TECHNISCHE UNIVERSITÄT WIEN
VIENNA UNIVERSITY OF TECHNOLOGY

DISSERTATION

Getting Past Passive Vision – On the Use of an Ontology for Situated Perception in Robots

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung von

A.o.Univ.Prof. Dipl.-Ing. Dr.techn. Markus Vincze
Institut für Automatisierungs- and Regelungstechnik (E376)

eingereicht an der Technischen Universität Wien
Fakultät für Elektrotechnik und Informationstechnik

von

MAG.PHIL. DIPL.-ING.
MATTHIAS J. SCHLEMMER
9825330
Schönburgstraße 9/3A
1040 Wien, Österreich

Wien, im März 2009

Abstract

Computer vision has achieved many highly complex techniques to solve specific problems with specific solutions. Research on robot companions, however, deals with varying and cluttered environments that pose challenges to perception, which cannot be overcome with niche solutions. For example the detection of object concepts defined by the common substrate that all instances share as opposed to the detection of specific instances only is of crucial necessity for autonomous robots.

This dissertation aims to discuss important issues and topics with respect to the goal of bringing these disciplines closer together again. It thus starts by exemplifying how work in computer vision is usually approached: by presenting a texture edge detection and an interest point tracking technique. We then give thorough theoretical considerations pertaining to what cognitive robotics – as a foundation for robot companions – is in need of and propose a theoretical framework in terms of cognitive functions that bridge to computer vision techniques. Besides giving a clear position on what the presuppositions of such a view on vision and robotics are, one of our main concerns that lie behind our functional approach, is the continuous incorporation of various disciplines. Hence, we give comprehensive references to philosophy, psychology and cognitive science in general throughout this thesis.

In particular, we present intentionality, prediction, abstraction, symbol binding and generalisation as functions that seem to be of crucial importance for situated perception in autonomous robots, as far as our focus on visual cognition is concerned. Another key finding is that those functions find their common connecting point in an ontology which might hold the needed object concepts, as well as further semantic information that allows for connecting to further cognition.

Experiments include the presentation of how such an ontology can be set up, introducing the use of a recently defined ontology language from the semantic web movement. We show how visual abstraction, one of the cognitive functions found to be important, can be performed and how different vision techniques can be exploited while simultaneously retaining grounded and situated scene understanding. In this way, we can show how an ontological representation accounts for explicit knowledge handling, the deployment of the presented cognitive functions, and the needed transition from quantitative data to qualitative information.

Zusammenfassung

Das Feld der Computer Vision hat in den letzten Jahrzehnten eine beachtliche Anzahl an komplexen Methoden hervorgebracht, die es ermöglichen, spezifischen Problemen mit spezifischen Lösungen zu begegnen. Das Gebiet der Heimrobotik sieht sich ihrerseits mit alltäglichen und variierenden Umgebungen konfrontiert, welche große Herausforderungen an ihre Wahrnehmungsfähigkeit stellen, die nicht mit Nischenlösungen zu behandeln sind. Beispielsweise ist für die autonome Robotik die Detektion von Objektkonzepten, durch das von allen Instanzen geteilte Zugrundeliegende definiert, eine entscheidende Fähigkeit, die im Gegensatz zur Detektion einzelner Instanzen ganz andere Schwierigkeiten birgt.

Diese Arbeit zielt darauf ab, wichtige Aspekte in Hinblick auf das (Wieder)Zusammenführen von Computer Vision und Robotik (als Artificial Intelligence) zu diskutieren. Sie beginnt mit typischen Beispielen, die zeigen sollen, wie in der Computer Vision üblicherweise Problemstellungen behandelt werden: mit einer Methode zur Texturkantenerkennung und einer weiteren zur Objektverfolgung mittels interest points. Anschließend wird die Brücke zu theoretischen Überlegungen zur kognitiven Robotik (als Grundlage für die Heimrobotik) geschlagen und ein theoretischer Rahmen im Sinne von kognitiven Funktionen vorgestellt. Die Vorannahmen hierfür werden dargelegt und es wird der Überzeugung gefolgt, dass eine solch komplexe Aufgabenstellung nur unter Miteinbeziehen unterschiedlicher Fachgebiete möglich ist. Daher wird in dieser Arbeit wiederholt auf Philosophie, Psychologie und Kognitionswissenschaft verwiesen.

Namentlich werden *Intentionalität*, *Prädiktion*, *Abstraktion*, *Symbol Binding* und *Generalisierung* als jene Funktionen vorgestellt und diskutiert, welche als entscheidend für die situierte Wahrnehmung in autonomen Robotern erachtet werden. Ein weiterer wissenschaftlicher Beitrag ist die Begründung einer von diesen Funktionen geteilten Ontologie, welche sowohl die Objektkonzepte als auch zusätzliche semantische Information zu beherbergen vermag, welche die Verbindung zu weiterer Kognition darstellt.

In einem abschließenden praktischen Teil werden Experimente vorgestellt, die einen möglichen Aufbau einer solchen Ontologie zeigen, wobei eine rezente Ontologiesprache (*OWL*) mit der Computer Vision verknüpft wird. Es wird gezeigt, wie visuelle Abstraktion, eine der genannten kognitiven Funktionen, auf diese Weise implementiert werden kann und wie verschiedenartige Vision-Techniken ausgenützt und verbunden werden können unter konsistenter Aufrechterhaltung eines Szenenverständnisses. Außerdem wird dadurch deutlich, wie auf diese Weise eine ontologische Repräsentation sowohl zu einem klaren Umgang mit explizitem Wissen als auch zur Entfaltung der präsentierten kognitiven Funktionen und der benötigten Transition von quantitativen Daten zu qualitativer Information führen kann.

Acknowledgements

I would like to express my gratitude to a number of people who supported me while working on this thesis. I would like to thank Markus Vincze for the freedom he provided me in following my own ideas and for his continuing support. Special thanks go to my external reviewer Georgi Stojanov from the American University in Paris for his time and encouragements. I would also like to thank Aaron Sloman from the University of Birmingham for endless fruitful discussions and ideas – I very much enjoyed my time in England. My gratitude additionally goes to Markus Peschl from the University of Vienna who drummed up my interest for cognitive science and thus influenced the direction of my thesis topic.

I am grateful to my office mates of the RobSens group who continuously cared for an enjoyable working environment. Thanks to Sara Lampe for being my „Phone-a-Friend“ whenever a question regarding English grammar came up. I am also indebted to the general public for funding my work via various EU projects.

Last but not least I would like to thank Denise for believing in me and my undertaking. Special thanks also go to my brother who always had an open door whenever I needed a downtime in the Alps and to my mother and sister for their unconditional support.

Contents

List of Figures

What we call the beginning is often the end
And to make an end is to make a beginning.
The end is where we start from.

Little Gidding
T.S. Eliot

Chapter 1

Introduction

Writing a PhD-thesis is a bit like composing a symphony. Both are cognitive acts, both are time-consuming undertakings and in both cases, the final product should sound good, be long enough, creative and – last but absolutely not least – something new. This is all difficult enough, but things get even trickier when the "œuvre d'art" tries to counterpoint established methods and to bring different styles together. As with music, science has a huge variety of separate disciplines, each following its own code of conduct, having its own experts and approved methods. Lateral thinking can run into a variety of troubles, starting from a rising demand of knowledge concerning the State of the Art to acceptance difficulties in the separated communities.

However, despite the difficulties just mentioned, this thesis tries to be of the latter kind (without claiming, however, to be as ingenious as Beethoven's Ninth) – introducing a different view on the recently famous movement of "cognitive vision", which includes the buzzword "cognitive" that – as other disciplines did, too – the computer vision community introduced to emphasise that purely technical approaches have come to a crucial point where *more cognition* (or – to use another buzzword that will be made clear – *more intelligence*) needs to be incorporated. Clearly, thinking in terms of autonomous robots, i.e., of *robotic* vision, these issues get even more important. This thesis thus aims at conceptualising vision as part of the cognition of a robot companion. Our approach is to consult disciplines that are interested in the workings of autonomous agents, such as artificial intelligence, cognitive science, developmental psychology, and philosophy.

In the following we will give the context of the work (Section 1.1) and the involved disciplines (Section 1.2), show different stances that can be taken with respect to the topic (Section 1.3) and give a personal motivation (Section 1.4). Then, we will present the research questions along with an outline of their answers in Section 1.5 and provide the structure of the thesis (Section 1.6). We conclude with final introductory remarks (Section 1.7).

1.1 Context of the Work

This thesis is situated in the broad field of "cognitive robotics". Our specific objective is a robot companion, i.e., an autonomously acting robot capable of accomplishing tasks for its human operator in a home environment. As our group is mainly dealing with perceptual issues, our concrete goal is the development of *situated vision*[1], denoting a robotic vision approach that aims at integration into the overall system – as in contrast to "computer vision" which is mainly interested in solving concrete problems in very narrow domains. In the following we will analyse these different fields in more detail and end this section with an overview of the requirements imposed by them.

1.1.1 The Field: Cognitive Robotics

We will devote a whole section to the clarification of the notion of "cognition" (along with "intelligence" and "consciousness"); for now it suffices to say that cognition in our understanding points to a principle that allows for situation-adapted behaviour, which could – from a third person perspective – be called "intelligent". Consequently, definitions of cognitive robotics or cognition show a broad spectrum of subproblems. For example, in [Com08], cognitive robotics is defined as being

> [...] concerned with integrating reasoning, perception, and action within a uniform theoretical and implementation framework (using methods drawn from logic, probability and decision theory, reinforcement learning, game theory, etc.). It is quite a young field of research. [...] Complex applications and the need for effective interaction with humans are increasing demand for robots that are capable of deliberation and other high-level cognitive functions. Models from cognitive science and techniques from machine learning are being used to enable robots to extend their knowledge and skills. Combining results from mainstream robotics and computer vision with those from knowledge representation and reasoning, machine learning, and cognitive science has been central to research in cognitive robotics.

As can be seen, "cognitive robots" thus need to be capable of quite a lot of tasks. Of course, parts of the mentioned capabilities (such as reasoning from perceptual data) might be found in systems that would not be called "robots" in the first place. Hence, cognitive robotics can be classified as a subset of "cognitive systems" which additionally comprise systems that do not have any ego-motion but rather deploy so-called "ambient intelligence". An example would be a room that reacts to the presence of a human in a personalised manner. The paradigm of ambient intelligence is related to the idea of ubiquitous computing, i.e., the constant presence of supporting artificial intelligence.

1.1.2 The Objective: Robot Companions

Our objective is the idea of a robot butler that serves the human. In our opinion, such a robot companion can only materialize when a minimum of "cognition" is present,

[1] This term has also been used by [Pyl01] for "[...] bidirectional contact with the world [...]".

therefore robot companion research can be classified as subset of cognitive robotics. Such a robot has a clearly defined niche: to serve next to the human, in the human operator's environment and hence having its own "life-world" there. Example tasks of such a system could include:

1. bringing the user an everyday item ("*James, bring me my cup!*"),
2. helping with the shopping,
3. noticing an emergency, such as getting help if the user tumbled and is unconscious,
4. cleaning the flat, or
5. feeding the cat.

And lots more – the number of possible scenarios is endless. The challenge lies in the requirement to prepare the robot with the necessary functions to learn these tasks and to adapt what it already knows to a given situation. It is obvious that one cannot expect all of these situations including all possible eventualities to be preprogrammed. *Situatedness* [Mat02] is one of the most prominent preconditions for such flexibility. With situatedness we refer to the fact that the system is embedded in its environment such that all information is bound to the current situation. This leads us to our specific focus, "situated perception", which we will tackle in the next section. Before doing this, we should discuss why robot companions should be located in the wider field of cognitive robotics.

Robot companion research is dealing with a lot of high-level goals, such as those mentioned above. Breaking these down to single steps and to specific engineering problems, brings out subtasks that are themselves challenging and even (partly) unsolved – such as mapping the user input to the representation in the robot, navigating in partly unknown environment, dealing with unexpected problems, planning, finding objects, grasping them, and so forth. But there is one thing that unites these challenges: The overall goal of a robot that is working intelligently in a home environment. Intelligence, we would say, can only be judged from a third person perspective, and almost always in a situation-dependent manner. Therefore, if we accept the working hypothesis that cognition is the overall principle that takes care of situation-adapted behaviour, and a robot companion needs to show such a behaviour, then a robot companion needs to "implement cognition" (for now we can stick to this very crude description) and is hence an example of a cognitive system.

1.1.3 The Goal: Situated Vision

One of the hard topics in designing a robot companion concerns the requirements for its perception. Vision is often judged to be of special importance to the perceptual apparatus of humans (e.g., [Nør94, p.191]), and as it is a distal sensor it seems very well suited for robotic tasks. However, work on computer vision and especially on vision for robotics during the last years has made clear that it is not at all a simple task to extract meaningful information from pixel-grids. We will see that one of the hard issues is exactly at the bottom of the word "meaningful". Human vision is far more complex than doing pattern matching.

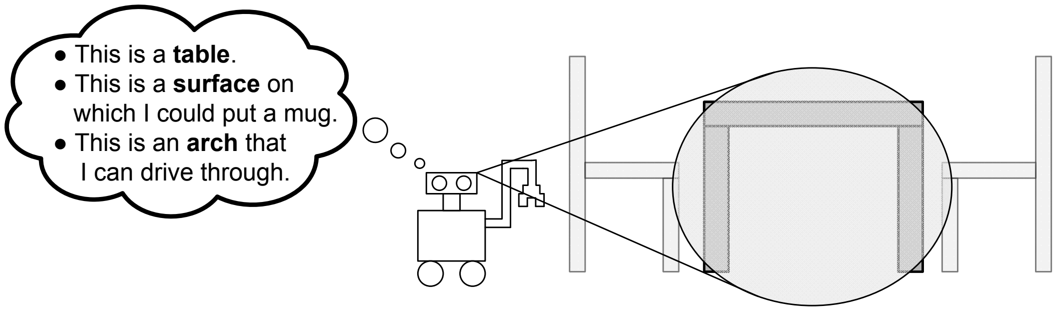

• This is a **table**.
• This is a **surface** on which I could put a mug.
• This is an **arch** that I can drive through.

Figure 1.1: Cognitive issues for a robot companion: where does computer vision end, where does cognition start?

In the last few years, the buzzword "cognitive vision" has become popular in order to account for the fact that the computer vision community is in need for cognitively inspired methods. However, it is often forgotten that cognition should not only serve as inspiration on *how* to tune the algorithms but rather on *what vision is for*. This means – again, spoken from a roboticist's point of view – that vision needs to be embedded into cognition (perception is in fact usually judged to be part of cognition in cognitive science). We refer to this embeddedness into the overall cognition of the robot and consequently the adaptability of vision to the current situation as *situated vision*. It uses the term *situated* which has been deliberately borrowed from the movement of "situatedness" as it shall emphasise that perception can only be thought of as perception *of* something *in* a given scene. Furthermore, it shall underline that perception of an autonomous agent (such as a robot companion) is always bound to a task and therefore situation- and context-dependent. We would like to make an effort that the notion of "situated vision" gets widely introduced into the community of researchers that are dealing with companion robots.

