a humanoid platform at the University of Karlsruhe, can be found in Asfour et al.[49] and on www.paco-plus.org.

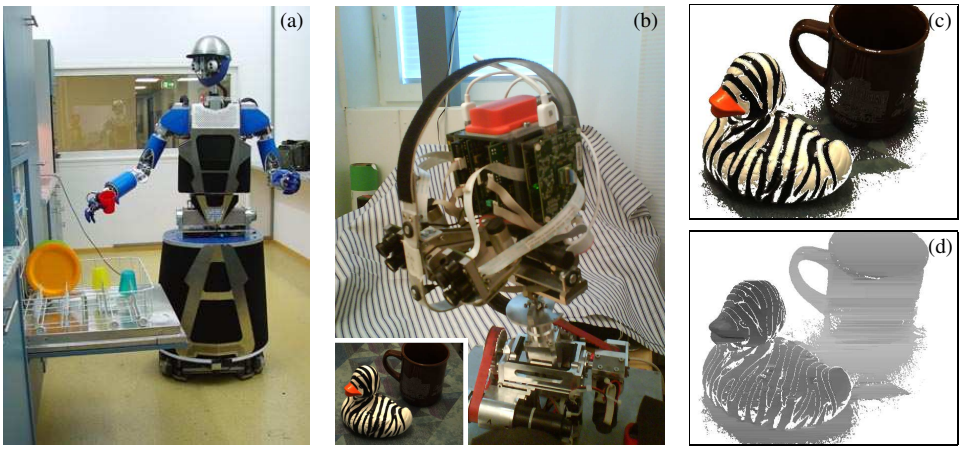


Fig. 5. (a) The Armar-III humanoid platform[49] used in the PACO-PLUS project. (b) A duplicate of the Armar-III stereo head, used in our lab, including a clipped region of an acquired rectified image. (c) Result of image differencing related to an image without the objects. This mask is applied in (d) to results from the disparity processor. Note that apart from white being the mask region, intensity corresponds to distance to the viewpoint.

eras for recognition and pose estimation, and two wide field cameras for attention. It has seven mechanical degrees of freedom: neck roll, pitch and yaw, head tilt and pan & tilt for each camera in relation to the neck. The attentional system keeps updating a list of scene regions that might be of interest to the rest of the system. The oculo-motor system selects regions of interest from the list and directs the head so that a selected region can be fixated upon in the foveal views. Redirection is done through rapid gaze shifts (saccades). As a consequence, the camera system always strives towards fixating on some region in the scene. The fact that the system is always in fixation is exploited for continuous camera calibration and figure-ground segmentation. There are two kinds of calibration needed for the perceptual system. One is a conventional *Eye-to-Hand Calibration*, i.e. the transformation between head and manipulator coordinate systems. The other is the *Eye-to-Eye Calibration*, or the calibration of extrinsic and intrinsic parameters of the binocular system. For more details regarding this see Björkman and Eklundh[26].

Object Segmentation in 2D. 2D object segmentation from a single image, as an optional part, will boost further performance of the perception system as local (2D) shape information aids the binocular fixation discussed earlier. Though there will 2.5D segmentation be discussed in Section 3.4, 2D segmentation in the image will already provide a focus on a *thing*. As an example, we assume the background image to be given in a static head scenario. The object is segmented by image differencing and the 3D point cloud from stereo can be masked easily to include only these points, as common uncertainties and noise in the environment can be removed. Additionally, an estimate of mean disparity can be computed from which the disparity algorithm benefits. More sophisticated methods for object segmentation in the 2D image have not been implemented in this system yet, but are clearly

available in the literature. Promising in this context are techniques like object segmentation from attention or object segmentation from manipulation. However, even a simple differencing subtraction method already demonstrates that a step of 2D segmentation is factually valuable for the whole system (see Fig. 5). At this stage, a *thing* can be seen as sets of pixels in image space which are assumed to belong to the same object.

3.3. *Grasping an Object based on 2D Shape*

As output from the attention module, we obtain a hypothesis about the existence of a physical object in form of segments in a stereo image pair. The following two sections will describe methods to infer grasp configurations for this potential object.

In this section, we predict prehensile parts of an object in a monocular image. Our approach is based on the hypothesis that two-dimensional visual attributes afford specific grasps. Here we consider *relative shape* which describes the global structure of an object relative to one of its parts. We represent this by applying *shape context* calculated based on the object's contour in a monocular image. We assume that we can apply grasping experience gathered from a set of known objects to grasp yet unknown objects that have similar shaped prehensile parts. To that end, we use a supervised learning technique on a database of synthetic images. The way of grasping novel objects by exploiting past experience with familiar objects has its correlate in the human nervous system. As discussed in Section 2, in these cases the ventral pathway responsible for object identification and the previously mentioned dorsal pathway work in a highly integrated manner. The result of the algorithm will be several point candidates in an image that are considered to be good places on an object at which the palm of a (robotic) hand can be applied to grasp it. In order to obtain a full grasp configuration consisting of 3D grasping point, hand approach vector and wrist orientation we use triangulation and a high-level approximation of the global object shape.

An overview of the whole subsystem described in this section is given in Fig. 6. It also shows interconnections to the reasoning system. For example, a task could be provided to the 2D grasp inference component determining which grasping point model should be applied. This model could be, e.g. dependent on one or several categories of objects or

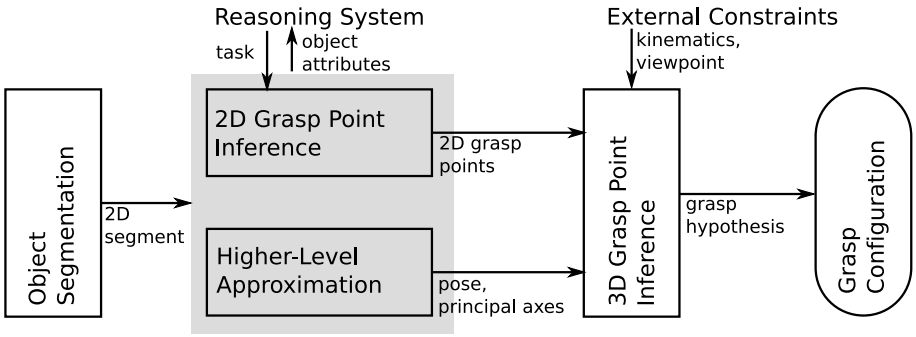


Fig. 6. Overview of the system to infer grasps from 2D shape information.

on the kind of higher level action (drinking from a cup requires another grasps than for example putting this cup into a dishwasher). This would require specific training data. However, in this work we apply a database that provides grasp experience dependent only on the object category. Different grasping point models are available either derived from a very specific subset of objects (e.g. only cups) or from all object categories, thus being very general. Information can also flow into the other direction, i.e. from the grasp inference system to the reasoning system. Object attributes are fed back to allow for potential failure analysis or re-planning.

In the following sections the two main subcomponents (inference of 2D grasping points and the global shape approximation) and their integration are explained. For a more detailed description, we refer to our previous work[81].

Inferring 2D Grasping Points. As mentioned above, the global object shape plays a significant role in the selection of an appropriate grasp. We need a local descriptor that relates this global property to each single point on the object. To encode this property of *relative shape* we apply the concept of shape context which we will briefly summarize in the following. For a more elaborate description, we refer to Belongie et al.[82].

The basis for the computation of shape context is an edge image of the object (Fig. 7b). N samples are taken with a uniform distribution from the contour (Fig. 7c). For each point we consider the vectors that lead to all the other sample points (Fig. 7d). These vectors relate the global shape of the object to the considered reference point. we comprise this information into a log-polar histogram to emphasize the influence of nearby samples (Fig. 7e). Shape context is invariant to translation, rotation and scale.

Grasping Point Descriptor. Given the segmented object in an image as input to our grasping point detection system, we compute the object contour by applying the Canny edge detector. This raw output is then filtered to remove spurious edge segments. A potential grasping point in an image is defined as a 10×10 pixels image patch. A descriptor for each patch serves as the basis to decide whether it is a grasping point or not. This descriptor is composed of the accumulated histograms of all sample points on the object's contour that lie in that patch. Typically only few sample points will be in a 10×10 pixel wide win-

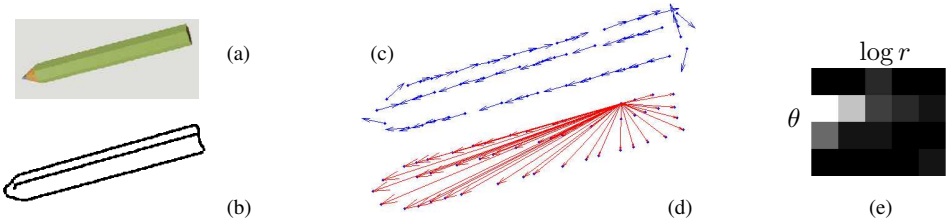


Fig. 7. Example for deriving the shape context descriptor for the image of a pencil. (a) Input image of the pencil. (b) Contour of the pencil derived with the Canny operator. (c) Sampled points of the contour with gradients. (d) All vectors from one point to all other sample points. (e) Histogram with four angle and five log-radius bins (θ and $\log r$) comprising the vectors depicted in (d).

dow. We therefore calculated the accumulated histograms in three different spacial scales centered at the current patch and concatenated them to form the final feature descriptor. In practice we applied 5 radius and 12 angle bins and sampled with 200 points as suggested by[82]. The overall dimension of the feature descriptor is thus 120.

Database. As a training set we applied the database by Saxena et al.[51] containing synthetic images of eight different object classes. Synthetic in this case means that a ray tracer was used to render images of different object models along with human-chosen grasp labels. Additionally, lighting conditions, object attributes (like color, texture and scale), camera positions and orientations are varied.