Consequently, one of the major stances of this thesis is that vision is wrongly assumed to be 2- or 3-dimensional (e.g., depending on monocular or stereo vision). It is rather *multi-dimensional*, comprising dimensions such as anticipations, theories, assumptions, primings, and much more. Vision can start from capturing pixels, but it must not end there – at least not for robotic objectives. It might be enough for pattern matching techniques like those needed for robotics as purely technical solution (as in industrial inspection tasks), but it definitely is not for vision of a robot companion. What is needed is a structured account of what the requisites from cognition are and what the glue is like that holds everything together. *From there*, we might be able to see vision with different eyes.

Having said that cognition is situation-dependent, on the vision side, the task can – very bluntly spoken – be reduced to "feeding all subtasks with adequate information" – finding objects, landmarks and features – in different lighting conditions and under various circumstances and clutter.

Here is the crux: *adequate*. Consider Figure 1.1: This fictitious home robot could have a variety of observations in this scene. Due to the chairs it might *predict* that there is a table to be seen; maybe it *generalises* all tables it has ever seen and can "re-cognise" this one as being an instance of table. Another possibility would be that it *abstracts* the surface of the table to an affordance for putting the mug down that it has retrieved earlier. And so on and so forth. We have deliberately chosen this set of possibilities as we will later see that all of them, *prediction*, *generalisation*, and *abstraction* seem to be necessary for "intelligent behaviour". All of these functions can be considered to be "cognitive", and robot companion research is in need for exactly these issues; furthermore, the line between "pure vision" and "higher-level cognition" becomes fuzzy quite fast: Obviously the meaning this table gets *for the robot* is dependent on some factors, e.g., its "priming" or its actual capabilities.

In short, all this implies that perception is situated in an overall system without the need to be confronted with a random image in which it should "re-cognise" learnt objects. All this points to the fact that specific assumptions can be made that reduce the ill-posed "general" to a more realistic "guided" scene understanding system. And the guidance is given by higher-level planning, another open problem yet crucial requirement for autonomous behaviour. Obviously, representations used by vision and those used by such an overall system need to be compatible and quite rich. With other words, we need to keep the whole system in mind.

1.1.4 Wrap-Up: Context, Requirements, and Goal

We have defined that situated perception would be a necessary means for a robot companion, i.e., a robot that serves next to the human in a home environment. There are some disciplines that evolved in the last years that deal with various aspects of such a robot butler. The top of Figure 1.2 tries to sketch our view how these disciplines relate to each other. Below, the different requirements of the fields are listed, where each area of research inherits the requirements of the broader fields. The list is, of course, not exhaustive and it furthermore argues on various layers. It shall give an impression of the breadth of necessary subfunctions and -capabilities.

Of course, the narrower the field, the more specific and capacious the requirements get. This is due to the fact that, e.g., a "definition" of a cognitive system can only be made via the least common denominator, such as "reacting intelligently" and "integrating various sorts of information". However, as can be seen from the wording, those are also the hardest to "just implement" as they are formulated quite fuzzily.

Having said this, nevertheless the boundaries between the requirements are not strict, of course, and shall only underline the special importance in the mentioned subfield. E.g., recovery from failure is also important for systems without self motion and there are cognitive robots that are not robot companions yet still moving, i.e., they need localisation as well.

In particular, for our objective of a robot companion, we grouped visual and non-visual aspects. It is obvious that perceptual issues are tightly interwoven with non-perceptual capabilities, e.g., predicting what is to be seen next will need the "understanding" of the situation (top-level) and in some cases additionally even the capability of human-robot communication (partly non-visual).

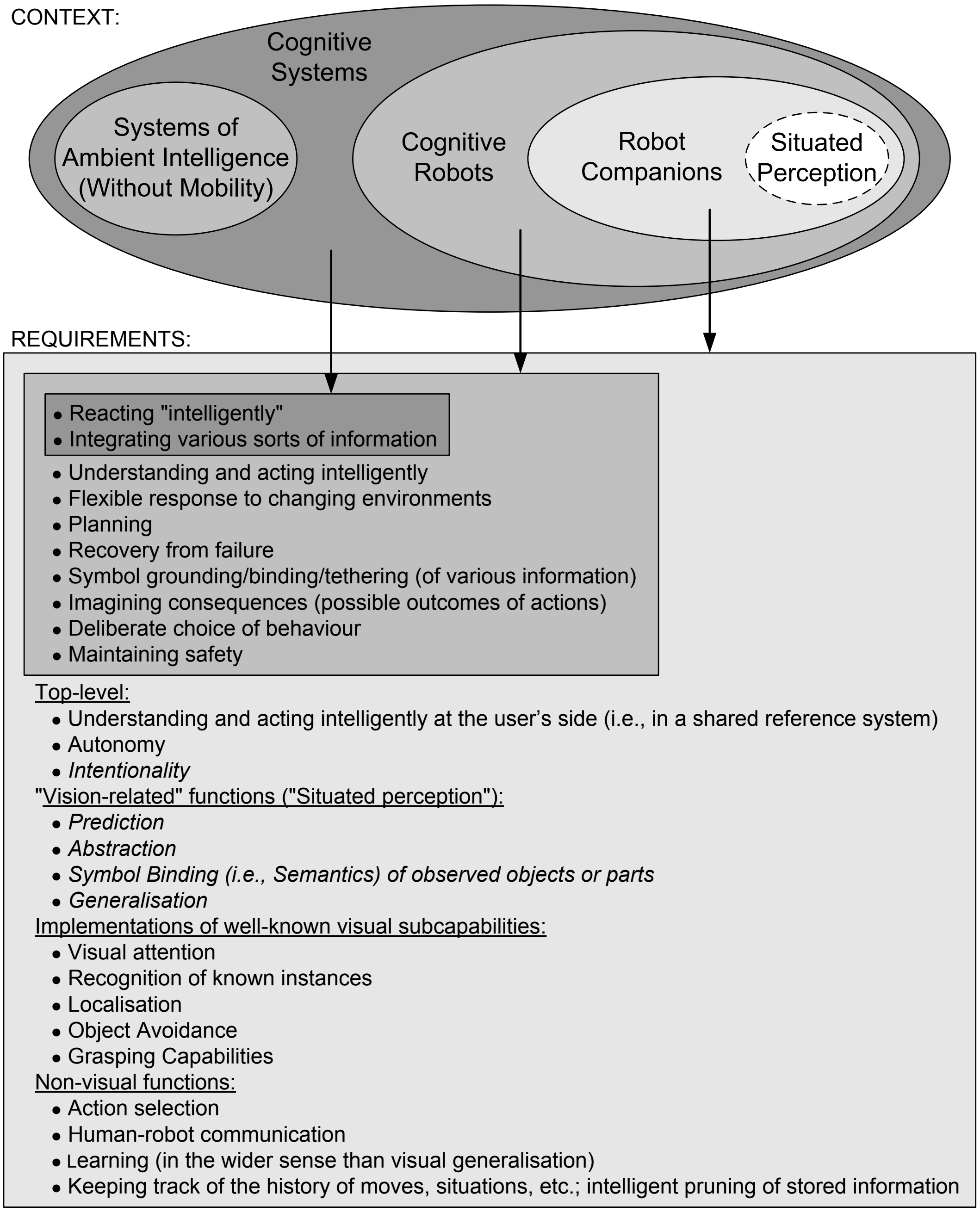

CONTEXT:
Cognitive Systems
Systems of Ambient Intelligence (Without Mobility)
Cognitive Robots
Robot Companions
Situated Perception
REQUIREMENTS:
• Reacting "intelligently"
• Integrating various sorts of information
• Understanding and acting intelligently
• Flexible response to changing environments
• Planning
• Recovery from failure
• Symbol grounding/binding/tethering (of various information)
• Imagining consequences (possible outcomes of actions)
• Deliberate choice of behaviour
• Maintaining safety
Top-level:
• Understanding and acting intelligently at the user's side (i.e., in a shared reference system)
• Autonomy
• Intentionality
"Vision-related" functions ("Situated perception"):
• Prediction
• Abstraction
• Symbol Binding (i.e., Semantics) of observed objects or parts
• Generalisation
Implementations of well-known visual subcapabilities:
• Visual attention
• Recognition of known instances
• Localisation
• Object Avoidance
• Grasping Capabilities
Non-visual functions:
• Action selection
• Human-robot communication
• Learning (in the wider sense than visual generalisation)
• Keeping track of the history of moves, situations, etc.; intelligent pruning of stored information

Figure 1.2: Context and requirements: Robot companion research is located in the wider fields of cognitive robotics and cognitive systems. Each field has its own requirements (not an exhaustive list); a growing number the narrower the field gets.

Our main concern here is to point to the crucial perceptual capabilities of a robot companion as this is our objective in this research project. There are requirements for the robot companion that are "especially" concerned with vision or perception in general, such as: predicting what is likely to be perceived next, abstracting seen things to more general object concepts, or straight-forwardly recognising objects especially those of the shared reference system. However, we will see that also "intentionality" (task-directedness) or "symbol binding" (giving percepts situational meaning) is of crucial importance because they provide tight interaction between visual and non-visual requirements. We put "well-known subcapabilities" in the Figure as well to point to special fields in computer vision research, on which we will not concentrate on in this thesis. Nevertheless, they, too, are important requirements for a future robot companion.

The just mentioned requirements of intentionality, prediction, abstraction, symbol binding, and generalisation (all in italics in the Figure), will be of special interest for our goal of situated perception in order to achieve the top-level goals of autonomy and we will subsequently call them "cognitive functions". This approach is somehow borrowed from *functionalism* [Put60] yet not identical to it (as will be explained). The functional approach will give us the opportunity to bridge between various disciplines and to take a position where implementation issues can take a back seat and the overall framework comes predominantly into view. We will devise the necessity and the role of each of the functions and will come up with the necessary solution that they connect to a shared ontology.

1.2 Methodology

It has been said that the thesis is located in the field of cognitive robotics with focus on visual perception of a robot companion. In order to get a grip on the complexity of the far goal of this kind of research, namely an autonomous, flexible, "intelligent" robot that acts and extends in varying environmental conditions, understands scenes and events and is possibly even able to interact with other non-human or human agents, an interdisciplinary approach is needed. This is also reflected by the different requirements listed in Figure 1.2. Cognitive robotics, in general, is the attempt to bring research of human intelligence and of artificial intelligence together [Chr99], with the aim of generating a new approach to enable robots to have higher-level cognition capabilities (such as beliefs, complex goals, attitudes and the like).

It is obvious that cognitive robotics is in principle an inherently interdisciplinary approach. Talking about complex goals, for example, immediately brings up the question what the input from the environment looks like and additionally, we have to think about the minimal set of cognitive functions that is involved in reasoning over this input. Talking about the environment, we need to include thoughts about internal representations, e.g., of object concepts[2]. This will be at the heart of this thesis: The question

[2] "Intelligence without representation" as Rodney Brooks has proposed [Bro91] is dubitable when it comes to explicit planning of future events, weighing options with simulated outcomes, and the like – which we do not want to exclude at the outset.

on reducing the gap between explicit knowledge (needed for some of the higher-level cognitive capabilities) and vision (which is the means to build up this knowledge in the first place or to later guide its processing).

The disciplines that we judge to be important for the undertaking just described are: artificial intelligence, cognitive science, psychology (especially developmental psychology), philosophy and, of course, computer vision. Computer vision alone is too fixated on low-level image plane processing, artificial intelligence often takes perceptual input as given and cognitive science has strongly shifted its focus towards neural networks during the past years. We will later on argue for a rehabilitation of symbolic representation, although the specific implementation is in principle not in the fore here. Anyway, all those disciplines together could provide an approach to tackle the hard theoretical and practical issues involved in our undertaking. Philosophy, finally, might be able to point to questions and considerations that have been thought about for years, decades and centuries – and although on a more theoretical or conceptual level, they might give us the essential hints on where to look for solutions.

To illustrate the different foci and questions and to expose our foundational starting points, we will shortly summarise the different lines of research taken from various disciplines in the following. We will constrain our review to the major strands that we started off with and will not go into details for reasons of space and clarity.

1.2.1 Vision Science and Computer Vision

Computer vision – as far as methods for dealing with objects are concerned – has achieved a great deal of specialised methods for specialised tasks. A typical example would be industrial applications, where visual inspection tasks can be considered as State of the Art. The usual goals of algorithms developed tackle the advancement of speed, of recognition rate, or of precision. Specific measurements are introduced and the rise of vision databases during the last years shows the growing wish of comparability of algorithms. Computer vision techniques for robotics have mainly focused on adapting stand-alone solutions for a robotic platform, following the same “scientific canon” as in pure computer vision. We would like to mention two foundational findings from vision research that computer vision builds on: Gibson’s “ecological optics” and Marr’s “sketches”.

J.J. Gibson’s *ecological optics* put focus on the question what kind of information the proximal stimulus provides about the distal stimulus [Pal99, p.53 f]. Being influenced by Gestaltism, Gibson underlined the importance of *texture gradients* on the retinal image (“ambient optical array”, AOA) as well as of information coded in the *optical flow*. Consequently, he argued that one must not forget the fact that visual perception in agents has evolved while the agent has been on the move (seeking food, etc.). A crucial insight is that the information of the outside world by means of optical flow is related to the optical flow constituted by the head’s movement.

The seminal work on which still most approaches of computer vision build, is the one of David Marr [Mar82]. He coined the modular view of visual perception as being composed of different “sketches” [BGG03, p.80 f]: The “primal sketch” represents the changes of light intensity over the image space, including information about boundaries (as they are edge steps in intensity). The "$2\frac{1}{2}$D sketch" then already comprises information about visible object surfaces (distances and orientation with respect to the observer) whereas the final, "3D model" stage determines the seeing of solid object shapes and helps for identifying their classes. The popularity of this data-driven approach is additionally supported by Marr's own demand that any theory about visual perception has to be expressible in an algorithm and thus constituting a computational approach to vision.

For our purposes, we will concentrate on how to integrate computer vision approaches into a cognitive system, namely our objective of a robot companion. This implies that the view on vision as being tightly bound to the necessities posed by the environment (such as Gibson pointed to), will be of overall concern for our considerations by simultaneously being aware that *up to a certain degree*, the bottom-up view of Marr is still necessary for a synthesising science such as computer vision.

Current computer vision research focuses on appearance-based approaches which try to solve the problem of recognising instances by learning methods. Either the system checks the current image against a specific instance (classical object recognition), or it uses stable features that can be found in a set of training images that constitutes an object category ("generic object detection").

The first part of the title of this thesis reads "Getting Past Passive Vision". By passive we refer to the classical computer vision approach to tackle specific problems without grounding the data used, instead of an integration of the perceptual information into an overall system. In order to further clarify what we mean with passive vision and to contrast it with what we call "situated vision", we will show two works done in the field of computer vision (Part I, i.e., Chapter 2 and Chapter 3). This shall also help to underline that by no means we want to talk down the use of computer vision techniques such as object recognition or tracking![3] They all have their niche. However, these techniques must not be confused with the truly needed methods for artificially intelligent systems. We will see that for a "cognitive system", usual object handling by computer vision techniques do not suffice, which can be illustrated by the following points:

- They lack semantic binding.
- They are constrained to handling specific object instances[4].
- They lack a common formalism for using vision results for higher level processing.
- They usually use knowledge about the world implicitly in the code instead of making it available to other (cognitive) subsystems.

The need for common ("passive") vision techniques is anyhow obvious. First, there are a lot of applications where the handling of specific instances is needed and where

[3] Likewise, when we later on use the term "low-level" computer vision, we do *not* want to imply "simple" or "uncomplex" regarding what we think about this kind of work. "Level" solely refers to a coarse view of cognition as a hierarchy of processing steps.

[4] Note that even generic object detection systems rely on "similar" appearances in a subset of the training images in order to be successful.

semantic information is not of importance. The two chapters of Part I show examples. Second, it must not be forgotten that low-level vision techniques are a necessary foundation for higher-level scene understanding. Without reasonably reliable vision preprocessing output, high-level tasks are hard if not impossible to achieve. This might as well hold for living agents, yet visual input can be much more distorted than (currently) in machines (e.g., concerning the viewing angle, coloured filters, rotation and scaling, etc.). We must be aware that the dream of the personal butler will not materialise until perception is rich enough to deal with cluttered environments, novel objects and all the richness of the world we humans are confronted with every day. As the neurobiologist Jaak Panksepp once put it (with reference to the need for a basic emotion system) [Pan07]: “Computer will be garbage in → garbage out. The brain: Garbage [...] in and → total coherence out.”

Therefore, in computer vision science, we have two main needs for good algorithms. The first need is the provision of highly specialized and nearly always functioning methods to solve *a specific problem*. Part I deals with such systems. The second need stems from another goal, namely to have the basic means for higher-level reasoning. Ideally, of course, there are overlaps between the two routes. We would argue, however, that robustness is more important than perfect accuracy in the latter systems. This argument is in line with what Sloman mentions in [Slo08a, p.74]: “A child or animal who is confronted with something uncertain, because of poor lighting, bad eyesight, dirty windows, occluding objects, distance of objects may not be able to adopt any of those engineering solutions. [...] However a child can learn other ways of coping with uncertainty, by using the epistemic affordances in the environment to remove or reduce uncertainty.”

Summarised, vision *alone* cannot succeed – there is more to vision than meets the eye.

1.2.2 Artificial Intelligence

The focus of research in artificial intelligence (AI) is on the workings of the mind (artificial or natural). John McCarthy’s definition of AI is that it is “[...] the science and engineering of making intelligent machines, especially intelligent computer programs” [McC00]. The notion of intelligence, however, is itself controversial and has changed with the years – we abstain from narrating the history of AI[5]. In short, after realising the obvious limitations of “generally intelligent” machines (such as Newell and Simon’s “General Problem Solver” [NS63], which reduced human reasoning to logical computation), the focus changed to a situated and context-dependent account of intelligence, e.g., in Rodney Brook’s “subsumption architecture”.

Brooks’ seminal work on *embodiment* tackled issues related to the study of “complete, integrated agents” that deploy autonomous intelligent behaviour physically in an environment, leading to the necessity of putting focus on their bodies. *Situatedness* is given when disturbances (such as people walking by or sensor drifts) happen and nevertheless function is equally well obtained [PS01, p.320]. Brooks’ organising principle of

[5] Excellent introductory books are [RN03] and [Nil98].

choice was the "subsumption architecture" which targets at the decomposition of complex modules into simple basic ones arranged in layers, the higher layers "subsuming" the lower ones. Each layer constitutes a complete route from input to motor output, and each layer takes the decision of the next lower layer into account [Cla01, p.101].

One could say AI has itself evolved to a more mature approach towards intelligence. Yet at the bottom line, the hunt for "intelligence" is still open. Apropos evolving: Work on evolving agents in AI ("A-Life") has succeeded in simulating robots that change and adapt their shape, learning systems, and even social interaction. Mostly simulation is used here, which means that typical difficulties of visual perception – such as recognising objects under varying lighting conditions, image blur and the like – are circumvented. Moreover, perception often takes a back seat – and if it does not, it is taken for granted. Objects are simulated as being detected as such, leaving the crucial question open on how an object gets a (semantically laden) object within the agent in the first place. The discussion on "symbol grounding" impressively shows deep problems with "getting meaning into the system" (cp. [Har90]). In A-Life, questions like the degree of fitness or the survival rate are in the fore.

For us, AI research is interesting as it follows the general goal of getting an artefact work "intelligently". Especially our later references to higher-level cognition (including planning, learning, action selection, and the like) point to typical topics of AI research.

1.2.3 Cognitive Science

Cognitive science tries to connect some disciplines by having the *human* as a whole in mind. Nevertheless, AI and cognitive science share a considerable part, both with respect to their goals as to their paradigms. Sticking to our example of *embodiment*, one might say that AI focuses more on "programs" as we saw in McCarthy's quotation further above, and consequently the embodiment movement was primarily accepted in cognitive science. Nowadays, cognitive scientists often presuppose that cognition is influenced if not even determined by the form of the agent's (human's) body. For the case of robotics, this seems trivial (except for some purely simulated approaches), as a robot's functionality is automatically affected by its sensory capabilities.

Let us have a look at a classical definition of cognitive science, which additionally shows us the variety of tasks we can find here [OL90, p.xi]:

> *Cognitive Science* is the study of intelligence in all of its forms, from perception and action to language and reasoning. The exercise of intelligence is called cognition. Under the rubric of cognition fall such diverse human activities as recognizing a friend's voice over the telephone, reading a novel, jumping from stone to stone in a creek, explaining an idea to a classmate, remembering the way home from work, and choosing a profession. Cognitive processes are essential to each of these activities; indeed, they are essential to everything we do.

So cognition seems to be *everything.* As disappointing for the hunt of a definition that can be implemented in a straight-forward manner as this may seem, it points us to a very important aspect: It is highly probable that the crucial questions regarding human perception are deeply nested within the bigger framework of human cognition. So one obvious question for us is why to start from pattern matching when talking about perception in an autonomous artificial agent. Cognitive science helps us to take a more general view on intelligence – especially from an analytic perspective (as opposed to the synthesising perspective of AI).

1.2.4 Developmental Psychology

Psychology investigates the experiencing and behaving subject. This implies various research topics, such as clinical psychology, psychological diagnostics, or developmental psychology. The latter can be considered one of its major strands and is of special importance for us. The rationale is that one branch of developmental psychology is interested in the development and use of percepts and concepts (mainly in infants), which will get important in the course of this thesis.

Although we will not tackle learning explicitly, which is for obvious reasons important in developmental psychology, we do find interest in how experience and already known concepts influence and might even guide future perceptions. Furthermore, also behaviour and intelligence is studied by developmental psychologists. On a meta-level, we also have interest in the "nature-nurture debate", which tackles questions on what knowledge or apparatus a child is equipped with at birth as opposed to the knowledge that he/she acquires throughout life.

1.2.5 Philosophy

Finally, we take philosophy as one of the disciplines important for our objective of situated perception in a robot companion. At first glance it might seem quite atypical or even disturbing to include a science like philosophy in a technical thesis. However, we firmly believe that philosophy has a special role, both historically and with respect to the way of questioning.

Philosophy has an outstanding role in the history of sciences, often termed as "mother of all sciences". Its unique approach can provide researchers with new perspectives by simultaneously retaining the big picture. The usual suspects of disciplines involved in the study of cognitive systems, namely biology/neuroscience, developmental sciences, psychology and artificial intelligence, have evolved from philosophy when sciences started to split into more detailed directions. Nevertheless, their roots are philosophical.

Second, at least since the ground-breaking article "Epistemology Naturalized" by Willard van Orman Quine [Qui69], natural sciences and philosophy started to move together again. Quine rejected the notion of a "prima philosophia" (first philosophy) in the sense of a philosophy that locates itself in the elitist atrium of sciences, but rather stood up for philosophy being seen as a science like any other, simply arguing in a more abstract sense [Sch98, p.153].

The last and maybe the most exciting point is that philosophy has the tendency to pose the hard and uncomfortable questions. There are some underlying issues in cognition that are however only tackled explicitly in philosophy (e.g., the discussion of free will). As we will see later, exactly the problem of what makes a cognitive system cognitive – sometimes with other terms – is one of the oldest philosophical topics.

Similar to psychology, philosophy has a variety of subdisciplines, such as epistemology, ontology, ethics, or language. We will focus on the first two:

Epistemology (in the philosophical sense) is the study of the nature and the acquisition-possibility of *knowledge*. It is thus connoted with a *subjective account* on what the world is like. The most radical account of this subjectivity has been taken by *constructivism* that strictly objects to any attempt of correlating one's knowledge about the world with the actual nature of the world of which we are unable to know anything. However, as [Sch87] points out, constructivists usually do not follow an *ontological solipsism*, i.e., the position that there actually exists no outside world at all.

Ontology is the philosophical strand that asks about what there is. It is thus the "objective" side of the coin in the hunt for an understanding of the nature of the world we live in. Ontology is concerned with questions about basic entities and their relations. The distinction between epistemology and ontology raises the complex question of how concepts are formed with respect to rational vs. empirical evidence. Especially the interplay between the perceiving subject and the objective world has been of much discussion throughout the history of philosophy. We will later focus on an attempt (in the aftermath of Immanuel Kant's work) to reunite both sides of the coin (idealism and realism, but also epistemology and ontology) by focusing on the *relation* between them.

Summarised, there are issues in our hunt for an intelligent robot companion that philosophy might help to address. We will encounter later on notions like "thing-in-itself" in the context of abstraction in cognitive systems and "intentionality" as the philosophical notion for a system's bias due to the task currently pursued. Also semantics are to be looked at – a topic well-known to artificial intelligence. To put it in a nutshell, it is our firm belief that the look at epistemological positions has the ability to bring up new approaches and perspectives to an understanding of cognition in engineering as well.

1.2.6 Wrap-Up: Methodology

As could be seen, there are a variety of disciplines involved if we want to take our research project seriously. It goes without saying that an exhaustive investigation of all the questions implied would cover more than one thesis. This short introduction shall merely underline that an interdisciplinary approach is imperative and motivate the search for a clear theoretical framework (cp. Chapter 4) on which the last practical chapter of the thesis (5) will then build. For example, artificial intelligence and cognitive science have evolved side by side; robotics and computer vision, however, despite their apparent relation to these fields have deepened their work more into solutions for detailed problems. The gap that had opened is exactly the reason for not having general AI-strategies that work robustly with vision data.

We can put it a different way: The goal of this thesis is to focus on perception, but instead of asking about how to enrich current techniques with "more cognition", we take a step back and question the requirements from a higher, more system-view like

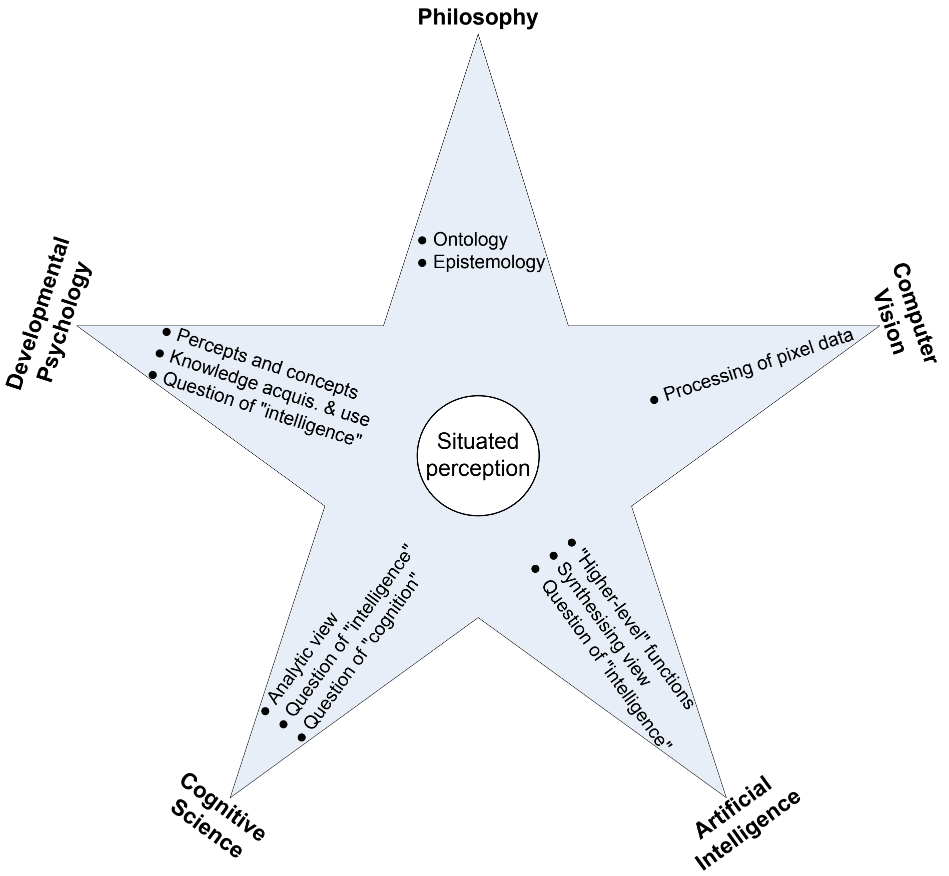


Figure 1.3: Involved disciplines and major issues taken.

position. We will investigate what capabilities are needed (we will call them “cognitive functions”) and how we can embed perception within them. The goal is not to do *cognition on top of vision* but rather *vision as part of cognition*. This will even allow us to every now and then link to sciences that enable totally new perspectives that are usually inaccessible for a synthesising science like robotics. The advantage for engineers lies in the analytical view, which engineers usually do not have at hand – both an introspective (or psychoanalytical) and an objective (neurological) view. The far goal is that engineers can take other approaches to better understand the mind and – hopefully – to break this down to the functional level where synthesis can begin. Of course, we will also tackle questions on the use of anthropomorphisms.

Figure 1.3 summarises the main points from the disciplines listed above that we will have a look at in the course of this work.