Classification. Let g_i denote the binary variable for the ith image patch in the input image. It can either carry the value 1 or 0 for either being a grasping point or not. The posterior probability for the former case will be denoted as $P(g_i = 1|D_i)$ where D_i is the feature descriptor of the ith image patch. To determine this posterior, we trained an SVM with a radial basis kernel.

Evaluation. We showed[81] that due to the application of global shape our grasping point inference system is robust against occlusion and strong texture. It generalizes well over novel object shapes.

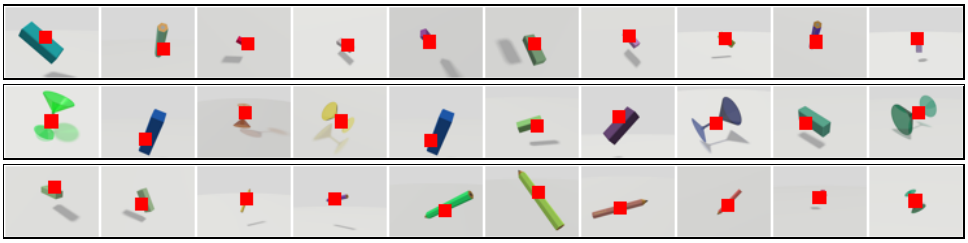
In Section 2 it is mentioned that graspable features are detected from visual data by the dorsal pathway. These features as encoded by our grasping point models trained on different sets of objects are depicted in Fig. 8. They are extracted by applying the Trepan Algorithm by Craven and Shavlik[83] to the learned classifiers. This algorithm builds a decision tree that approximates a concept represented by a given classifier. Although originally proposed for neural networks, Martens et al.[84] showed that it is also applicable for SVMs. One row in Fig. 8 shows samples from one leaf node of the induced decision tree that classifies the patches as grasping points (red squares in relation to the complete object shape). We see them as representatives of prototypical graspable features.

We can observe that when trained on different object classes, each prototype correspond mainly to one specific object, e.g. the set consisting of a pencil, a white board eraser and a martini glass has one leaf node for each of the object. One prototypical feature corresponds directly to one grasp type coupled to an object.

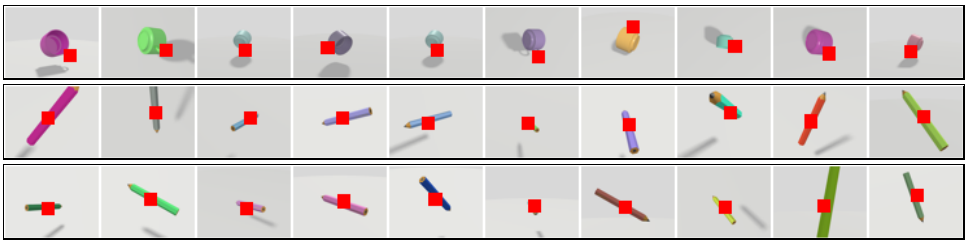
Approximating the Object Pose. As a manipulator we are considering a three-fingered Barrett hand[85] in a pinch grasp configuration (the two fingers are in parallel and opposing the thumb). Given a 2D grasping point detected with the method described above, we want to infer an appropriate 6 DoF grasp configuration i.e. the position and orientation of the Barrett hand. For that purpose, we need to roughly approximate the object pose. In the following we will briefly outline our approach that is described in more detail in our previous work[81].

According to Cuijpers et al.[60], humans grasp a cylindrical object highly dependent on the position of the major and minor axes of its cross section provided that a pinch

(a) Pencil



(b) Pencil, martini glass and white board eraser



(c) Pencil and mug

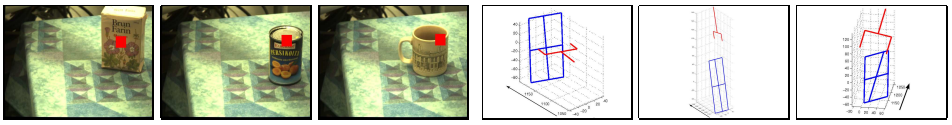
Fig. 8. Samples for prototypical grasp features given different training sets.

grasp (grasp with index finger and thumb) is applied. Here, we generalize this approach to arbitrarily shaped artifacts by fitting an ellipse to the segmented object in the image plane. We determine its orientation in 3D by applying stereo matching to a point on the major and minor axis and the centroid of the segment. The objects pose is then associated with the three dimensional position of its segment centroid and the orientation of the plane.

A byproduct of this process is the instantiation of extrinsic object attributes (*position* and *orientation*) as well as of intrinsic ones such as *shape* (e.g. approximated by the ellipse). *Color*, *texture* or *size* can be directly derived from the segment in the image. This information can be fed back to the reasoning system presented in Section 4 to allow for later analysis of potential failure or re-planning. The applied *actions* are another attribute associated to the object. How these are derived is presented in the following.

Generation of Grasp Hypotheses. After we run the classifier on each image of the stereo image pair, we have to associate the resulting 2D grasping hypotheses to each other in order to obtain a 3D point via triangulation. For this purpose, we are considering a set of local maxima regarding the grasping point detection in the left image and via stereo matching derive the corresponding points in the right image. The product of their grasping point probabilities forms the base for ranking 3D grasping points.

The next step is to infer the wrist orientation and approach vector of the Barrett hand for



(a) Objects with grasping points (b) Grasp Configurations

Fig. 9. Examples for generated grasp configurations. (a) Right image of the stereo camera with grasp point labeled. (b) Related grasp configuration with a schematic gripper and the plane with the axes approximating the object pose. the viewing direction is indicated by the arrow.

grasping the object at the best 3D grasping point. This is done by considering the relation of this point to the plane that approximates the object pose. The axis of the ellipse that is better aligned with vector from the 2D grasping point to the segment's centroid is chosen as the approach vector. The normal of the surface determines the wrist orientation. If the best grasping point is very close to the object's centroid, then the normal of the surface is the approach vector and the minor axis determines the wrist orientation. Examples for grasp configurations are given in Fig. 9. Previously, we showed how this approach is executed on our robotic platform.[81]

3.4. *Grasping an Object based on 3D Shape*

In this section, we will present an instance of a 3D perception system, using 3D dense stereo information, rough shape approximation, low-level grasp planning and grasp quality learning. Fig. 10 shows the instance of the system and its parts which will be described in the following paragraphs, or have been described previously in terms of stereo image acquisition and 2D Segmentation in Section 3.2. This process is accompanied by continuous gathering of attributes and emergence of symbols which transform *things* into *objects*. As discussed from the viewpoint of manipulation, robot grasping capabilities are necessary to actively execute tasks, interact with the environment and thereby reach versatile goals. These capabilities also include the generation of stable grasps to safely handle even objects unknown to a robot. In earlier work[86], it was motivated that the key to this ability is not primarily to select a grasp depending on the identification of a selected object, but rather on its shape. In this work, it was also claimed that 3D information is highly valuable for the purpose of grasping.

Object Segmentation in 2.5D. For the thereby motivated purpose of grasping on 3D shape, the 2D segmentation as discussed in Section 3.2 may be helpful, but not sufficient. We already presented our ideas for grasping from 2D image data. In general, however, and as long as there is no high-level reasoning system to infer 3D shape properties for unknown objects from a 2D image only, a mug on the cover of a magazine will not be distinguishable from a real cup on the table without any further analysis of 3D data. Additionally, estimation of an object's *size* or *shape* in three dimensions is intuitively valuable for its manipulation. 2.5D segmentation will help us to distinguish between objects in three dimensions with options beyond those of the presented 2D segmentation.

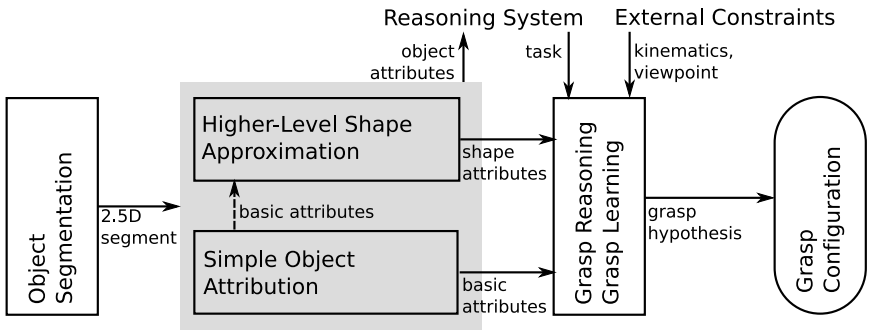


Fig. 10. Sketch of the open-loop 3D perception cycle providing shape representation from stereo image data.

General high-dimensional segmentation, be it in 3D space or even enriched with color space information, has high complexity and drawbacks. However, efficiently shortcutting this problem was successfully demonstrated through the assumption of planar surfaces [87,88]. In a number of manipulation scenarios, as also in ours, we can assume that manipulable objects are very commonly placed on a horizontal plane, e.g., a table. In our current system and scenario, where there is only one table for reasons of simplicity, detecting the table plane can either be done by Hough Transformation in 3D, or, and both much more efficiently and online, by integrating the vector of gravity. The vector of gravity corresponds to a good estimate of most table planes' normals, and can be deduced with minor effort from either the acceleration sensor or the kinematic chain of the head (see also Fig. 5b). Given a table plane, the 3D scene can further be purged by removing points lying on or below this plane. See Fig. 11 for an example.

Additionally, 2.5D segmentation and basic object attribution (e.g., *area*, *height*, etc.) become accessible by such plane definition [89]. Hence for the purpose of grasping, these attributes do already provide a valuable base for spatial object relationships and basic grasp planning, e.g. in terms of reachability (*position* on the table) and graspability (*size* related to the gripper) estimation. Of course, if objects are standing close to each other, 2.5D segmentation will detect them as one. We see this issue to be approached by the 2D segmentation discussed above or even explorative manipulation through expectation and surprise.

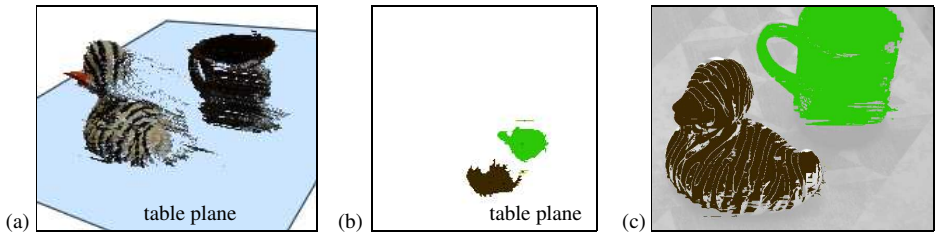


Fig. 11. (a) Scene reconstructed from Fig. 5d, purged by table plane assumption. (b) Objects deduced from 2.5D segmentation on table plane projection. Note that the table is not detected, but just the (infinite) table plane visualized. (c) Reprojection of the segmented objects to the image.

Higher-Level Shape Approximation. In general, the procedure followed in this section is the evaluation of *3D shape* for grasping. In case a point cloud belonging to an *object* (note that as mentioned, an object here might truly be a composition of closely neighboring objects on the table) has been extracted from the whole scene, one can estimate a *shape* representation of that object to generate grasp hypotheses on this representation. Which representation to choose is still an unanswered question and a large range is applied and explored in the literature, e.g. by point clouds [90], geometric shape primitives [91,92,62], or superquadric parametrizations [64,65,69,63]. In our earlier work [93,86], we have shown that even a *box* approximation of the point cloud can yield grasp hypotheses, as also interfaces to reason about task, viewpoint, graspability, and more.

Our representation efficiently approximates a 3D point cloud by a constellation of Minimum Volume Bounding Boxes (MVBBs). The fit-and-split approach starts with fitting an oriented root bounding box (see Fig. 12 Root) and estimates a best split by using the 2D projections of the enclosed points onto the box surfaces. Depending on a volume gain parameter t, two child boxes might be produced and then be tested for splitting. Due to this procedure, a binary-tree-shaped *box hierarchy* of object shape emerges, like the one shown in Fig. 12. For more details, we refer to our earlier work [86].

A main issue of ongoing work to make this approach more feasible in a real scenario is the hallucination of an object's backside which is commonly not visible from one viewpoint. Pointers from this problem are directed to superquadric estimation of the detected sub-parts of the point cloud, higher-order moment computation, symmetry generation or 3D surface extrusion from 2D projections [94], or even haptic exploration. In addition, subparts of interest could later on be described by the same methods to estimate a shape representative, e.g. *ellipsoid*, *cylinder* or *cone*.

Generation of Grasp Hypotheses. Through the approximation we have basically reached a shape representation stage. *Shape* is one of the intrinsic properties of an object which extend a basic object attribution. Besides the instantiation of extrinsic attributes (like *position* and *orientation*) and intrinsic attributes (like *color*, *texture* or *size*), *shape* is another intrinsic attribute. However, returning to mind that we want *action* to become an attribute by describing the close relationship between actions and objects or action and perception, shape definitely is one of the most important. In addition, any primitive shape representation enables low-level grasp planning and learning about shape closely related to actions in a number of different aspects:

Reducing Grasp Hypotheses from Geometrical Heuristics. In our case of a box-based representation, we can directly connect grasp hypotheses to the box faces in the decomposition, i.e. align grasp approach vectors with face normals and grasp orientations related to face edges. Experiments for this approach have been presented in [86]. A basic geometry-based heuristic framework to reduce the number of hypotheses has been demonstrated [93], where grasp hypotheses were selected related to a given task, or rejected because of geometrical occlusion, i.e. in the box constellation, or viewpoint. The heuristics were mainly

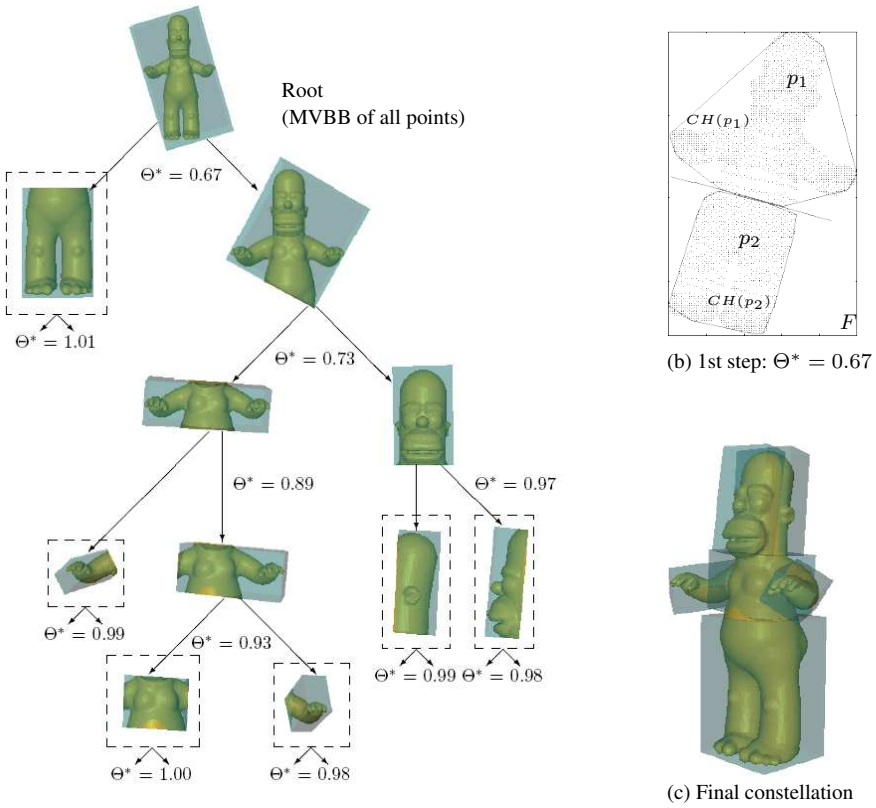


Fig. 12. Example of a decomposition hierarchy, using a gain parameter of t=0.98. With Θ^* below t, a valid cut is detected. Each split minimizes the volume based on the convex hull area projected onto the box faces, exemplified in (b). If t is exceeded, the box is a leaf box ('dashed), i.e. a part of the final constellation in (c).

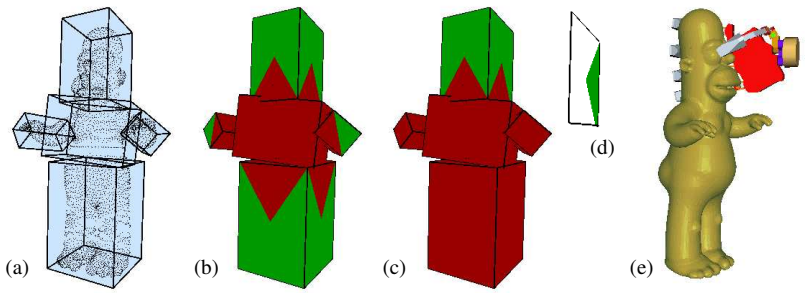


Fig. 13. (a) Point cloud and box constellation as in Fig. 12. (b) Resulting hypothesis restrictions from viewpoint, occlusion and blocking. Invalid hypotheses are depicted by dark (red) triangles, valid ones by light (green), where the inward vertex correlates to the hand orientation (the thumb). (c) According to a task like 'show', the head box was preferred. (d) The highlighted triangle corresponds to the most stable hypothesis of those in (c). (e) Visualization of the final grasp on the original model. Viewpoint has only been changed to show the finger contacts.

based on intrinsic geometric box attributes like size or orientation, but connected to extrinsic properties like task and viewpoint. To finally pick one grasp from the reduced set of hypotheses, a neural network approach was used to learn grasp qualities from box face representations. GraspIt![95], a grasp simulation environment, was used as a supervised network trainer for this purpose.

A sequence of such a heuristic decision process can be tracked in Fig. 13. First, the whole set of hypotheses from all faces is reduced to those that are graspable from a geometrical consideration, i.e. not occluded or blocked by other boxes. A grasp task like 'pick', 'show' or 'poke' is directly connected to a geometrical criterion (e.g.'outermost' box, i.e. the most distant one to the overall estimated center of mass, for 'show' in this example), with which the hypotheses are ranked, and a grasp pre-shape (e.g. pinch grasp for 'show' in this example). Finally, the network is trained with a set of examples to select the most stable hypothesis as the grasp to perform.

Introducing Grasp Kinematics. Since in these experiments, we only used basic grasp pre-shapes, i.e. power-grasp and pinch-grasp, we introduced hand kinematics in a later step. This was done by connecting 2D grasp mechanisms from Morales[96], Speth et al.[97] with the 3D box representation[98], in fact with the same face representations that we already briefly mentioned above. Each box allows to project the point set it envelopes onto each of its surfaces to produce a 2.5D projection, see Fig. 14. While we used a normalized pattern for learning grasp qualities using pre-shapes and fixed orientations in the heuristic approach above, we now use it for a contour-based grasp mechanism. This grasping mechanism rates triplets of contact points (in case of a 3-finger hand that we used in that work) using different stability criteria. Thus, this approach relies on knowledge about the gripper kinematics for the benefit of a better adapted stable grasp. Additionally, the grasp configuration holds a grasp configuration in terms of all degrees-of-freedom of the hand, in contrast to a combination of approach vector, orientation vector and grasp pre-shape.