1.3 Stances of the Thesis

Having outlined the different disciplines that provide motivational questions for our research project, we could ask about the different stances that can be taken up on these

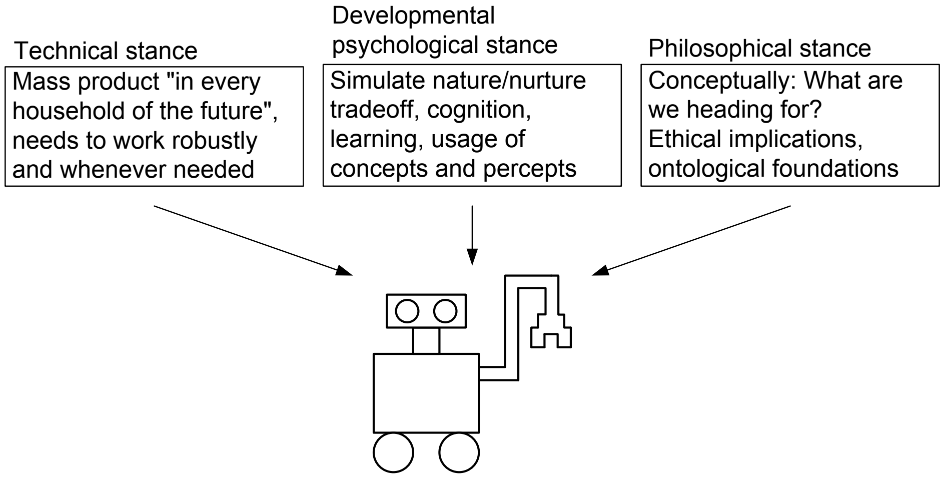

Technical stance
Mass product "in every household of the future", needs to work robustly and whenever needed
Developmental psychological stance
Simulate nature/nurture tradeoff, cognition, learning, usage of concepts and percepts
Philosophical stance
Conceptually: What are we heading for?
Ethical implications, ontological foundations

Figure 1.4: Stances of the thesis.

issues (Figure 1.4 depicts them):

Firstly, from the purely technical stance, which views robotics as being only “robotics as industry”, interest lies in the wish to get an artefact work in a truly “intelligent way”, judged from a third person – human’s(!) – perspective. The far goal is the mass construction of a robot companion that is able to work in a home environment next to the user; among other things by simultaneously giving arbitrary objects the right situation-dependent meaning (this is for sure a very crude description).

Secondly, we can take the developmental psychologist’s stance where interest lies in the wish to simulate (parts of) those systems that we would widely judge to be really cognitive ones, namely humans (and maybe sometimes animals). Furthermore it can be rewarding to learn from organisms’ learning methods, because truly intelligent systems possibly need to develop within their specific ecological niches in order to deploy their intelligence. The other way round, a good model might deliver new insights about how living organisms might work.

Finally thirdly, there is the philosophical stance, asking questions concerning our topic that have been asked for the last thousands of years, yet that have not been answered fully due to the complexity of the issues involved. Furthermore, the philosophical stance allows us to take a step back and view the undertaking from a very conceptual standpoint.

Not surprisingly, this thesis is guided by all three stances, borrowing results but also approaches from all of them. The reason is trivial to say yet tricky to follow: It is our firm belief that we need to think all of them together to see the big picture and not to get lost in details. The downside, however, is that working on the big picture leaves too less space to work on the details – but a truly “cognitive” robot asks to be designed on all levels. Hence, this thesis tries to think conceptually on quite a theoretical level and only gives some paradigmatic showcases in order not to loose the technical aspects out of sight and to show feasibility where necessary. The main argument, however, stays quite high level as we should start to tackle the right problems, not solutions.

1.4 Personal Motivation

I would like to finally give a personal motivation which also explains the structure of this thesis. My own work started with typical computer vision tasks, such as texture edge detection (Chapter 2) and model-based tracking (Chapter 3). Those techniques, as much as they are interesting from a purely technical point of view, are, however, limited in their generality and are too detail-affixed for the far goal that our group pursues: a truly autonomous robot companion that assists humans in their everyday tasks. The simultaneous studies of philosophy enhanced my way of questioning for more than well-designed algorithms and to pose meta-questions that resulted in considerations about human functions that seem to be important for our goals (cp. Chapter 4).

Theory alone, however, is not enough for a student pursuing a technical PhD. Therefore, besides the vision techniques done previously (Part I), preliminary showcases have been programmed that investigate different aspects of the theoretical considerations. They are meant to investigate...

- ...how visual *abstraction* can be performed on the border between quantitative computer vision output and qualitative descriptions in the ontology.
- ...how a *logic-based* ontology (borrowed from computer science) could serve as *cognitive* ontology in an agent. Here, the major challenge lies in the bridge from “low-level” vision output to “high-level” (symbolic) objects or proto-objects.
- ...how different layers of vision techniques (from appearance-based instance recognition to qualitative visual abstraction) can be *glued together* for a comprehensive robot companion scenario.

The showcases allowed me to dig deeper into foundational questions that in my opinion future robotics research will need to pose, tackle and take serious – at least if it claims to be based on considerations about cognition. This also implies questions that inquire the philosophy of science and how theories about our world are established and maintained. In my opinion, this goes much further than paradigmatic discussions and debates on principles – as could be observed in the history of AI concerning the correct paradigm of implementation: whether the “only right way” is symbolic, subsymbolic, dynamical or embodied and situated. Of course, at some point these discussions are necessary, but without a common notion of what we want to achieve, they are meaningless. This might point to one of the advantages of work that has a robot companion as far goal: We always talk about a task, about something that we wish to have got done. The personal driving force of this thesis has been the wish to think laterally on issues of computer vision for cognitive robotics with having this pragmatic approach in mind. Interdisciplinarity, for me, seems to be the approach of choice to do that – although this implies facing the ghosts of more than one discipline.

1.5 Research Questions

We will now summarise the research questions and results of this thesis.

The meta-questions posed by our approach are:

- What can we learn from various disciplines that investigate or tackle “cognition”?
- Can we locate a common “layer” of argumentation?
- Where are useful starting points for an interdisciplinary work between philosophy and computer vision for robotics? Searching for motivation in philosophy has been the goal of artificial intelligence for quite some time, yet the special focus on visual perception still has a lot of open issues.
- What is *computer vision* missing in order to be considered as situated robotic vision (i.e., integrated in the overall cognitive system)?

These and similar questions will dominate our theoretical Chapter 4, where we will have a look on relevant philosophical traditions, on anthropomorphic considerations, and, of course, on various notions connected with our research project. We will result in a *functional* approach, which, once taken, implies further questions:

- What is the *minimal set of functions* that seems to be necessary for situated perception in a robot companion?
- What is the *glue* that holds this minimal set together?
- How does the knowledge look like that we need to store?

The answers to these questions will also be given in Chapter 4 in terms of five cognitive functions especially relevant for situated perception (intentionality, prediction, abstraction, symbol binding, and generalisation). As connecting glue, we will present our view on requirements for a shared ontology and as knowledge we will plead for the storage of concepts – both object concepts in the visual sense and general concepts that allow for linking various sorts of information.

To put it in a nutshell, Chapter 4 introduces a general theoretical framework for situated vision of a robot companion which we transfer into a showcase implementation in Chapter 5. There we see that it allows us to take specialised vision tools and to link their output to higher-level functions usual to AI (such as planning or reasoning). It is this step from quantitative “in-the-code” *data* to qualitative, explicit *information.* This is highly related to the attempt to get from the detection of a “proto-object” (e.g., a cylinder) to the notion of “object” (e.g., a mug).

Additionally, Chapter 5 puts the following novel goals forward:

- The use of a State of the Art ontology on the basis of which we can discuss drawbacks and features of the use of a logical formalism to robotic vision.
- The presentation of a possibility to link quantitative pixel data to qualitative information in an ontology.
- Connected to this, we show how visual abstraction can be performed, in order to solve tasks as depicted in Figure 1.5: We show how the “constituting substrate” (which all of these different arches have in common) could be stored in an ontology and exploited when confronted with an image the robot has not seen before.

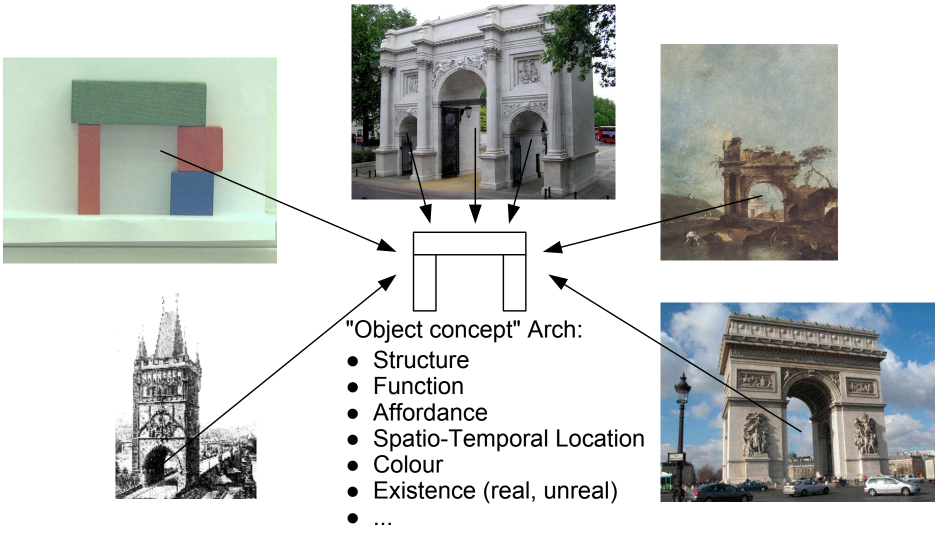


Figure 1.5: Variety of arches – what is the common constituting substrate?

- We are showing how complex tasks that most probably are in need for various, situation-dependent vision tools could be done by integrating cognitive functions and explicit information in an ontology. This is what we would then call "situated perception".

Besides these main topics, this thesis presents novelty with respect to a texture edge detection method (Chapter 2) as well as a monocular interest point tracking approach (Chapter 3). Both chapters shall demonstrate how "typical computer vision tasks" are tackled and solved. They will later on serve as paradigmatic examples for our questions on the bridge to cognitive functions.

1.6 Structure of the Thesis

The organisational structure of this thesis – also following the author's own realisations and growth along the hard topics of the field – is somewhat different to usual academic works. A well-known strategy is to begin with the most general assumptions and questions of the field and then to specialise to a narrow domain, gradually detaching from the profound questions underlying the whole field of investigation. Moreover, it could be claimed that this is the only feasible way that "synthesising sciences" such as computer science or probably any technical science can perform. We believe, however, that one of the main problems when talking about cognitive robotics and about perception for situated robotics in general, lies in the fact that for too long, existing vision tools have been more and more fine grained and then – as the "cognitive revolution in computer science" has taken place, those tools have been taken and it has been tried to just augment some cognitive methodology on top of them. A. Sloman pointed out that "[...] we have not yet developed visual mechanisms that come close to matching those produced by evolution. In part, this is because the requirements for human-like visual systems have not been analysed in sufficient depth [...] E.g. there is a vast amount of research on object recognition that contributes nothing to our understanding of how 3-D spatial structures and processes are seen or how information about spatial structures and processes is used, for instance in reasoning and acting" [Slo08b].

Instead, we claim that starting from the specialised techniques and knowing what is possible, from which the author of this thesis has also started (see Part I), we need to step back and understand what cognition actually means and what role perception plays. K. Nelson wrote with reference to the so-called "nature-nurture-debate": "It is possible to break the system apart, to examine brain function, for example, but then the system is no longer functioning as a system (i.e., it is no longer active in its environment)" [Nel99, p.188]. We believe that this holds fully for vision as part of a cognitive system as well. We need to investigate the *system*, *then* we can have a look at vision techniques and judge which tools are appropriate and which are possibly well enough developed for our purposes.

Therefore, this thesis will pursue the following path:

It starts from special vision techniques in order to exemplify what computer vision is typically concerned with in Chapters 2 and 3, which together form Part I of this thesis.

Part II starts afterwards with Chapter 4 – the central component of this thesis – where we will present our general theory. We begin with the challenges of cognitive robotics – especially pertaining to vision, of course – (Sections 4.1 and 4.2) which then entails that we have to straighten out what our underlying assumptions and constraints are (Section 4.3), including the inseparability of ontology and epistemology in robotics. Similarly important is the clarification how we understand notions such as "intelligence" or "cognition" (Section 4.4) and finally why this all leads to our conviction that we need to start from a functional layer for synthesising the findings from the various disciplines (Section 4.5). Besides presenting the five functions that we focus on (Sections 4.5.2 to 4.5.6), Section 4.5.6 will particularly also bring our plea for using symbolic computing. Section 4.6 is finally concerned with our proposal to use an ontology for glueing the different visual and non-visual functions together, which entails considerations on *what* to store and *how* to store it. Here again, we will focus on object concepts that can be perceived. Section 4.7 is concerned with preparing for the last chapter of this thesis by introducing what "ontology" means in the computer science sense and presenting what an innate specification of abstract concepts mean. Section 4.8 summarises the outcomes so far.

Chapter 5 is then again a "practical" chapter, in which we investigate the use of an ontology in the strict engineering sense, which includes the representation of object concepts with predicate logic. We will present the State of the Art tools and methods that we use (Section 5.2), before we show our basic example that allows us to show how visual abstraction can be performed in order to get *from quantitative pixel data* and computer vision techniques to *high-level qualitative information* (Section 5.3). Before facing typical critics in Section 5.5, we will tackle the complex and typical example of a robot companion that tries to bring the user his/her cup, which not only resembles our group goal but shows how an ontological approach like the one taken in this thesis might help to achieve this goal (Section 5.4).

In the conclusion (Chapter 6) we will summarise the main topics of this dissertation and discuss missing parts and possible further extensions.

1.7 Final Introductory Remarks

It must be said that the goal of this thesis is not to provide a concise and consistent theory of human cognition. It is rather pursuing a possible advancement in cognitive robotics that is *on a functional level* inspired by findings from other disciplines. Consequently, we are not aiming at using human intelligence as blueprint or – even worse – trying to build an artificial human-like creature. It is rather our will to abstract cognitive functions, that we as designers judge to be probably useful for our long-term goal of a cognitive robot companion and try to investigate how far those could be implemented. Our position is that it is completely wrong to try to rebuild humans. Besides the general "What for?" question, ethical implications would be questionable.

The second part of the title of this thesis, the explanation of which we have deprived the reader so far, tries to summarise our main goals: It is called "On the Use of an Ontology for Situated Perception in Robots". Besides the focus on robotics, we are concerned with a theoretical account of cognitive functions that we judge to be crucial for autonomous robots of the future – but only focusing on perceptual issues, not on robotics in general. In this regard, the notion of "ontology" becomes fundamental and we aim to show that such a common "glue" is indispensable.

At last, the most important disclaimer: As can be seen, a lot of the points mentioned above are quite theoretical. All too often, comprehensive theoretical work is left out, especially in the technical domain, leading to confusion and an absent basis on which can be built on. We hope to provide a countersteering approach to that for the field of cognitive robotics. For this thesis, we tried to bring conceptions from different disciplines together so to provide a basis for future work regarding the use of ontologies in computer vision – with the far goal of a truly autonomous robot companion in mind. This implies that a lot of "results" are not code snippets or ready-to-use tools, but rather often take the form of *ideas* and *conceptions*.

Related Work will be given in the respective chapters as the broad topic of this thesis would make it confusing to have a separate section dealing with all the related work for the different subtopics. Where necessary, we provide pointers to further work along the lines presented in this thesis. All images used, unless indicated otherwise, have either been taken by the author himself or are public domain.

Part I

Low Level Robotic Vision

2.011 It is essential to things that they should be possible constituents of states of affairs.

Tractatus logico-philosophicus
Ludwig Wittgenstein
[Wit98]

Chapter 2

Texture Edge Detection

As mentioned in the Introduction, this and the next chapter present works that tackle typical tasks encountered in computer vision. We will not give considerations on situatedness or ontologies but rather focus on a specific problem and its specific solution – the intended application area is therefore quite narrow. Hence, this part gives two characteristic examples for technical solutions that are highly tuned for a specific task.