Summary of Section 3. Concluding this section in terms of vision based grasp inference, we summarized our approaches on directly linking purpose-dependent manipulative actions with two different kinds of shape representations. The first approach considers 2D

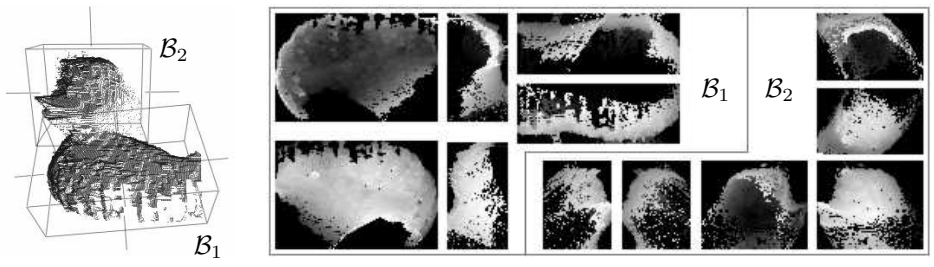


Fig. 14. Box decomposition of one of the objects presented in Fig. 11 and face projections of these two boxes. Intensity inversely corresponds to depth, thus projections can be interpreted as 2.5D depth maps.

shape and analyzes it based on previous experience that can be object or task dependent. The second approach derives a very fundamental primitive shape approximation from a 3D point cloud and directly links actions to parts of this approximation dependent on purpose, the object or kinematics. During this process, and more as a byproduct from the process of segmentation and simple geometrical shape representation, a number of various object attributes emerged. Such *attributes* and *symbols* are central for planning and reasoning on objects, actions, and affordances, and can be fed into that system (see Fig. 2). Especially, the 3D shape approximation provides a rich base for further analysis. Thus, concluding this section in the context of attributes, and planning and reasoning, we have currently a lot of options to proceed with in terms of objects, actions and combination of both:

(i) Planning and reasoning about basic attributes like *location, color* or *size* in terms of reachability and graspability (low-level[89] or high-level),
(ii) planning and reasoning about box *shape* constellations in terms of intuitive, task-dependent grasps (low-level[93] or high-level),
(iii) learning of grasp qualities on face projections with[98] or without[93] the explicit use of hand kinematics to find a stable grasp,
(iv) learning of or extension towards part-based shape primitives in future work,
(v) or learning of affordances, manifold actions and effects in future work.

All of those share aspects in which the proposed perceptual system can take advantage of an adjusted object shape approximation, which we believe is closely intertwined with object manipulation, at the low-level of grasp generation. An additional drive, however, is to port immanent high-level aspects to a high-level planning and reasoning system that can make use of such symbols. Such a system will be described in the next section.

4. Linking Low and High-Level Representations for Symbolic Reasoning

Reasoning is an important component of any robotic system for a number of reasons. Firstly it provides an agent with a way in which to explore the world and allows for learning of new skills such as grasping. In this case given the task of exploring, the reasoning system can drive the attention and perception systems to generate new grasp hypotheses not previously considered. On the other hand, when given a plan it can provide a means of understanding the world by examining causes for unexpected events, which we will refer to as failure. Finally, it can provide possible solutions for these failures in order to help the agent learn what went wrong during the execution cycle and how to proceed from its current state. In turn, the reasoning system aids in building a knowledge base of useful information for the planning system to use for building a set of possible planning routines. As such, the reasoning system is an important component not only for the attention and perception systems, but for planning and task execution as well. Although a discussion on planning is beyond the scope of this paper, we believe it is important to note that the use of vision coupled with the use of affordances for conducting high-level reasoning helps to facilitate planning. However, in this section we will only focus on the use of vision for learning affordances in order to conduct high-level reasoning.

In order to engage in high-level planning and reasoning tasks, sensory information collected through the visual stream must be re-formatted into a more useful representation. As presented earlier from a biological perspective, reasoning and planning are handled in the Prefrontal lobe and F6 regions and impact many other areas of the brain including the attention, perception, and execution regions. Therefore in order to accommodate data transfer between the perceptual areas of the brain and the reasoning system, a continuous mapping between low-level continuous data to high-level discrete symbols must take place. This mapping in turn enables the agent to plan actions and to reason about changes in the world. Assuming an existing planning system has been developed and is in use, we can focus our attention towards the role reasoning plays in the perception-action cycle by examining how knowledge about the world can be represented and used.

4.1. *Representing Sensor Data*

To facilitate continuous control of robotic systems we need representations that differ from the classical discrete symbolic AI representations commonly used for planning. Typical robotic data, which we denote here as low-level representations, can generally be characterized as vectors of continuous values, representing information such as relative and absolute positions, joint angels, force vectors and different image descriptors. Symbol-level representations, tend to be composed of discrete symbols related directly to objects and actions. These symbols are intended to capture low-level conceptual state changes as a result of some change in the world. However, neither of these types of representational systems alone covers all the requirements for execution of tasks in realistic settings. There is evidence to support the need for both types of low and high-level representations to produce human level behavioral control[99]. As such this section provides an overview of both low-level and high-level representations and discusses the use of high-level representations as a means for conducting symbolic reasoning.

Low-Level. Low-level representations for objects such as pixels and edges processed through the visual stream are typically represented using a common set of features such as color, texture, intensity, orientation, and position, as described earlier in Section 3. Using either monocular or stereo images, information about the world can be extracted and represented in either 2D or 3D. If the goal of the system is to only attend to points of interest in a scene, then 2D information provides a rich enough representation for where the agent should attend to. However, if an agent is required to act in the world, such as grasping an object, a more specific object representation may be required. In this case the use of 3D visual information provides an opportunity to obtain additional object details. Using information relating to the object such as depth, height, and width, *shape* emerges as part of a set of more advanced object descriptors[89]. Although establishing the shape of an object remains a challenge[81,86,30], this additional knowledge provides insight into the potential functionality of the object. As such, we must examine how to transition from low-level to high-level representations in order to use sensory information produced by the perception system more effectively.

Symbolic Level. At the symbolic level objects can take on a number of more meaningful associations including shape, size, and functionality. For instance what objects of a certain shape and size can be used for. In this case the object itself does not need to be identified, more simply only how the object can be used is of importance. For example, planar objects such as cuboids can be stacked while spherical objects cannot. In this case objects with a more cuboid appearance can be said to afford stacking, or have 'stack-ability'. Likewise, objects that appear hollow can be used as vessels for holding flowers, coffee, or used for cooking tasks (i.e. pots). Hollow or concave objects can therefore be said to afford filling, or have 'fill-ability'. This concept of affordances is not new and has been studied extensively since it was first coined by J.J. Gibson in the 1970's [100]. Using certain higher-level object features such as shape, size, or whether the object is textured or smooth, we can map specific objects to certain actions and model this at the symbolic level by examining how objects, actions, and the effects of those actions relate to each other. For example, spherical objects that are pushed have the effect of rolling, planar objects can slide, and roughly textured objects may be impervious to motion.

Context. At the symbolic level we must also take context into account. In order to understand the importance context plays in reasoning, we must first examine how context can be represented and how it can be used to assist inferencing. Context can include the task to perform, the embodiment of the agent (i.e. the physical constraints of the agent) or focus primarily on objects and places in the world (i.e. the physical constraints of the agent's environment). Focusing on the latter, objects processed through the visual stream are typically found within rich surroundings, often embedded in a context with other related objects. A context can therefore infer information that enhances interactions with objects. That is, how objects should be perceived and used.

It has been shown that humans relate objects to each other in various ways [101] (i.e. dimensions). These dimensions are encoded in different brain regions and one centralized component serves all of the relation classifiers on demand with a detailed object representation. The relationship types that are known to be encoded in humans are: physical appearance, basic level categories, contextual relations and semantic relations [102]. Torralba and Sinha [103] for example show that probable object locations and object scales can be inferred from a simple holistic representation of context based on the spatial layout of spectral components. The technique can be used to infer the area of interest when performing object detection. From our perspective this shows that context can be established without a prior object classification.

Context also facilitates the recognition of related objects even if these objects are ambiguous when seen in isolation. An ambiguous object becomes recognizable if another object that shares the same context is placed in an appropriate spatial relation to it [102]. Studies with sequential observations of objects without background shows that objects infer context. The first object establishes the context and subsequent objects are recognized momentarily if they are associated with the same context. If however the object presented does not belong to that context then the brain has to firstly recognize the object and sec-

ondly establish the new context. Reviewing previous work in this area gives a foundation for modeling context as a set of related objects with relations between them. For contextual information to assist the recognition process, it has to be extracted rapidly in order to generate guiding expectations. The work by Torralba and Sinha [103] provides evidence that such a mechanism for rapid context recognition is possible to construct. From here, objects can then be mapped to specific contexts in order to constrain the types of actions that can be applied to them (i.e. grasp types).