In this chapter we present work on texture edge detection applied to a model-based tracking tool which has been developed at our institute [Ayr03]. Those systems – although having the mentioned narrowness in their applicability – are often needed for the purpose of exactly handling objects, e.g., when doing visual servoing or grasping. Especially for the domain of home robots this is relevant as once the object of interest is identified, it must be tracked accurately for allowing its handling. Higher-level semantics are naturally taking a back seat for the execution of assignments like this.

Our task [SV05] was to include an additional cue for edge detection based on texture information in order to enhance robustness in case of cluttered surfaces. We chose to use statistical feature matrices as they are suited to detect changes in texture, which may occur on the transition from object to background. The model of the object to be tracked is given a priori manually by the user in a quantitative description.

After reviewing the State of the Art in texture for detection and tracking (Section 2.1), we shortly present "Vision for Robotics", the tool for which the texture cue has been developed (Section 2.2). Then, the chosen mathematical representation (Section 2.3) and the algorithmic implementation (Section 2.4) are given. Results (Section 2.5) and a discussion (Section 2.6) conclude this chapter.

2.1 State of the Art

The work described in this chapter finds its application in a model-based tracking tool, described in Section 2.2. We will confine the review of the State of the Art to texture for edge detection and leave out the huge pile of literature that deals with tracking of objects through a sequence of images in general. A comprehensive survey of monocular model-based 3D tracking approaches can be found in [LF05]. A short general overview of texture analysis can be found in [TJ05].

[FP03, p.287] describe texture as "[...] a phenomenon that is widespread, easy to recognise and hard to define." Definition is especially hard because the appearance and even the existence of texture are dependent on the scale at which it is looked at. We will see that for the goals presented here, these general problems can be eased. For the sake of completeness, let us have a look at the usual tasks for which texture is used (following [ibid.]):

- *texture segmentation*, i.e., breaking an image into components with nearly the same appearance,
- *texture synthesis*, i.e., constructing large regions of texture from small example images, and
- *shape from texture*, i.e., recovering surface orientation or shape from image texture.

Additionally, *texture classification* can be mentioned, which deals with assigning a specific texture class to a portion of an image – applications are typically inspection tasks, e.g., for aerial observations and land-use categorisation – a reference work in this context is [HSD73]. This, however, requires a priori knowledge of the classes of textures that can be observed. The (usually large) co-occurrence matrices, used by [ibid.], can further be used to compute features ("Haralick-features") that reduce the complexity of the image patch description for faster classification.

Texture analysis methods can broadly be divided into statistical and syntactical methods. Whereas the latter are based on the fact that a (large) textured surface has some kind of repetition of (small) so-called "textons" or its composition is at least following some kind of *grammar* and are therefore well suited for classification and synthesis tasks, the former describes the appearance of the texture based on features extracted. The resulting feature vector can then be used for comparing different textures, and hence for classification. Statistical methods comprise both those based on spatial frequencies (e.g., Gabor filters) and those using directly the spatial arrangement of pixels (e.g., co-occurrence matrices of intensity distributions). We will concentrate here on statistical methods that have been applied to detect texture edges. This is, of course, a related task of texture segmentation.

[MB03] use the compass operator, originally introduced by [RT99] for comparing colour distributions and performing edge detection in RGB images, and apply it to the output of a multi-dimensional image that is generated by different texture filter outputs. The specific choice of the underlying texture filters is not constrained; the mentioned examples are all statistical: spatial (Laws filter, Gabor filters) or frequency-based (such as Fourier, discrete cosine or wavelet transforms). With this approach, the authors are able to detect edges between regions that have the same colour distribution, yet are perceptually totally different – a pure compass filter could not detect those edges. [GD06], too, use a bank of filters (Gabor filters and wavelets), resulting in a total of eight images that are subsequently smoothed and are input to a self-organising map (SOM) that reduces the eight-dimensional feature vectors of each pixel to a scalar. On

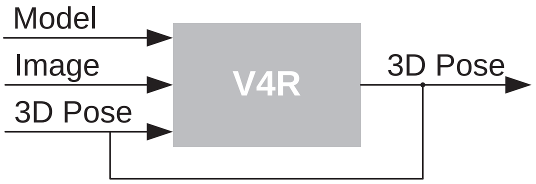


Figure 2.1: Top-level view on *V4R*. Left: cues as input; right: 3D-pose as output.

the resulting image, canny edge detection and edge linking is performed. This way, advantages from Gabor filtering and wavelet transform respectively can be combined. [SDF04] aim at higher speeds than the approaches presented so far. Their goal is the integration of texture boundary detection into a real-time pose estimation algorithm. Therefore their approach is similar to the one presented in the following. Underlying is the use of a Markov model applied to 1-D search lines perpendicular to the assumed object boundary. The texture is modelled as a statistical process generating a sequence of pixels. Consequentially, two distinct textures, one being on the object and the other formed by the background, generate a changepoint at the boundary of the object. For the choice of the priors that define which process has generated the given intensity distribution on either side of the changepoint, the authors apply Markov models of 0th and 1st order, meaning that the intensity distribution is either fully independent or each pixel intensity depends on the preceding one only. They show that both models deliver accurate results, however, the 0th order Markov model is more sensitive to initialisation.

2.2 *V4R* – Vision for Robotics

In order to support the rationale for using the chosen approach of texture edge detection, we will shortly present the functionality and basic principle of Vision for Robotics (called *V4R* in the following), a model-based object tracking tool for which the texture edge detection has been developed. A detailed explanation of the cue integration method used can be found in [Ayr03], related work on model-based object tracking is given in [VSGA05].

2.2.1 Basic Functionality

The goal of *V4R* is the tracking of rigid objects in unknown surroundings in order to deliver the 3-D pose of the object as accurately as possible, which can be subsequently used as input for a robot's end effector. Main requirements are speed (tracking at camera frame rate) and robustness. As the usage of monocular (2-D image) cues is prone to deliver inaccurate results, cue integration is performed that takes additional, non-image cues (such as the previous pose or the model of the object) into account. Figure 2.1 shows the different inputs (cues) on the left and the output pose on the right side of *V4R*.

The model-cue is provided by the predefined model of the object that shall be tracked and is stored in a file containing a wire-frame description. This implies that only a subset of possible objects, namely those describable by such a representation, can be handled. Pose is the cue resulting from the previous tracking step or the initialisation respectively. The "image cues" stem directly from the 2-D image and are of interest for us here.

Image cues are edge segments that are possible candidates for parts of the object boundary. As we are dealing with a 3-D object, a "visibility check" and a "projection" technique are first performed that deliver the approximate regions of interest (ROIs) where (only the visible) object boundaries are likely to be found in the current 2-D image. Those ROIs are then warped in a manner that the (supposed) edge lies vertically, which speeds up the detection later on. For keeping feature processing time constant, a windowing approach similar to the approach of Hager and Toyama [HT98] is applied. This ROI-window is also used for the texture edge detection described in Section 2.4. After extracting one or more edgel candidates[1] in each of the different modes (colour, intensity, texture), voting following a modified RANSAC-scheme (Random Sample Consensus, [FB81]) is applied to extract one or more feature candidates. Those candidates are input to a global voting scheme that takes the model cue into account. As more than one view candidate is generated this way, the final decision is made by including the previous pose (i.e., the pose cue).

2.2.2 Texture as Additional Cue for Edge Detection

While most of the techniques developed for texture analysis have a quite large portion of the object surface in mind (e.g., because classification is their goal), we need to deal with small patches. The reason is that we are aiming at detecting edges in a small local neighbourhood and rely on the higher-level ("model-" and "pose-") cues that *V4R* already provides when it comes to the whole object. *V4R*, as explained, performs cue integration in order to enhance robustness. Hence, the use of a texture edge detection cue aims at helping out in situations where illumination and lighting conditions weaken the edge detection that is purely based on colour or intensity.

For a seamless integration of a texture cue to *V4R*, the following constraints are important to be met. The chosen technique needs to be:

1. *Fast:* *V4R* aims at tracking objects at real-time; therefore the use of complex filter banks is not suited.

2. *Local:* Only local textural properties are needed, so a large portion of constraints applying to texture classification is weakened (the object's model and pose are provided from higher-level cues already implemented in *V4R*).

3. *Free from presuppositions* concerning the actual texture: We do not want to pre-model the texture but rather detect texture edges without any a priori knowledge about the appearance of the object, as the model only consists of the *shape* in a wire-frame manner.

[1] "Edgel" denotes an edge pixel of locally maximum gradient of a cue value.

4. *Able to deliver most probable edgel locations:* This means that one or more than one candidate pixel for an edge can and should be delivered – this is important in order not to lose the correct edge in presence of texture edges *on* the object.

Note that this approach does not demand the same texture across the whole object but rather only a textural difference to the background near the boundary of the object.

2.3 Statistical Feature Matrices (SFM)

The basis for the approach taken in our exercise comes from the paper of [WC92] in which the authors introduce the use of a statistical feature matrix (SFM) that measures the statistical properties of pixel pairs at several distances. They emphasise three advantages in comparison with co-occurrence matrices normally used:

- The size of the matrix is given by the maximum distance used (and does not depend on the depth of the image),
- expansion of the matrix is easy and
- some physical properties can be evaluated.

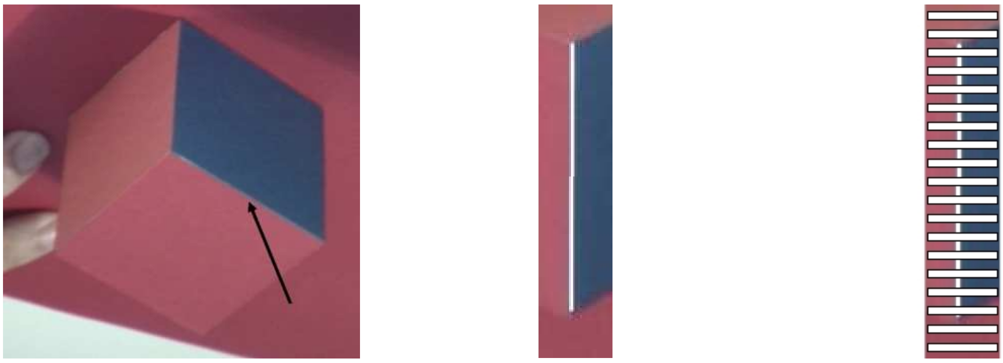
SFMs store distances between pixels in a matrix following well-known statistical features, which has been used by our approach as well. These features comprise contrast (`con`), covariance (`cov`), and dissimilarity (`dss`) and are defined as follows [ibid.] (with $E(\cdot)$ being the expectation operator):

> Let δ be the intersample spacing distance vector and η be the average gray-level of an image I, the δ *contrast*, δ *covariance*, and δ *dissimilarity* are defined as
> δ *contrast*: $\mathrm{CON}(\delta) \equiv E\{[I(x,y) - I(x+\Delta x, y+\Delta y)]^2\}$
> δ *covariance*: $\mathrm{COV}(\delta) \equiv E\{[I(x,y) - \eta][I(x+\Delta x, y+\Delta y) - \eta]\}$
> δ *dissimilarity*: $\mathrm{DSS}(\delta) \equiv E\{|I(x,y) - I(x+\Delta x, y+\Delta y)|\}$

These statistical features are computed for each cell of the SFM which serve as descriptions of the texture and are later on used for texture classification in the original paper.

2.4 Edge Detection Using SFM

In the cited paper of Wu and Chen, SFMs are used for texture analysis in order to classify Brodatz textures as well as ultrasonic liver images. However, for our purpose, the given algorithm shall be used for detecting edges. Although the goal is different, the advantages, especially the first two, mentioned in Section 2.3 can be exploited for our real-time tracking environment: Computation time of SFMs is only dependent on the chosen size of the matrix and can usually be kept far below 255 – the number of grey levels in an 8-bit image. This means that the size of the matrix can be chosen according to available time.

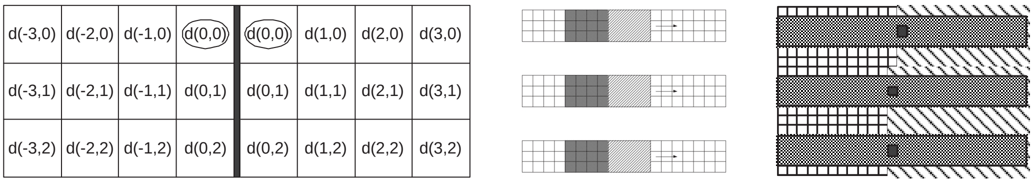


(a) The arrow points to the edge currently searched for.

(b) Warped window; the edge is manually enhanced for demonstration purposes.

(c) Strip positions in which the texture matrices (SFMs) will slide.

Figure 2.2: Tracking a colour cube: one of the edges that are warped for detection.



d(-3,0)	d(-2,0)	d(-1,0)	d(0,0)	d(0,0)	d(1,0)	d(2,0)	d(3,0)
d(-3,1)	d(-2,1)	d(-1,1)	d(0,1)	d(0,1)	d(1,1)	d(2,1)	d(3,1)
d(-3,2)	d(-2,2)	d(-1,2)	d(0,2)	d(0,2)	d(1,2)	d(2,2)	d(3,2)

(a) Layout of the matrices.

(b) Sliding matrices.

(c) Edge pixels found.

Figure 2.3: Sliding window technique for texture edge detection: (a) Layout of the two adjacent modified SFMs; (b) The matrices are sliding from left to right in each strip (three are shown); (c) Edge pixels are detected in each strip according to maximal matrix-difference.

Hence, this approach seemed suited and we adapted it in order to be used in *V4R*. Recall the requirements listed in Section 2.2.2: Texture edge detection needs to be fast and local, without the need of presuppositions on how the actual texture looks like. Texture is just another cue, so that the detection technique needs to be transparent for the overall system. Therefore, we directly used the ROI that is also used for colour or intensity edge detection and that has been warped so that the assumed edge lies vertical (as described in Section 2.2.1). Figure 2.2 shows a simple scene in which the edge of a colour cube shall be detected (this example only demonstrates how the matrices will be used and does not include strong texture information). On the left (Figure 2.2(a)), the whole image from the tracking sequence is shown, Figure 2.2(b) displays the warped ROI of the supposed edge along with the manually enhanced edge and Figure 2.2(c) depicts the "strips" in which the SFMs will later slide.

As explained, cue integration in *V4R* is performed by taking hypothetical edge pixels and fitting the most probable edge. We thus need to detect (one or more) texture edge pixels that are input to the cue integration engine. Note that this is done locally, i.e., we have no a priori knowledge of how the texture on either side of the edge looks like. Hence, we are applying a windowing technique as sketched in Figure 2.3.

Figure 2.3(a) shows the layout of two (slightly adapted) adjacent SFMs for a very small matrix size (4 × 3). The "centre pixels" are marked with an ellipse. Such a matrix pair is sliding horizontally at different vertical positions ("strips") in the ROI (Figure 2.3(b) sketches this for three strips). Figure 2.3(c) finally depicts the detected edge pixels that are input to the cue integration technique as hypothetically belonging to the edge.

The calculation of the distance between the matrices adapted to the layout shown in Figure 2.3 is then

$$D_{sf} = \|\mathbf{M}_{left} - \mathbf{M}_{right}\| = \sqrt{\sum_{i,j} |\mathbf{M}_{left}(i,j) - \mathbf{M}_{right}(cols - i, j)|^2} \tag{2.1}$$

with $cols$ being the number of columns and $\mathbf{M}_{left,right}$ being matrices of contrast, covariance or dissimilarity, following the definitions in Section 2.3. The idea is that the difference in contrast, covariance and dissimilarity between the matrices get maximal with the assumed edge being between them.

Additionally, another feature defined in [WC92] can be exploited: a (very coarse) regularity measure (`reg`). Therefore we separately compute for each matrix the feature

$$f_{reg} = \sum_{(i,j)} DSS(i,j) \tag{2.2}$$

and compute the joint regularity measure:

$$F_{reg} = exp(-\sigma) \tag{2.3}$$

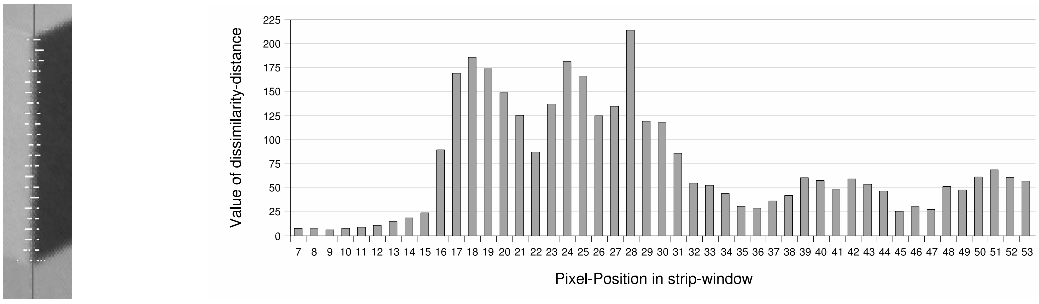
with σ being the standard deviation of the matrices. The idea here is that regularity of the area spanning both matrices is minimal when the searched edge is in the middle of the two single matrices – in this case, the regularity feature (Equation 2.2) reflects the maximum regularity for each separate matrix, resulting in a high variance of regularity between the two matrices (Equation 2.3).

To summarise, we have four different statistical measures to compute the probable location of an edge pixel. It is obvious that this method of detecting the correct edge pixel gets better with increasing matrix size. Due to timing constraints, however, the matrices are not supposed to get too large. One advantage of using cue integration as explained in Section 2.2 is that more than one pixel may vote for being a correct edge element. Consequently, we allow *local* maxima instead of only the *global* maximum to be input to the edge fitting process.

Especially in cluttered scenes this is a necessary means in order not to lose the advantages of this technique by still using only small matrices.

2.5 Results

We wanted to investigate the best application of using statistical feature matrices for edge detection as described. To this end, the three channels of both the RGB- and the YT_1T_2-colour model have been tested on the colour cube example with little texture information. The rationale is that the responsiveness of the texture edge detection should be maximised – both for simple colour scenes and for cluttered ones. The latter model



(a) Edge pixel candidates.

(b) Measure of dissimilarity vs. x-position.

Figure 2.4: Results of Figure 2.2: (a) Edge pixel candidates with fitted edge in grey; (b) a typical distribution graph for one of the strips.

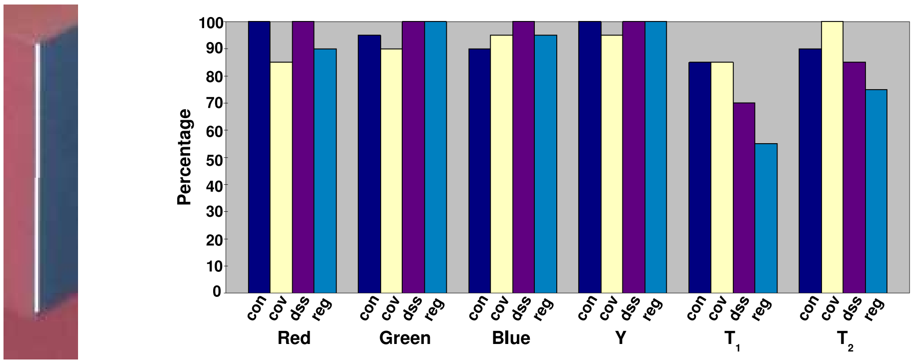


Figure 2.5: Percentages of found edge pixel candidates for the edge. In R,G,B and Y channels, dissimilarity and regularity measures work best. For Y_1 and Y_2, performance is generally worse, but covariance dominates.

(YT_1T_2) separates the luminance (Y) channel from the chrominance channels (T_1 and T_2) and is used by *V4R* as it is superior to HSI and other non-linear models [Kub99]. All four statistical measures that have been described are computed in each of the channels. Note that this approach differs from the usual application of texture algorithms as they are typically only applied to grey-level images or only to the luminance channels of colour images. Indeed, when using the YT_1T_2 model, our investigations showed that the chrominance channels are significantly less suited for texture edge detection than the luminance channel. However, using the RGB model, which does not have a dedicated luminance channel, significant distinctions could be shown.

In order to be able to evaluate the performance in the different channels, the deviation of detected edge pixels from their manually observed correct position is computed. An edge pixel being within two pixels from the true position is considered as correctly found.

Figure 2.4 shows results: Figure 2.4(a) depicts the resulting first five local maxima for the edge of Figure 2.2. In Figure 2.4(b) the distribution of the matrix differences using the dissimilarity measure for the intensity channel in one of the strips is shown.

For the same simple example boundary of the cube, the resulting percentages of found edge pixels are shown in Figure 2.5. Note that the percentages reflect how many

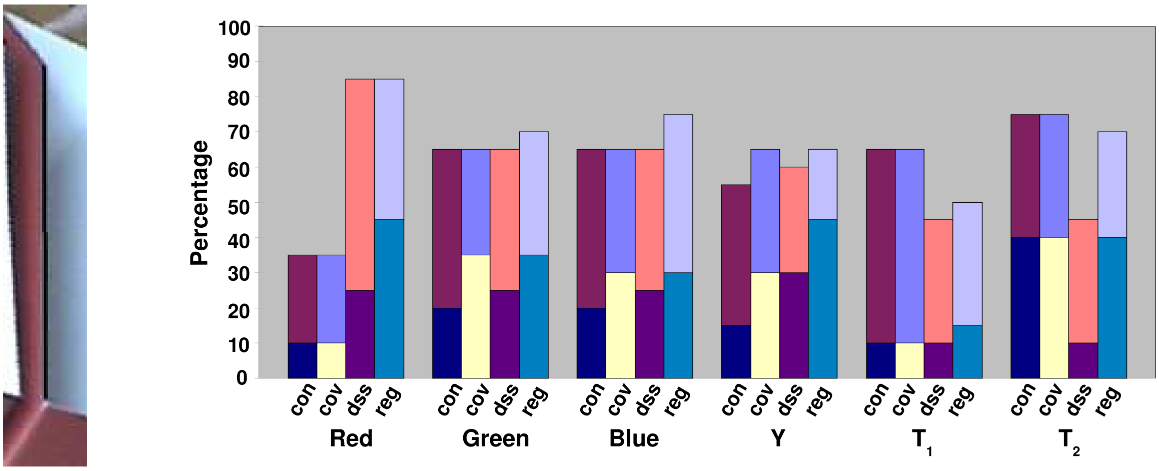


Figure 2.6: Detecting an edge with local clutter: the lower bars show first *two* maxima, the upper bars the *next three* maxima; dissimilarity and regularity are best in the first four channels, contrast and covariance dominate the chrominance channels.

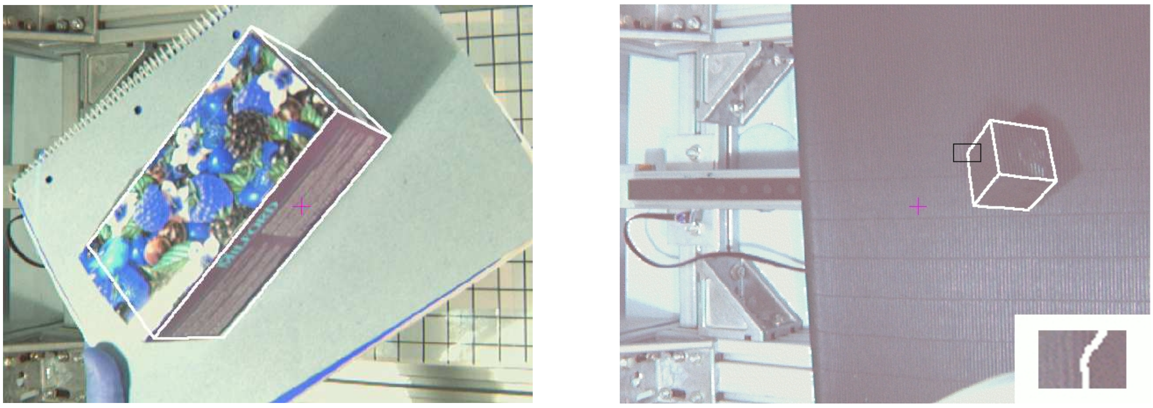
of the loci of *the first two maxima* of the corresponding statistical feature hit the correct edge (i.e., here, only two edge pixel candidates are provided by each strip and each statistical feature). As can be seen, for this simple case texture edge detection works very well although the difference in texture of the surfaces is not easily observable.

Another example is shown in Figure 2.6. Now, both the results for *the first two* and the gain when allowing *the first five* maxima are shown.

Summarised, these results clearly vote for the dissimilarity measure to be used along with the luminance channel (the YT_1T_2 model is already used by *V4R*). Additionally, it is essential to work with relative instead of absolute maxima and to handle local clutter by *relative* thresholding to choose a reasonable number of edge pixel candidates: only extrema that are at least one third of the highest peak of the distribution (cp. Figure 2.4) are valid extrema. With these found settings, complex tracking situations with textured objects could be improved. Figure 2.7 shows two example sequences where the usage of colour and intensity alone failed whereas the usage of texture could significantly delay loss of tracking. Especially with the cube on the right, only textural computations can locate the boundary of the object which is even hard to identify for humans.

2.6 Discussion

Textured objects provide rich appearance cues for various visual tasks. In this chapter, we presented how texture could be used for finding the borders of objects. The reason was the provision of an additional cue to an existing cue integrating approach to model-based rigid object tracking. We used a matrix-approach, describing two adjacent local texture patterns. Combined with a sliding window approach, the maximum difference between the two is searched and a possible edge is assumed at this position. Precondition for this approach is, however, an already existing approximate estimation of the possible edge location.



(a) Tracking a tea box; the pose (model reprojected in white) is not lost yet not recovered perfectly (left end).

(b) Black object on black (yet differently textured) background – the zoomed region highlights the difficulty in noting the differences.

Figure 2.7: Two examples where intensity alone fails; only texture prolongs tracking.

A drawback of this approach is that in principle the responsiveness and consequently the quality of the edgel estimation is dependent on the ratio between the size of the used matrices and the size of the texture (meaning the size of the repeated texture subpatterns). For example, imagine an object with a checker grid texture whose grid size is either very small, so that the texture description of the matrix is perfectly reflecting this special layout, or is larger than the matrix size in which case the description would be quite poor. Furthermore, this also implies that the distance of the camera to the object is crucial. In our case, the size of the matrices has predominantly been chosen for performance reasons, in order not to affect tracking at frame rate. The pragmatic reason was that texture shall only serve as an additional cue, in the best case avoiding loss of tracking, in the worst contributing nothing. To summarise, this approach is really only meaningful as an *additional cue* as has been done in this project, but not as a standalone tracking solution.

Concerning the overall topic of this thesis, namely where the differences between a cognitive approach to perception for mobile robots and low-level computer vision work lies, we can for now state that the only “cognitive flavour” of such a research project lies in the paradigm of using texture to detect the grouping of objects (in this case: where the object under inspection ends). We will need to think later about what “object” actually means to the system (namely in this case: nothing). In the next chapter, we will also use texture; however, we will abstain from searching for edges but rather use the surface features of objects to be handled. Furthermore, we will integrate this information into a broader framework that also includes learning and detecting the object. This also has its own specific disadvantages, of course, especially concerning the objects that can be handled. However, we will see that it is a bit more “cognitive” with respect to the fact that the system acquires its own representation.

> Appearances [are] to us in four ways: for either things appear as they are; or they are not, and do not even appear to be; or they are, and do not appear to be; or they are not, and yet appear to be. Further, in all these cases to form a right judgement is the office of an educated man.
>
> *Discourses, Chapter 27*
> Epictetus
> [Epi87]

Chapter 3

Interest Point Tracking

A second typical work of "low-level computer vision" that we did and would like to present here, is the deployment of a monocular object tracker that uses appearance information directly derived from the scene [SBV06, BSV06, SBV07]. It is part of an integrated system that has been developed as joint work with Georg Biegelbauer.

We have indicated in the last chapter that tracking rigid objects is an important capability for many computer vision systems, especially in the robotics domain for visual servoing. Tracking approaches might consider only two-dimensional surface patches or comprise a full 3D-model of the target object. The system presented in this chapter is of the latter kind with having the goal in mind of being able to track simple geometric shapes through image sequences in cluttered scenes, where a precise pose of the object should be obtained. Insofar, this approach is similar to the one presented in the last chapter.

Now, though, we want to deploy a system that is additionally able to retrieve the shape of the object under inspection without parameter tuning and furthermore to detect it at the beginning of the tracking sequence in the scene. This way, we were able to minimise the steps needed for manual interaction. In this respect, such a system can – from a conceptual point of view – be seen as a bit more "cognitive" as the example in the previous chapter, where the object model had to be provided manually. Now, the system builds its own model of the object at runtime. It is a sort of learning, however, without the need for exemplars and for defined lighting conditions in each step.

Furthermore, the whole system combines two different and complementary kinds of visual information, namely shape and appearance. The first two steps (the learning of the model and the detection of the learnt object in the scene) are done exclusively with shape information; the subsequent tracking part only uses appearance (i.e., the textured surface). We will only present the part that is concerned with the monocular tracking and will only very shortly describe the learning and detection steps. Details about the latter two (including the respective State of the Art) can be found in our colleague's dissertation [Bie06].

Section 3.1 gives a review of related work, then the overall system design is presented (Section 3.2). Afterwards, Section 3.3 describes the tracking method developed in detail before experimental results show the feasibility of the approach (Section 3.4). Again, a discussion (Section 3.5) completes this chapter.

3.1 State of the Art

So-called “interest points” have become prevalent in recent years, not least because the rise of computational power allows for their fast computation. Recently they enjoy great popularity in object recognition and stereo matching tasks. In general, they are better called “interest regions” as they mostly include information about a local neighbourhood rather than a single pixel. Usually, detection of specific features in the image is separated from the description of their local neighbourhood.