Understanding Failure. Along with context, failure during task execution is another area we are interested in examining. In the perception-action cycle there are a number of failure situations that may occur. Firstly, failures in the attention system can arise if specific objects being searched for cannot be found in the current scene. Failures of this type could be potentially resolved using two approaches: (i) the attention system can attend to the scene from a different viewpoint in order to re-assess the image (i.e. moving the camera position), or (ii) the reasoning system issues an exploration task to the robot. In the latter case the robot may engage in a series of pushing actions involving the perception system to try and separate objects located on a table in order to reduce occlusions. By reducing occlusions, the attention system increases the likelihood of matching the desired search criteria. Secondly, action failures can occur if the control system cannot execute the action issued by the planning system as expected. For instance, if the planner issues a command for the robot to close its grippers around an object and lift that object up off of a table, there may be several failures encountered. One type of failure may include closing the robot's fingers and making an unwanted only partial contact with the object. This can result in grasp failure being detected through both the visual and haptic streams. The reasoning system therefore must reason as to the most probable cause for the failure. In this case an incorrect grasp type selected by the perception system may be the most likely cause. From this point once the cause for the failure has been established by the reasoning system, this new knowledge can then be fed back to the perception system and re-planning can then take place.

This process of identifying symbols for objects, actions, effects, context and potential causes of failure is an important step in order to provide a useful framework within which to perform reasoning. Once an agent has a way in which to reason effectively about events that take place, it can then perceive and act more appropriately in the world.

4.2. *Probabilistic Approach to Modeling Affordance Relations for Reasoning*

The use of symbols provides a higher-level discrete representation for objects and their potential usage. From here we can model what certain objects can and cannot afford and reason when unexpected events take place in the world. As a first step towards building such a reasoning system we introduce the work by Montesano et al. [104], as an initial starting point with which to build on for our future work. We present Montesano et al.'s [104] probabilistic model in this paper in order to illustrate how symbolic reasoning can be accommodated, and how this approach can be used to design a reasoning system that uses the concept of affordances to perform inferencing at the symbolic level more effectively.

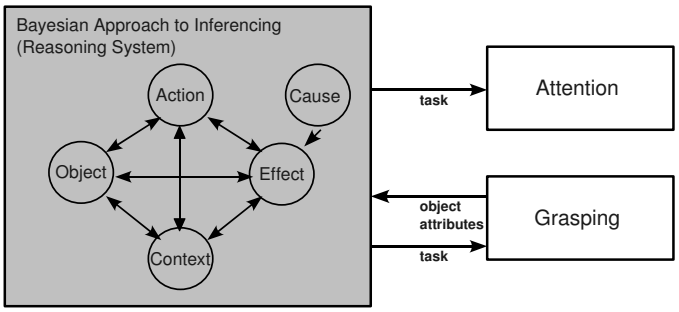


Fig. 15. A conceptual design of the reasoning system, and its relationship to the vision system.

In the model presented by Montesano et al.[104] affordances are learned and modeled as relations between objects, actions, and effects using a Bayesian network (BN) similar to the conceptual ideas of a reasoning system presented in Fig. 15 (i.e. their work excludes the context and cause nodes shown here).

In their representation objects are defined according to their shape and size and associated with specific actions, and the resulting effects of those actions. The actions they define include grasping, tapping, and touching objects. Objects are represented using either a ball or box shape and classified as either small, medium, or large in size. Colors such as green, yellow, or blue are also included as part of the object descriptor. The resulting effects of actions applied to objects included contact duration between the object and the end-effector during a grasping operation, object velocity (post-contact), and hand velocity (pre-contact). Each type of action (i.e. grasp, tap, touch), object descriptor (i.e. shape, size, color), and effect is represented as a single node in their Bayesian network. The relations (i.e. edges) between each of these nodes is considered an affordance and is learned using an imitation learning approach.

Although Montesano et al.[104] do not take into account context and potential causes for failures, we believe that these aspects are an important addition to the model as shown previously in Fig. 15. In our representation, context is defined according to the embodiment of the agent (i.e. physical limitations of the robot), a specific task to be performed (i.e. task execution), and where those tasks should be conducted (i.e. environment restrictions). Using this information, we can then narrow down the set of possible grasping choices from which to choose from. In other words, the inclusion of context provides an additional constraint to the search space and can therefore reduce the workload placed on the reasoning system. This constraint also limits what the attention system should and should not attend to, subsequently reducing the amount of computational processing the perception system must perform. Likewise, examining the potential cause for failure is also an important component of the system in order to understand what went wrong and possibly to correct unwanted outcomes.

If we take into account context and potential causes of failure then Montesano et al.'s[104] probabilistic model of affordances can be expanded to include these additional variables.

In this case the main variables of the reasoning system would then include *object*, *action*, *effect*, *context*, and *cause*, as shown in Fig. 15, with each of these variables being assigned a number of discrete states.

Each discrete state can then be represented as a single node in the Bayesian network. For example, *Object* would be represented by six nodes (i.e. shape, size, weight, texture, type, pose). *Action* represented by one node (i.e. action). *Context* represented by three nodes (environment, task, and embodiment), *Cause* represented with one node (i.e. action failure cause), and so on. Determining each of the discrete states for actions, objects, effects, context and cause is done by extracting perceptual data and mapping this data to higher-level symbolic descriptors. We assume that learning of sensory-motor coordination has already taken place as described by Lopes et al. in their developmental roadmap approach [105].

Once the low-level sensory-motor mapping and symbolic components for the reasoning system are in place, reasoning can use the perception, attention, and control systems to query for support and guiding information. For example, querying the reasoning system for an action, given an object and a certain task, using symbolic representations would have the form: $P(Action \mid ObjectShape, Task)$. As such, the result from the query is a set of actions and an associated probability distribution, with the most probable action to execute being the final output of the system. Using the same basic querying mechanism we can answer a number of different questions such as: (i) What environment is the robot acting in? (ii) What is the task being performed? (iii) Observing a certain effect (i.e. rolling), what type of object is the robot interacting with (i.e. ball)? and so on. This approach therefore provides both a framework for reasoning about how certain actions affect certain objects and a way in which to diagnose possible causes for failure encountered during task execution.

Example Reasoning Scenario. In order to illustrate the role of the reasoning system more clearly we can explore the task of picking up a bottle within the context of a kitchen setting. In this scenario the initial goal for the attention and perception systems are to attend to points of interest in the environment that may have objects. From here, objects that appear cylindrical in shape can be extracted/segmented from the scene and mapped to an appropriate grasp type in order for a grasp attempt to be made. If however the grasp fails (i.e. haptic feedback detects no contact between the object and the end-effector), as illustrated in Fig. 16, then the reasoning system must use feedback from the visual stream to: (i) confirm the failure detected (i.e. confirm object is not being grasped by end-effector), and (ii) to understand the cause of the failure. In this case the list of possible causes may include object slippage, or incorrect grasp type selected. If the reasoning system returns the most likely cause for failure was due to slippage, then this may be a result of underestimating the objects weight (i.e. object heavier than expected), resulting in insufficient grip-force being applied during the initial failed grasp attempt. In other cases the wrong grasp type may have been selected as a result of an incorrect segmentation of the scene. Objects that are classified as cylindrical for example may be more complex in shape and require more advanced grasp configurations. In this event the reasoning system can prompt the vision system to extract additional details from the scene in order to learn more about

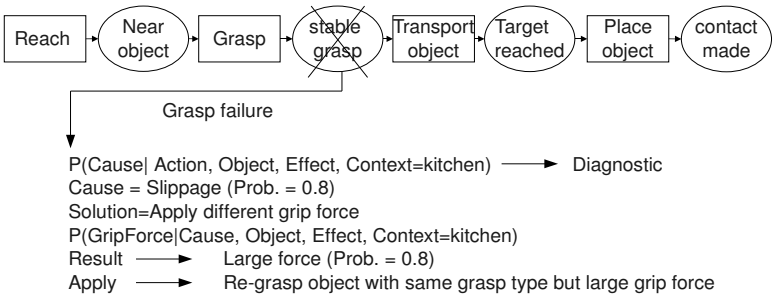


Fig. 16. Probabilistic inferencing approach for understanding and reasoning about failure.

the object's structure. This may result in an exploratory stage where the vision system and the end-effector work together in a perception-action cycle to examine the object more closely. For example, using a pinch grasp to hold a small sub-section of the object in order to prevent unnecessary occlusions, the vision system can gather more detailed information about the object that it did not previously have. This new information can then be used by the perception system to select alternative future grasp types.

5. Example Scenarios

After we have presented and described the components to use in our framework, we will now motivate them in a two-level scenario. The common task will be to 'assemble a tea set,' while both sub-scenarios differ in complexity of the scene and use the connected components, respectively. We assume the following general assumptions for the scenario:

(i) The plan to 'assemble a tea set' is known and described by partial actions 'put the saucer on the tray', 'put the cup on top of the saucer', and 'put the spoon into the cup'.
(ii) The robot is within the working space, i.e. having a natural close viewpoint onto the table where all objects are placed.
(iii) The goal objects that are needed for the task (tray, saucer, cup, spoon) are on the table.
(iv) The objects' appearance models are known to the system, i.e. they are supposed to be identified with high probability.

Given these assumptions, the difference between our two scenarios will be the degree of occlusion of the known objects by completely unknown objects. In the first scenario, the known objects are the only objects on the table, and thereby clearly visible and unoccluded. Therefore, this scenario will mainly show the *task-based capabilities* of our system in terms of planning and reasoning. In the second scenario, when objects are occluded, not only a visual search is needed, but also an interactive search for the known objects by putting the unknown aside. Within this scenario, the *grasp-based capabilities* of our system get into focus, but are still connected to the plan-reason-system on grasp level.

We will demonstrate the approaches for these two problems along the presented components, which are

(i) our plan-reasoning module $\mathcal{PR}$, presented in Section 4,
(ii) our attention-segmentation module $\mathcal{AS}$, presented in Section 3.2,
(iii) both of our perceptual grasp hypotheses generation modules $\mathcal{G}^{2D}$ and $\mathcal{G}^{3D}$, presented in Sections 3.3 and 3.4, respectively.

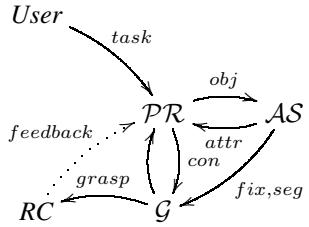
5.1. *The Task-Oriented Scenario*

Initializing the system with a the task 'assemble a tea set,' $\mathcal{PR}$ will connect a plan to this task, taking into account constraints, e.g. about the embodiment and experience. These might also include knowledge about the perceptual modules of the system, i.e. their assumptions and complexity. Since the goal objects are known both from the plan and from experience, $\mathcal{PR}$ will trigger $\mathcal{AS}$ to detect them sequentially in an order preferable for the task. Since we assume an unoccluded, uncluttered scenario with appearance-wise well trained goal objects, $\mathcal{AS}$ will provide visual attributes of these objects (e.g. texture, color), as also their locations (in the image) and information about their fixation and segmentation:

By feeding back information about the visual attributes of each object to $\mathcal{PR}$, a decision can be made about which perceptual grasp system shall be triggered. Embodiment, experience and visual attributes are main keys to make this decision. In our framework, criteria could look like the following:

(i) Embodiment: Monocular Vision $\rightarrow \mathcal{G}^{2D}$; Stereo Vision $\rightarrow \mathcal{G}^{3D}$
(ii) Experience: On 2D Shape $\rightarrow \mathcal{G}^{2D}$; On 3D Shape $\rightarrow \mathcal{G}^{3D}$
(iii) Attributes: Non-Textured $\rightarrow \mathcal{G}^{2D}$; Textured $\rightarrow \mathcal{G}^{3D}$

Imagine the cup is textured, it would be valuable to use the texture-driven dense stereo system $\mathcal{G}^{3D}$. In contrast to that, for a uniformly colored spoon, the use of the contour-driven shape context system $\mathcal{G}^{2D}$ will be motivated. While $\mathcal{PR}$ will decide on which module to choose in this context, it also has to provide part of the constraints about the task to it. As presented, $\mathcal{G}^{2D}$ and $\mathcal{G}^{3D}$ keep interfaces on the task level, e.g. which grasp or hand configuration to prefer. The selected $\mathcal{G}$ component's input will be enriched by the fixation and segmentation from $\mathcal{AS}$. The produced grasp hypothesis will be provided to $\mathcal{PR}$ for reasoning purpose, and to the robot controller (*RC*) to be performed:



To be able to cope the case when an action fails, concepts for failure and surprise have to be incorporated into this open-loop perception-action cycle. Assuming a grasp failure occurs, the *RC* detects the failure through haptic feedback and reports the event to $\mathcal{PR}$. From here reasoning is performed to conjecture about the cause of failure. Failure is defined as a mismatch between an expected and an actual event taking place and can be further

confirmed through visual feedback. Once the cause for the failure has been determined, the reasoning system works together with the planner to re-plan a new sequence of actions. This action sequence also drives how the attention and perception systems should proceed from the current failed state.

5.2. *The Grasp-Oriented Scenario*

The previous scenario is strongly supported by the object detection capability of the $\mathcal{AS}$ component. Due to this capability, it would theoretically be possible for a model-based perceptual grasp generator $\mathcal{G}^{Mod}$ to identify grasp hypotheses directly. However, we have to assume that the goal objects may be occluded by unknown objects in a real scenario. In such a case, the goal objects are not directly perceivable. When occluded by unknown object, those have to be interactively put aside. Anytime several objects are segmented as one, manipulative actions might be needed to distinguish between one real object or a composition of many.

However, the outline emerging from the scenario above does not have to be changed for this purpose. If $\mathcal{AS}$ reports to $\mathcal{PR}$ with having not found any of the goal objects while the task is not finished yet, $\mathcal{AS}$ can be triggered bottom-up without having task information and context included. It will provide $\mathcal{G}$ with information about the unknown object, triggering an action to put it aside. $\mathcal{G}^{2D}$ and $\mathcal{G}^{3D}$ are not dependent on knowledge about the object to produce grasp hypotheses, but can work on any segment that $\mathcal{AS}$ will provide them with. Since $\mathcal{PR}$ has been briefed by $\mathcal{AS}$ beforehand, it will interpret the grasp hypothesis from $\mathcal{G}$ as a 'put-aside' of the unknown object. By following this procedure, we will fall back to the task-based scenario where known objects are unoccluded and unclustered.

6. Conclusions and Future Work

The field of robotics is continuously expanding. The question is no longer if robots will take the leap out of the factories and into our homes but when and to what extent. Thus, the future applications of autonomous agents require not only the ability to move about in the environment, but also the ability to interact with objects. 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. For such applications, it is clear that vision is the most important sensing modality. As such, robot vision extends methods of computer vision to fulfill the tasks given to robots and robotic systems.

The work presented in this paper dealt with the aspects of vision based processing for object grasping and manipulation applications. In particular, we presented how to use attention and perception more effectively in order to reason about optimal grasp choices. The system design has been presented from both a neuroscience and robotic systems perspective. We have discussed different strategies for vision based attention considering both top-down and bottom-up strategies. To allow grasping of known and unknown objects, we considered several approaches based both on 2D and 3D visual information. The output of

the visual system has also been studied in the context of data representation and symbolic reasoning for the purpose of task execution. As an additional result, a number of various object attributes emerged. Such attributes and symbols are central for planning and reasoning on objects, actions, and affordances, and can be fed into that system. In particular, the work on 3D shape approximation and grasping, represents a direct link between the purpose-dependent manipulative actions and a very fundamental primitive shape approximation.

Our current work deals with the further development of both individual modules and systems integration. The system presented here opens also a whole set of questions for the future. In the development of robot systems that can act and interact in natural environments, we need to put a lot of effort on the representations that allow the robot to cope with the knowledge uncertainty and incompleteness, using learning and reasoning. During the recent years, it has been realized that these representations need to be grounded in the sub-systems such as object grasping and manipulation, vision based processing, or spatial modeling. In other words, there is still a significant need for the development of the sub-systems and understanding of how modeling of sensors and different hardware components of the robot affects the higher level reasoning and representations. We believe that our current and future work provide contributions in this direction.

Acknowledgements

This work was supported by EU through the projects PACO-PLUS, IST-FP6-IP-027657 and GRASP, IST-FP7-IP-215821. The authors would like to thank Mårten Björkman for valuable comments to improve this article.

References

1. J. Gray, C. Breazeal, M. Berlin, A. Brooks, and J. Lieberman. Action Parsing and Goal Inference using Self as Simulator. In *IEEE Workshop on Robot and Human Interactive Communication*, pages 202–209, 2005.
2. T. Kyriacou, G. Bugmann, and S. Lauria. Vision-Based Urban Navigation Procedures for Verbally Instructed Robots. *Robotics and Autonomous Systems*, 51(1):69–80, 2005.
3. P. McGuire, J. Fritsch, J.J. Steil, F. Rothling, G. A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter. Multi-Modal Human-Machine Communication for Instructing Robot Grasping Tasks. In *IEEE/RSJ Int. Conference on Intelligent Robots and Systems*, pages 1082–1089, 2002.
4. 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*, 111(2):246–252, 1996.
5. G. Rizzolatti, G. Luppino, and M. Matelli. The Organization of the Cortical Motor System: New Concepts. *Electroencephalography and Clinical Neurophysiology*, 106(4):283–296, 1998.
6. A. H. Fagg and M. A. Arbib. Modeling Parietal-premotor Interactions in Primate Control of Grasping. *Neural Networks*, 11(7-8):1277–1303, 1998.
7. 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*, 95(2):709–729, 2006.
8. Z. Li. A Saliency Map in Primary Visual Cortex. *Trends in Cognitive Sciences*, 6(1):9–16, 2002.