The early reference work of a detector of interesting image features is the “Harris Corner Detector” [HS88] which is based on the idea that a significant change in all directions (using an auto-correlation matrix and the computation of its eigenvalues) indicates an interesting image region. Other examples are Laplacian of Gaussians (LoG) or Difference of Gaussians (DoG), both using a convolution of the image by a Gaussian kernel with subsequent application of either the Laplacian operator or its approximation, the difference between two Gaussian images. A quite different approach is used when detecting “maximally stable extremal regions” (MSER) [MCUP02], where stable *regions* (not points) are found by applying thresholds on different levels and subsequently search for regions that stay stable over a large range of thresholds. A comprehensive summary of well-known detectors can be found in [MS05].

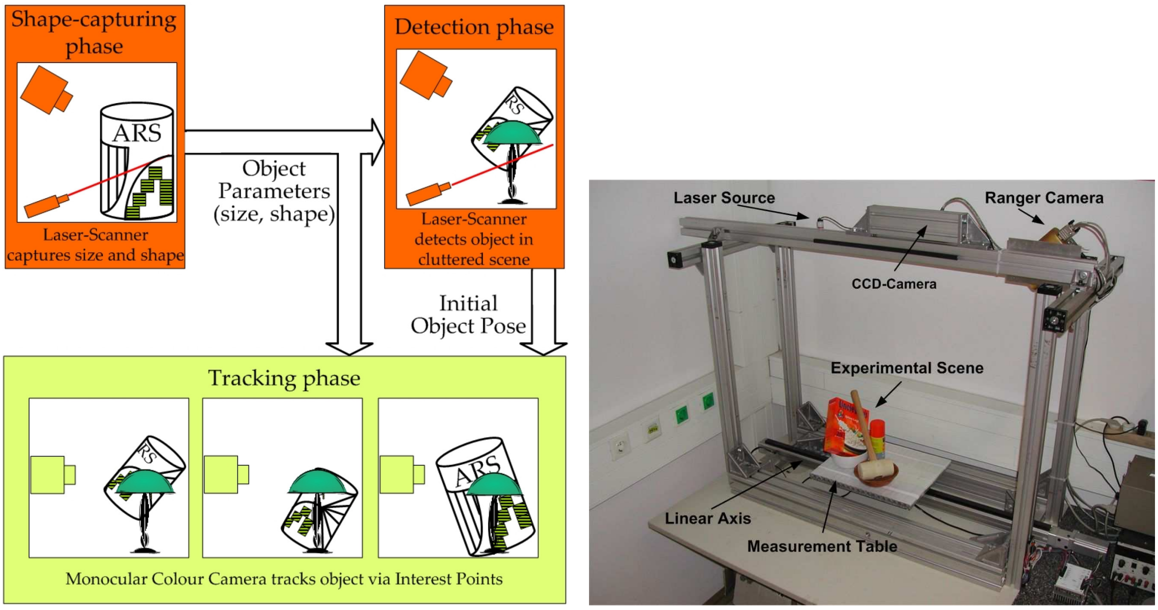
After having detected interesting points in the image, a so-called “descriptor” is applied that tries to capture information about the local neighbourhood of the point. There are some important properties that such a descriptor should have: invariance to image scale and rotation, robustness to affine distortions, changes in viewpoint, noise and illumination. Most widely used is the “Scale Invariant Feature Transform” (SIFT) by [Low04]. SIFT is a multi-dimensional feature vector that results from computing the gradient magnitude and orientation in a region around the (previously by DoG) detected keypoint location with subsequent Gaussian weighing and accumulation into orientation histograms. Usually, a 128-dimension vector is generated by using a 4x4 array of histograms with 8 orientation bins each. One of the various examples for extensions to SIFT is “Gradient location-orientation histogram” (GLOH) [MS05], which aims at increasing robustness and distinctiveness by using a log-polar location grid with three bins in radial direction. The resulting 272 bin histograms are dimension-reduced by applying principal component analysis (PCA). “Speeded Up Robust Features” (SURF) [BTG06] are descriptors that especially target at higher speeds by using integral images for image convolutions. A comprehensive overview and evaluation of main descriptors give [MTS+06].

Object Tracking has a long history in computer vision and we will confine ourselves here to model-based approaches that use predefined object features without the goal of doing self-localisation of the camera. An extensive overview of tracking in computer vision gives [LF05]. Similar to the system presented in this chapter, is the work of [KS03] who also do monocular Superquadric tracking. Their approach is to extract edges in the image, then fit the projection of the Superquadric model onto the image and minimise a cost function. A paper by [YKBK05] shows a combination of a laser scanner and a camera for tracking complex objects (e.g., a toy lorry). However, the selection of line features needs to be done manually from the range image. A system similar to the setup proposed here has been presented by [TK04] and consists of a laser range scanner and stereo cameras. Detection of geometrically primitive objects (boxes, bowls, and cylinders) is performed without previous learning and requires a previous scene segmentation using surface curvatures. Regarding interest point tracking, most approaches rather use a fast detection algorithm in every scene than a truly time-dependent tracking algorithm, e.g., [PLF05]. An exception to this rule is the work of [KK05] who combine line features and interest points in an IEKF (Iterated Extended Kalman Filter)-framework for tracking rigid objects. In [KK06], the authors use a CAD model and corner features for dealing with cylindrical and conical objects.

Concerning the general idea of combining different vision modalities, [KC02] illustrates the importance of fusing shape and appearance for robotic servoing and grasping tasks, because the robustness of model-based techniques for tracking line features of highly textured objects is lacking. Their solution to overcome these problems lies in the usage of training images with subsequent projection into eigenspace, which only slightly reduces the sensitivity to illumination conditions in open environments. Information about the objects of interest for the different tasks is currently often provided off-line by manually storing them in a database. For example, [HYMLJ05] use complete solid model representations in an object database for all objects that can be manipulated. The SIFT features that describe the objects and graspability or accessibility information are stored therein as well. In [KKK03], off-line computed Zernike moments around interest points are stored in a database for later being matched via a probabilistic voting during the on-line process. Additionally, recognition is verified by aligning model features to the input scene.

3.2 Overall System Design

Before going into details regarding the chosen tracking approach, we present the overall conception. Figure 3.1(a) shows the system and its three phases. The two red boxes on top depict shape-capturing and detection of the captured object in the scene, both done with the laser scanner. Shape-capturing is done without clutter, the detection then already takes place in the everyday scene in which occlusion and different lighting conditions are probable. Whereas the first step delivers the parameters (size and shape) of the object used subsequently, the second step is liable for the initial pose for the monocular tracker. All information is used later on, sketched in the green lower box. For the actual tracking the object, only interest points are used; for the computation of the pose in each frame, however, the other information is needed, as will be explained in detail in the next section. In Figure 3.1(b) our experimental setup is shown. Scanning the object with the laser is in this case done by moving the linear axis; on a real robot


Shape-capturing phase
ARS
Laser-Scanner captures size and shape
Object Parameters (size, shape)
Detection phase
RS
Laser-Scanner detects object in cluttered scene
Initial Object Pose
Tracking phase
RS
ARS
Monocular Colour Camera tracks object via Interest Points
Laser Source
Ranger Camera
CCD-Camera
Experimental Scene
Linear Axis
Measurement Table

(a) Shape-Capturing and Detection is done with the laser, tracking with the colour camera.

(b) Our experimental setup.

Figure 3.1: Conceptual design and sensor setup of our integrated system.

a pan-tilt-unit would be used.

The rationale to use these three steps and two different vision sensor modalities can be summarised as follows: In a home robotic scenario, it is wishful to be able to show an object once to the robot, which builds up its own internal representation (a Superquadric in the system proposed) which can be used later on to detect and track the object in different scenes with different lighting conditions. The learning of the object's size and shape is done with the laser, because this could take place under totally different conditions than the handling of the object later on. The appearance information used by the tracking is not retrieved before tracking is really necessary, i.e., in the actual cluttered scene. Another advantage to include such a shape-capturing step is the abandonment of usually used large databases to store hand-coded information of surface features of specific instances of objects. For example, retrieving a cereal box with the same size but different layout is possible in our case, but is not if a database with stored appearance information is used[1]. Finally, such a system accommodates the primary goal of autonomous robotics: no need for user interaction between those steps.

As representation we use Superquadrics, a family of parametric shapes first introduced by [Bar81], because they allow for a compact and straightforward description of primitive geometric shapes that can be easily passed along the different phases. The recovery of Superquadrics is a well-studied problem and global deformations can be accounted for, too [SB90]. At first glance it might seem to be too restricted to use such simple geometric primitives, however, a lot of everyday objects can be described by Superquadrics (such as cereal boxes or cans) and in principle, an extension of our system to objects that are composed of several Superquadrics, is straightforward.

The surface of a Superquadric is the spherical product of two parameterised quadric

[1] Note, however, that with the chosen Superquadric approach, the size must be identical in this case.

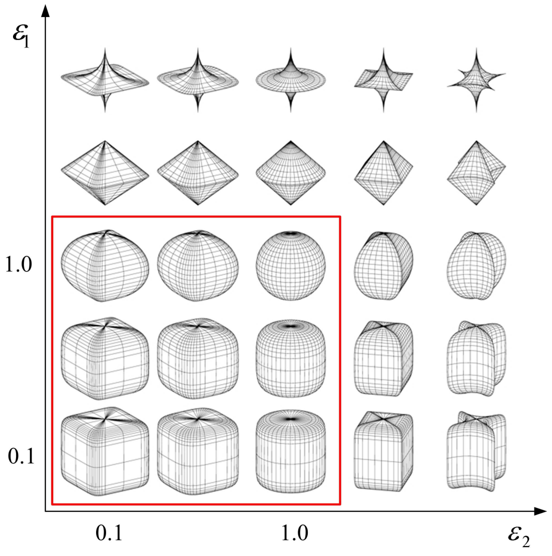


Figure 3.2: Overview of basic Superquadric shapes; the subset used in our system is marked.

```
DoG-detect IPs and store in *newpoints*
**for all** points in *newpoints* **do**
    Compute SIFT-description
    Search for SIFT-match in *oldpoints*
    **if** match found **then**
        Associate 2D coords with 3D model coords of
          matched keypoint
        Add both information to list of *posepoints*
    **end if**
**end for**
**for** $i = 0$ to *iterationnr* **do**
    RANSAC: Take 4 to 8 points from *posepoints*
    Compute new object pose
    Use remaining points to vote for this pose
**end for**
Choose best pose
Empty *oldpoints*, *posepoints*
**for all** points in *newpoints* **do**
    Compute 3D model coords with chosen pose
    Add SIFT-description and 3D model coords to
      *oldpoints*
**end for**
```

Figure 3.3: Tracking loop algorithm.

curves. Superquadrics consist of five parameters. Three (a_1 to a_3) contain the size in x-, y- and z-direction and the last two, named ϵ_1 and ϵ_2, describe the shape, which can be seen in Figure 3.2. The explicit formula of a Superquadric is given by

$$\boldsymbol{x}(\eta, \omega) = \begin{bmatrix} a_1 \cdot cos^{\epsilon_1}(\eta) \cdot cos^{\epsilon_2}(\omega) \\ a_2 \cdot cos^{\epsilon_1}(\eta) \cdot sin^{\epsilon_2}(\omega) \\ a_3 \cdot sin^{\epsilon_1}(\eta) \end{bmatrix}, \quad \begin{matrix} -\frac{\pi}{2} \leq \eta \leq \frac{\pi}{2} \\ -\pi \leq \omega < \pi. \end{matrix} \tag{3.1}$$

Substituting η and ω leads to the implicit formulation given by

$$F(x, y, z) = \left(\left(\frac{x}{a_1} \right)^{\frac{2}{\epsilon_2}} + \left(\frac{y}{a_2} \right)^{\frac{2}{\epsilon_2}} \right)^{\frac{\epsilon_2}{\epsilon_1}} + \left(\frac{z}{a_3} \right)^{\frac{2}{\epsilon_1}}, \tag{3.2}$$

known as Inside-Outside-Function [ibid.] used for Superquadric recovery.

3.3 Tracking Method

We will now focus on the tracking part; the other two steps are explained in more detail in [SBV07]. From the learning step, the system knows the parameters of the Superquadric from which a computation of the line features is in principle possible. However, edge tracking is prone to fail in case of textured objects due to the high number of spurious edgels (cp. [VSGA05]). Therefore we decided to rather exploit the textured surface that usually can be found on everyday objects. As interest points to be detected and described, we chose the combination recommended by [Low04] – Difference of Gaussians and Scale Invariant Feature Transform – because of having high repeatability and accuracy.

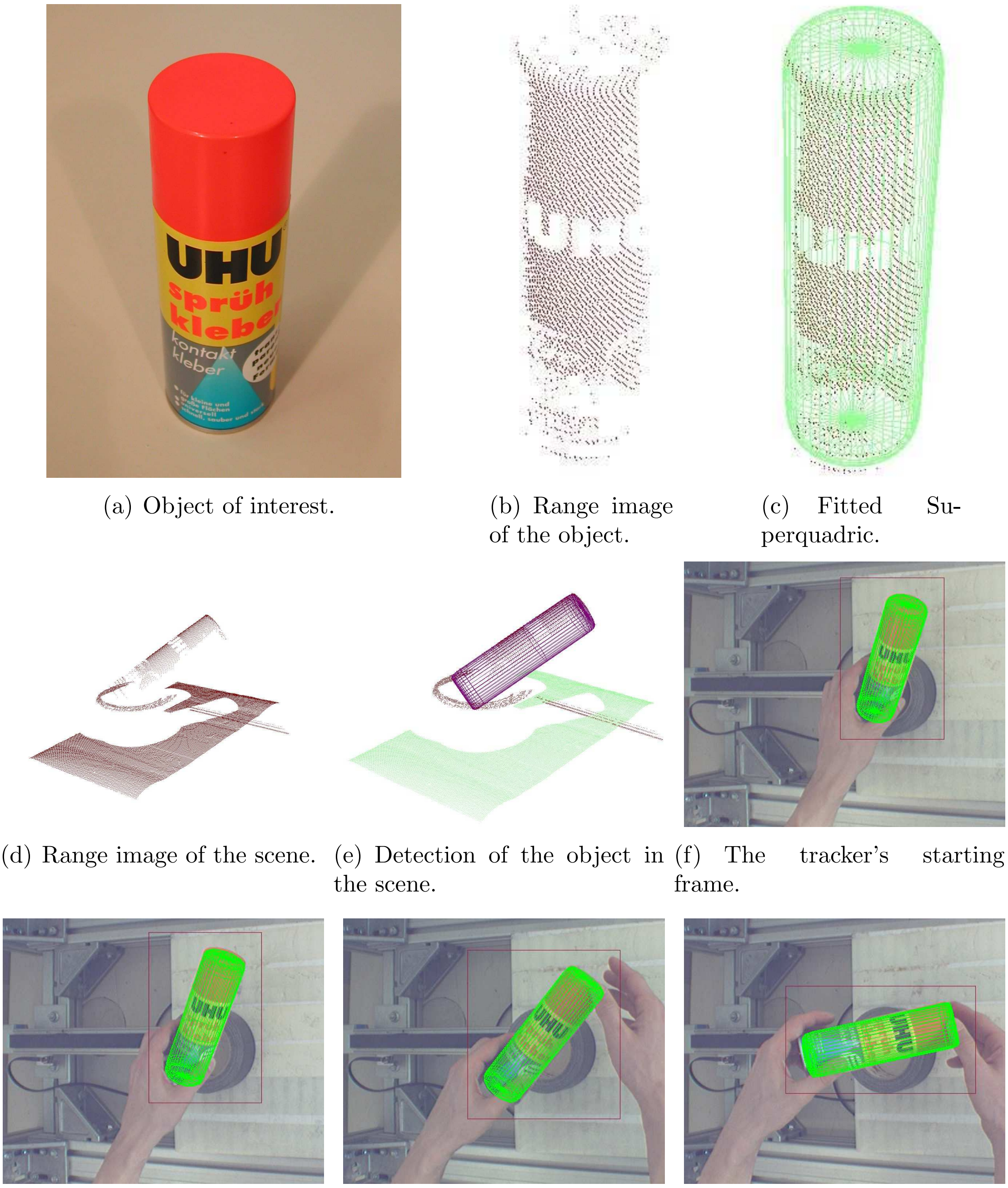
As input for the tracking, the object's parameters (from the learning step) and – for the first frame – the initial pose (from the detection phase) are used. Of course, a calibrated system is needed, using a reference coordinate system (see [SBV07]). The Superquadric is projected into the camera frame, and its convex hull is computed, resulting in a region of interest in the image. Within this region of interest, interest points are detected and described by SIFT. Note that the interest points are located not before now – the actual scene, which totally cancels out the need for same lighting conditions as in the previous learning step. Given the projected Superquadric, the computation of the 3D-location (in model-coordinates) of interest points in the image plane is straight forward. Here, we can see an additional advantage of the use of Superquadrics: Their description is in fact a closed surface, easing the computation of 3D model coordinates by intersecting the ray of sight of the located interest point in the image plane with the surface of the Superquadric. This step delivers a list of 2D image plane points that are associated with 3D model coordinate points.