9. M. Goodale. Separate Visual Pathways for Perception and Action. *Trends in Neurosciences*, 15(1):20–25, 1992.
10. M. Lu, J. Preston, and P. Strick. Interconnections Between the Prefrontal Cortex and the Premotor Areas in the Frontal Lobe. *Journal of Comparative Neurology*, 341(3):375–392, 1994.
11. G. Luppino, M. Matelli, and G. Rizzolatti. Cortico-cortical Connections of Two Electrophysiologically Identified Arm Representations in the Mesial Agranular Frontal Cortex. *Experimental Brain Research*, 82(1):214–218, 1990.
12. D. Ingle, G. Schneider, C. Trevarthen, and R. Held. Locating and Identifying: Two Modes of Visual Processing. *Psychological Research*, 31(1):42–43, 1967.
13. E. Chinellato and A. del Pobil. *Neural Coding in the Dorsal Visual Stream*, pages 230–239. Springer-Verlag Berlin / Heidelberg, 2008.
14. A. Murata, V. Gallese, G. Luppino, M. Kaseda, and H. Sakata. Selectivity for the Shape, Size, and Orientation of Objects for Grasping in Neurons of Monkey Parietal Area AIP. *Journal of Neurophysiology*, 83(5):2580–2601, 2000.
15. T. James, G. Humphrey, J. Gati, R. Menon, and M. Goodale. Differential Effects of Viewpoint on Object-driven Activation in Dorsal and Ventral Streams. *Neuron*, 35(4):793–801, 2002.
16. E. Chinellato, Y. Demiris, and A. P. del Pobil. Studying the Human Visual Cortex for Achieving Action-perception Coordination with Robots. In *IASTED Int. Conference on Artificial Intelligence and Soft Computing*, pages 184–189, 2006.
17. U. Castiello and M. Jeannerod. Measuring Time to Awareness. *Neuroreport*, 2(12):797–800, 1991.
18. M. J. Webster, J. Bachevalier, and L. G. Ungerleider. Connections of Inferior Temporal Areas TEO and TE with Parietal and Frontal Cortex in Macaque Monkeys. *Cerebral cortex*, 4(5): 470–483, 1994.
19. G. Rizzolatti and G. Luppino. The Cortical Motor System. *Neuron*, 31(6):889–901, 2001.
20. M. Petrides and D. N. Pandya. Projections to the Frontal Cortex from the Posterior Parietal Region in the Rhesus Monkey. *Journal of Comparative Neurology*, 228(1):105–116, 1984.
21. M. A. Arbib, A. R. Iberall, and D. Lyons. Coordinated Control Programs for Control of the Hands. In A. W. Goodwin and I. Darian-Smith, editors, *Hand function and the neocortex*, experimental brain research supplemental 10, pages 111–129. Springer-Verlag, Berlin, 1985.
22. T. M. Preuss and P. S. Goldman-Rakic. Connections of the Ventral Granular Frontal Cortex of Macaques with Perisylvian Premotor and Somatosensory Areas: Anatomical Evidence for Somatic Representation in Primate Frontal Association Cortex. *Journal of Comparative Neurology*, 282(2):293–316, 1989.
23. G. Recatalá, E. Chinellato, Á. del Pobil, Y. Mezouar, and P. Martinet. Biologically-Inspired 3D Grasp Synthesis Based on Visual Exploration. *Autonomous Robots*, 25(1):59–70, 2008.
24. R. S. Johansson and G. Westling. Roles of Glabrous Skin Receptors and Sensorimotor Memory in Automatic Control of Precision Grip When Lifting Rougher or More Slippery Objects. *Experimental Brain Research*, 56(3):550–564, 1984.
25. A. Ude, C. Gaskett, and G. Cheng. Foveated Vision Systems with Two Cameras per Eye. In *IEEE International Conference on Robotics and Automation*, pages 3457–3462, 2006.
26. M. Björkman and J-O. Eklundh. Attending, Foveating and Recognizing Objects in Real World Scenes. In *British Machine Vision Conference, BMVC'04*, 2004.
27. J.K. Tsotsos. Analyzing Vision at the Complexity Level: Constraints on an Architecture, An Explanation for Visual Search Performance, and Computational Justification for Attentive Processes. Technical report, University of Toronto Vision Laboratory, 1987.
28. J.K. Tsotsos. A 'Complexity Level' Analysis of Immediate Vision. *International Journal of Computer Vision*, 1(4):303–320, 1988.
29. J.K. Tsotsos. The Complexity of Perceptual Search Tasks. Technical report, University of Toronto Vision Laboratory, 1989.

30. S. Ekvall and D. Kragic. Learning and Evaluation of the Approach Vector for Automatic Grasp Generation and Planning. In *IEEE International Conference on Robotics and Automation*, pages 4715–4720, 2007.
31. A. Morales, P. Azad, T. Asfour, D. Kraft, S. Knoop, R. Dillmann, A. Kargov, C. Pylatiuk, and S. Schulz. An Anthropomorphic Grasping Approach for an Assistant Humanoid Robot. In *37th International Symposium on Robotics*, pages 149–152, 2006.
32. L. Itti. *Models of Bottom-Up and Top-Down Visual Attention*. PhD thesis, California Institute of Technology, 2000.
33. 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*, 13:4700–4719, 1993.
34. A.M. Treisman and G. Gelade. A Feature Integration Theory of Attention. *Cognitive Psychology*, 12:97–136, 1980.
35. C. Koch and S. Ullman. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, 4:219–227, 1985.
36. T. Koike and J. Saiki. Stochastic Guided Search Model for Search Asymmetries in Visual Search Tasks. *Biologically Motivated Computer Vision*, pages 408–417, 2002.
37. O. Ramström and H.I. Christensen. Object Detection using Background Context. In *International Conference of Pattern Recognition*, pages 45–48, 2004.
38. S.B. Choi, S.W. Ban, and M. Lee. Biologically Motivated Visual Attention System using Bottom-Up Saliency Map and Top-Down Inhibition. *Neural Information Processing - Letters and Review*, 2:19–25, 2004.
39. L. Itti, C. Koch, and E. Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
40. V. Navalpakkam and L. Itti. Sharing Resources: Buy Attention, Get Recognition. In *International Workshop Attention and Performance in Computer Vision*, 2003.
41. K. Lee, H. Buxton, and J. Feng. Selective Attention for Cueguided Search using a Spiking Neural Network. In *International Workshop on Attention and Performance in Computer Vision*, pages 55–62, 2003.
42. A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson. Top-Down Control of Visual Attention in Object Detection. In *International Conference on Image Processing*, pages 253–256, 2003.
43. S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, volume 3899. Springer, 2006. ISBN 978-3-540-32759-2.
44. J. Moren, A. Ude, A. Koene, and G. Cheng. Biologically-Based Top-Down Attention Modulation for Humanoid Interactions. *Int. Journal of Humanoid Robotics*, pages 3–24, 2008.
45. A. M. Okamura, N. Smaby, and M. R. Cutkosky. An Overview of Dexterous Manipulation. In *IEEE International Conference on Robotics and Automation*, pages 255–262, 2000.
46. K.B. Shimoga. Robot Grasp Synthesis Algorithms: A Survey. *International Journal of Robotic Research*, 15(3):230–266, 1996.
47. M. Ciorcarlie, C. Goldfeder, and P. Allen. Dexterous Grasping via Eigengrasps: A Low-Dimensional Approach to a High-Complexity Problem. *Robotics: Science and Systems Manipulation Workshop*, 2007.
48. C. Borst, M. Fischer, and G. Hirzinger. Grasping the Dice by Dicing the Grasp. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3692–3697, 2003.
49. T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann. ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control. In *6th IEEE-RAS International Conference on Humanoid Robots*, pages 169–175, 2006.
50. J. Tegin, S. Ekvall, D. Kragic, B. Iliev, and J. Wikander. Demonstration based Learning and Control for Automatic Grasping. *Journal of Intelligent Service Robotics*, 2(1):23–30, January

2009.
51. A. Saxena, J. Driemeyer, J. Kearns, and A. Y. Ng. Robotic Grasping of Novel Objects. *Neural Information Processing Systems*, 19:1209–1216, 2007.
52. A. Morales, E. Chinellato, A. Fagg, and A. del Pobil. Using Experience for Assessing Grasp Reliability. *International Journal of Humanoid Robotics*, 1(4):671–691, 2004.
53. M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele. Functional Object Class Detection Based on Learned Affordance Cues. In *6th International Conference on Computer Vision Systems*, volume 5008 of *LNAI*, pages 435–444. Springer-Verlag, 2008.
54. J. Grezes and J. Decety. Does Visual Perception of Object Afford Action? Evidence from a Neuroimaging Study. *Neuropsychologia*, 40(2):212–222, 2002.
55. M. Tucker and R. Ellis. On the Relations Between Seen Objects and Components of Potential Actions. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3): 830–846, 1998.
56. A.M. Borghi. *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*, chapter Object Concepts and Action. Cambridge University Press, 2005.
57. S. H. Creem and D. R. Proffitt. Grasping Objects by Their Handles: A Necessary Interaction between Cognition and Action. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1):218–228, 2001.
58. J.J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1979.
59. M. A. Goodale, J. P. Meenan, H. H. Bülthoff, D. A. Nicolle, K. J. Murphy, and C. I. Racicot. Separate Neural Pathways for the Visual Analysis of Object Shape in Perception and Prehension. *Current Biology*, 4(7):604–610, 1994.
60. R. H. Cuijpers, J. B. J. Smeets, and E. Brenner. On the Relation Between Object Shape and Grasping Kinematics. *Journal of Neurophysiology*, 91:2598–2606, 2004.
61. M. Gentilucci. Object Motor Representation and Reaching-Grasping Control. *Neuropsychologia*, 40(8):1139–1153, 2002.
62. A. T. Miller, S. Knoop, H. I. Christensen, and P. K. Allen. Automatic Grasp Planning Using Shape Primitives. In *IEEE Int. Conf. on Robotics and Automation*, pages 1824–1829, 2003.
63. C. Goldfeder, P. K. Allen, C. Lackner, and R. Pelossof. Grasp Planning Via Decomposition Trees. In *IEEE International Conference on Robotics and Automation*, pages 4679–4684, 2007.
64. G. Biegelbauer and M. Vincze. Efficient 3D Object Detection by Fitting Superquadrics to Range Image Data for Robot's Object Manipulation. *IEEE Int. Conf. on Robotics and Automation*, pages 1086–1091, 2007.
65. L. Chevalier, F. Jaillet, and A. Baskurt. Segmentation and Superquadric Modeling of 3D Objects. *Journal of Winter School of Computer Graphics, WSCG'03*, 2003.
66. D. Katsoulas. Reliable Recovery of Piled Box-like Objects via Parabolically Deformable Superquadrics. In *9th IEEE Int. Conf. on Computer Vision*, volume 2, pages 931–938, 2003.
67. E. Lopez-Damian. *Grasp Planning for Object Manipulation by an Autonomous Robot*. PhD thesis, Laboratoire d'Analyse et d'Architecture des Systmes du CNRS, 2006.
68. E. Lopez-Damian, D. Sidobre, and R. Alami. Grasp Planning for Non-Convex Objects. In *36th International Symposium on Robotics*, 2005.
69. S. El-Khoury and Anis Sahbani. Handling Objects By Their Handles. In *IROS-2008 Workshop on Grasp and Task Learning by Imitation*, 2008.
70. N. Amenta, S. Choi, and R. Kolluri. The Power Crust. In *6th ACM Symposium on Solid Modeling and Applications*, pages 249–260, 2001.
71. G. Barequet and S. Har-Peled. Efficiently Approximating the Minimum-Volume Bounding Box of a Point Set in Three Dimensions. *Journal of Algorithms*, 38:91–109, 2001.
72. B. Draper and A. Lionelle. Evaluation of Selective Attention under Similarity Transforms. In *International Workshop on Attention and Performance in Computer Vision*, pages 31–38, 2003.

73. A. Tavakoli Targhi, E. Hayman, J.O. Eklundh, and M. Shahshahani. The Eigen-Transform and Applications. In *Asian Conference on Computer Vision*, pages 70–79, 2006.
74. B. Rasolzadeh, A. Tavakoli Targhi, and J.-O. Eklundh. An Attentional System Combining Top-Down and Bottom-Up Influences. In *Workshop on Attention and Performance in Computational Vision*, pages 123–140, 2007.
75. B. Rasolzadeh. Interaction of Bottom-up and Top-down Influences for Attention in an Active Vision System. Master's thesis, Royal Institute of Technology, Stockholm, Sweden, 2006.
76. A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in Context. In *International Conference on Computer Vision*, pages 1–8, 2007.
77. T.M. Strat and M.A. Fischler. Context-Based Vision: Recognition of Natural Scenes. In *Asilomar Conference on Signals, Systems and Computers*, pages 532–536, 1989.
78. T.M. Strat and M.A. Fischler. The Use of Context in Vision. In *Workshop on Context-Based Vision*, 1995.
79. Y. Hu, X. Xie, W-Y Ma, L-T. Chia, and D. Rajan. Salient Region Detection using Weighted Feature Maps based on the Human Visual Attention Model. In *IEEE Pacific-Rim Conference on Multimedia*, pages 993–1000, 2004.
80. A.K.C. Wong and P.K Sahoo. A Gray-Level Threshold Selection Method based on Maximum Entropy Principle. *IEEE Trans. Systems Man and Cybernetics*, 19:866–871, 1989.
81. J. Bohg and D. Kragic. Grasping Familiar Objects Using Shape Context. In *International Conference on Advanced Robotics*, 2009.
82. S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.
83. M. Craven and J. Shavlik. Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems (NIPS-8)*, pages 24–30. MIT Press, 1995.
84. D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen. Comprehensible Credit Scoring Models using Rule Extraction from Support Vector Machines. *European Journal of Operational Research*, 183(3):1466–1476, December 2007.
85. W. T. Townsend. The BarrettHand Grasper – Programmably Flexible Part Handling and Assembly. *Industrial Robot: An International Journal*, 27(3):181–188, 2000.
86. K. Huebner, S. Ruthotto, and D. Kragic. Minimum Volume Bounding Box Decomposition for Shape Approximation in Robot Grasping. In *IEEE International Conference on Robotics and Automation*, pages 1628–1633, 2008.
87. R. B. Rusu, Z. C. Marton, N. Blodow, M. E. Dolha, and M. Beetz. Functional Object Mapping of Kitchen Environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3525–3532, 2008.
88. R. Triebel, R. Schmidt, Ó. Martínez Mozos, and W. Burgard. Instance-based AMN Classification for Improved Object Recognition in 2D and 3D Laser Range Data. In *International Joint Conference on Artificial Intelligence*, pages 2225–2230, 2007.
89. K. Huebner, M. Björkman, 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*, volume 5008 of *LNAI*, pages 13–22. Springer-Verlag, 2008.
90. A. Saxena, L. Wong, and A. Y. Ng. Learning Grasp Strategies with Partial Shape Information. In *23rd AAAI Conference on Artificial Intelligence*, pages 1491–1494, 2008.
91. D. Aarno, J. Sommerfeld, D. Kragic, N. Pugeault, S. Kalkan, F. Wörgötter, D. Kraft, and N. Krüger. Early Reactive Grasping with Second Order 3D Feature Relations. In *ICRA Workshop: From Features to Actions*, pages 319–325, 2007.
92. C. Borst, M. Fischer, and G. Hirzinger. Grasp Planning: How to Choose a Suitable Task Wrench Space. In *IEEE Int. Conference on Robotics and Automation*, pages 319–325, 2004.
93. K. Huebner and D. Kragic. Selection of Robot Pre-Grasps using Box-Based Shape Approximation. In *IEEE Int. Conference on Intelligent Robots and Systems*, pages 1765–1770, 2008.

94. A. Nealen, T. Igarashi, O. Sorkine, and M. Alexa. FiberMesh: Designing Freeform Surfaces with 3D Curve. *ACM Transactions on Computer Graphics, SIGGRAPH 2007*, 23(3), 2007.
95. A. T. Miller and P. K. Allen. Graspit! A Versatile Simulator for Robotic Grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004.
96. A. Morales. *Learning to Predict Grasp Reliability with a Multifinger Robot Hand by using Visual Features*. PhD thesis, Dep. of Computer and Engineering Science, Univ. Jaume I, 2004.
97. J. Speth, A. Morales, and P. J. Sanz. Vision-Based Grasp Planning of 3D Objects by Extending 2D Contour Based Algorithms. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2240–2245, 2008.
98. S. Geidenstam, K. Huebner, D. Banksell, and D. Kragic. Learning of 2D grasping strategies from box-based 3D object approximations. In *Proceedings of Robotics: Science and Systems*, Seattle, USA, June 2009.
99. C. Geib, K. Mourao, R. Petrick, N. Pugeault, M. Steedman, N. Krüger, and F. Wörgötter. Object Action Complexes as an Interface for Planning and Robot Control. In *IEEE-RAS Humanoids-06 Workshop: Toward Cognitive Humanoid Robots*, 2006.
100. J. Gibson. The Theory of Affordances. In R.E. Shaw and J. Bransford, editors, *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Erlbaum, NJ, 1977.
101. M. Bar. Visual Objects in Context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.
102. M. Bar and S. Ullman. Spatial Context in Recognition. Technical report, Mathematics & Computer Science, Weizmann Institute of Science, 1993.
103. A. Torralba and P. Sinha. Statistical Context Priming for Object Detection. *International Conference on Computer Vision*, pages 763–770, 2001.
104. L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor. Learning Object Affordances: From Sensory–Motor Coordination to Imitation. *IEEE Trans. on Robotics*, 24(1):15–26, 2008.
105. M. Lopes and J. Santos-Victor. A Developmental Roadmap for Learning by Imitation in Robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(2):308–321, 2007.

Jeannette Bohg holds a M.Sc. in Computer Science from the Technical University Dresden, Germany and a M.Sc. in Applied Information Technology with a focus on Art and Technology from Chalmers in Göteborg, Sweden. Since 2007 she is a Ph.D. candidate in the field of Computer Vision and Robotic Grasping at the department of Computer Science at the Royal Institute of Technology (KTH).

Carl Barck-Holst received his M.S. degree in Computer Science from the Royal Institute of Technology (KTH), Stockholm, Sweden in 2003. He currently holds a scholarship position for Ph.D. studies in the department of Computer Science at KTH. He has also held several positions in industry focusing on IT-architecture, transformation and technical program management. His research interests include machine learning, computer vision, and robotics.

Kai Huebner received his M.S. degree in Computer Science in the Natural Sciences from the University of Bielefeld, Germany, and his Ph.D. degree in Computer Science from the University of Bremen, Germany, in 2001 and 2006, respectively. Since 2007, he is a Research Assistant in Computer Science at the Royal Institute of Technology (KTH), Stockholm, Sweden. His research interests include computer vision and grasping for robot applications and service robotics.

Maria Ralph is a post-doctoral fellow in the department of Computer Science at the Royal Institute of Technology (KTH), in Stockholm, Sweden. She received her B.Sc. degree in Computer Science from Ryerson Polytechnic University, Toronto, Canada, and her Ph.D. degree in Systems and Computer Engineering from the University of Guelph, Guelph, Ontario, Canada. She has also worked as a Software Developer and Researcher for MDA Space Missions. Her research interests include robotic grasping, user-adaptive robots, human-robot interaction, and machine learning.

Babak Rasolzadeh recieved his M.Sc. degree in Electrotechnical Engineering from the Royal Institute of Technology (KTH) in Stockholm, Sweden. He holds an scholarship position for Ph.D. studies in the field of Computer Vision at the department of Computer Science at KTH. He has worked as an research engineer at the National ICT Australia working with Machine Learning in Computer Vision. His interests are active vision and online object learning.

Dan Song received her B.S. degree in textile engineering from Donghua University, China, and her Ph.D. degree in Biomedical Engineering from the University of Southern California (USC), Los Angeles, United States of America in 2000 and 2008 respectively. She is currently a post-doctoral fellow in the department of Computer Science at the Royal Institute of Technology (KTH), Stockholm, Sweden. Her research interests include control of reaching and grasping in both neuroscience and robotics, and particularly how to adapt neural models of grasping control to robotic applications, and learning by imitating human demonstrations.

Danica Kragic received her B.S. degree in mechanical engineering from the Technical University of Rijeka, Croatia, and her Ph.D. degree in computer science from the Royal Institute of Technology (KTH), Stockholm, Sweden in 1995 and 2001, respectively. She is currently a professor in computer science at KTH and chairs the IEEE RAS Committee on Computer and Robot Vision. She received the 2007 IEEE Robotics and Automation Society Early Academic Career Award. Her research interests include vision systems, object grasping and manipulation and action learning.