The tracking loop is summarised in Figure 3.3 and works as follows: In each frame, interest points are searched. In doing so, the tracking assumption is followed, meaning that the object is assumed to only slightly move. Consequently, points are only searched in a limited neighbourhood from their previous positions, which increases robustness. The descriptions of found points are compared with those from the previous step, using the matching technique suggested by [Low04]: Only those closest neighbours are judged to be good ones and are accepted the distance of which to the second-closest neighbour is significantly larger. As the 3D model coordinates of the points have been computed in the previous step, the newly found keypoints can now be associated with model coordinates, finally resulting in a list of 2D points that are associated with 3D points. This information is input to the pose estimation algorithm following [LHM00]. To overcome problems with possible outliers (wrong associations or points not belonging to the object), a few RANSAC-based (Random Sample Consensus) [FB81] subsets are used for pose computation. To deal with pose ambiguity difficulties an extension to the algorithm following [SP06] is applied. All points that have not been selected for the current estimation cycle are used as votes for the pose quality. Finally, the best voted pose is used or, in case of multiple equal votes, the mean of them is taken. The tracking loop ends with a reprojection of the points onto the Superquadric in order to compute the 3D model coordinates for the next step. Using such an approach that does not rely on points found in the first but rather only on those found in the previous frame, newly appearing points can be seamlessly integrated, allowing for rotations and occlusions throughout the tracking phase.

An alternative to this kind of tracking would be to use the learnt Superquadric and re-detect it using the laser scanner in each frame. This, however, is usually much slower than a monocular tracking approach.

3.4 Experiments

Figure 3.4 shows a rather simple scene for tracking a cylindrical object. The object shown in Figure 3.4(a) is scanned (3.4(b)) and a Superquadric is fitted (3.4(c)).



(a) Object of interest.

(b) Range image of the object.

(c) Fitted Superquadric.

(d) Range image of the scene.

(e) Detection of the object in the scene.

(f) The tracker's starting frame.

(g) Tracking Fr. #7.

(h) Tracking Fr. #11.

(i) Tracking Fr. #16.

Figure 3.4: Handling a cylinder in an uncluttered scene with hardly any occlusion: (a) to (c): capturing the shape of the object in an empty scene; (d) to (e): detection of the object in the actual scene where tracking will take place, the output is the starting pose for the tracker, depicted in (f); (g) to (i): some frames of the tracking step. (The reprojected Superquadric is depicted as green mesh-grid.)

The laser scanner is used a second time for detection of the object in the scene (Figure 3.4(d)), locating the learnt Superquadric (3.4(e)). This provides the starting pose (3.4(f)) for the subsequent tracking phase, some frames of which are shown in Figures 3.4(g)–3.4(i) (the model is superimposed in green).

Figure 3.5 shows a second, more complex example. Again, we chose an everyday commodity item as object to be retrieved and tracked, this time a rectangular rice box. The learning step is somehow deficient on the shortest side of the object: the superimposed model (white lines) is a bit too "thick". However, we can see that tracking is not affected, because of the mentioned separation of visual modes during detection and tracking: The interest points are derived from the actual scene, i.e., from actually visible surfaces of the object. An edge-detector would be misdirected, as it would try to locate an edge in the image where only background is visible.

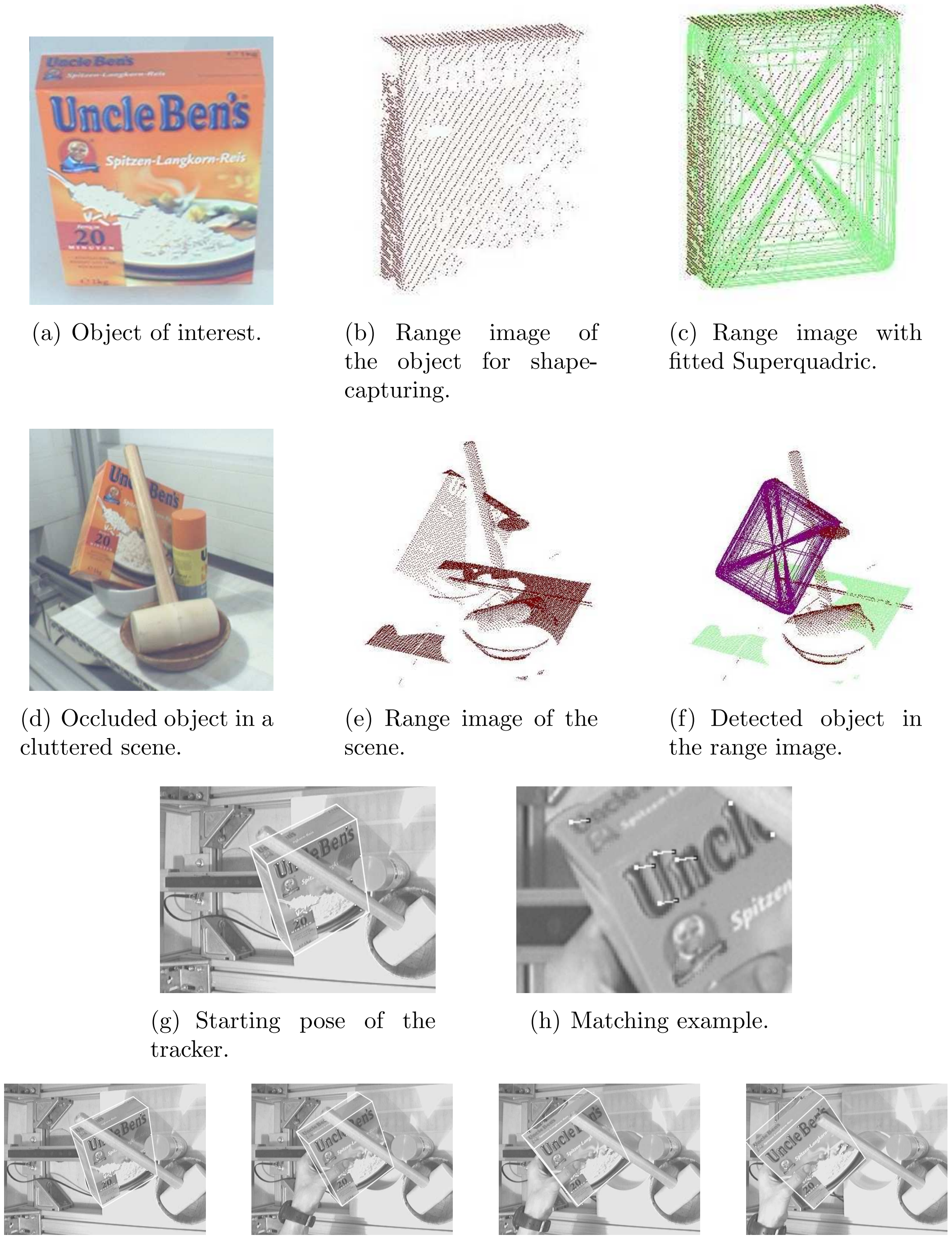
Note that during detection (second row of Figure 3.5), the rice box now lies in an arbitrary position and is partially occluded by the white bowl, the tin can as well as the hammer shaft. The white points are the locations of the interest points. The matching example on the right of the third row is a zoomed clip of frame #18. The black dots indicate the positions where interest points have been found in the previous step, the white dots the locations of the points in the current frame. Note that there are some white points that have no match with black ones (no white chain). Nevertheless, these points are stored for the next iteration as they may possibly be matched with points of frame #19.

Furthermore the occlusion caused by the hammer shaft is dynamic during tracking due to the motion of the rice box. Additionally, the hand coming from the left also occludes a part of the box. Finally, even the number of visible faces of the box changes. Nevertheless, the pose is recovered with sufficient accuracy.

3.5 Discussion

This chapter showed a tool with which objects of basic geometrical shapes (more concretely: shapes that can be described by the parameters of the Superquadric model) can be learnt, detected and handled using different vision modalities. The main reason was to overcome usual problems that occur when appearances of objects are learnt and later tried to be retrieved. Although interest points (such as SIFT used in this project) are quite robust against illumination and viewpoint changes, with our approach, even objects could be handled the parameters of which have been learnt in a flipped position. This means, using only appearance information in every of the steps shown, would rely on the fact to have had learnt all visible surfaces of the objects. We were able to circumvent this problem. The second contribution was the closing of the chain learning – detecting – tracking due to the exploitation of a calibrated setup and the use of a common object representation in a compact format.

This work was conducted mainly as a showcase on how a laser scanner and a colour camera can be combined; before the setup could be used on a mobile robot, work on speeding up the algorithms would be needed (a typical task to shift from "robotics as science" to "robotics as engineering"). Additionally, a nice future work would be



(a) Object of interest.

(b) Range image of the object for shape-capturing.

(c) Range image with fitted Superquadric.

(d) Occluded object in a cluttered scene.

(e) Range image of the scene.

(f) Detected object in the range image.

(g) Starting pose of the tracker.

(h) Matching example.

(i) Fr. #1.

(j) Fr. #18.

(k) Fr. #23.

(l) Fr. #28.

Figure 3.5: Handling a rectangular box in a cluttered scene with static and dynamic occlusions: (a) to (c): object in empty scene for learning its parameters; (d) to (f): detection in the actual scene; (f) and (g): the starting pose seen from the laser scanner and from the colour camera respectively; (h): matching example: black dots from frame #17 are matched with white dots from frame #18; (i) to (l): some frames of the tracking step (the reprojected object is this time depicted as white lines).

the extension of the possibly manipulable objects by using models that could connect more Superquadrics together, e.g., the mallet that can be seen in Figure 3.5 could be modelled as being composed of two cylindrical objects.

Finally, the tracking approach using only interest points and 3D model coordinate-information from the *directly preceding* step is not optimal. Although the advantage is the handling of full rotations (due to the integration of newly appearing points and the rejection of disappeared ones), drift occurs, meaning that the pose-error accumulates the longer the tracking sequence lasts. In this case, the additional use of, e.g., edges in order to "snap back" to a better pose estimation would be helpful – similar as to what [KK05] does.

As in the previous chapter, we can now reflect on a meta-level on how much "cognitive" this approach is. First, we can assert that the system now does not rely on a pre-coded model to be handled but rather is able to construct its own representation. This model is, however, also a bit constrained, namely to a range of possible shapes, but at least it is constructed on its own. Secondly, the system is more "autonomous" as it builds up its knowledge that is needed at hand, and then automatically passes this information to the other parts. Concerning the question what "object" actually means, we again have to be honest and admit that it is rather a kind of proto-object (or even more sobering: a set of shape parameters) rather than any kind of semantically laden object notion that is used by the system.

Part II

Bridging to Cognitive Functions

Das gnoseologische [erkenntnistheoretische] Problem involviert das ontologische und ist ohne dasselbe nicht zu behandeln. Aber es fällt mit ihm nicht zusammen. Deswegen muß die grundlegende Betrachtung des Gegenstandes derjenigen der Erkenntnis desselben vorausgehen.

Grundzüge einer Metaphysik der Erkenntnis
NICOLAI HARTMANN
[Har65]

Chapter 4

Theoretical Issues for Situated Vision

In the Introduction we have tackled some of the crucial questions involved when thinking about the perceptual capabilities of a robot companion (i.e., an autonomous cognitive robot working at the human's side in a home environment). In this chapter, we will present a structured account to these theoretical investigations and will focus on concepts and a rich ontology in order to show their importance – not only for the complete agent, but especially for its vision capabilities. Therefore, we want to tackle these issues on a broad basis and address the need and type of concepts for an autonomous home robot, because we believe that what [Vie06] states for concept *formation* holds just as well for the *usage* and *handling* of concepts: "When addressing concept formation in AI, what can be called the 'system level' is often overlooked, which means that concepts and categories are rarely studied from the point of view of a system, autonomous and complete, that might need such constructs and can acquire them only by means of interactions with its environment, under the constraints of its cognitive architecture."

After giving a motivation of the theoretical considerations of this chapter by reviewing related work (Section 4.1), we will explain why computer vision research should be radically re-integrated into work on artificial intelligence – establishing the link between the two parts of this thesis (Section 4.2). Afterwards we need to clarify the explicit and implicit assumptions that our approach takes (Section 4.3) and some of the notions implied (Section 4.4). The reason is that they are all prone to have different connotations depending on the field of research one is involved with. The core of this chapter focuses on cognitive functions that we define to be important to keep an eye on (Section 4.5) and on how to glue them together by using an ontology (Section 4.6). Section 4.7 addresses the issues not explicitly tackled so far and establishes the link to the subsequent practical chapter. A short summary of the topics discussed concludes this theoretical chapter (Section 4.8).

4.1 A Motivational Review of Related Work

In the following we will review related work of theoretical considerations about the topics discussed in this chapter. As it is a theoretical one, in principle all philosophical, developmental-psychological and even computer science traditions could be consulted as related work, which would lead to endless enumerations; so we will rather focus here on quite general remarks with respect to cognitive robotics and give the more concrete related sources of the chosen issues in the respective sections.

[SWH^{+}06] list long term requirements of cognitive robotics that serve as good overview of the complexity of the field. We will cite here some of the points that are relevant for our focus on *vision* for robot companions (which we classified as subset of cognitive robots in Chapter 1). The first citation very nicely states the main issues[1]:

> In the 1978 paper of Barrow and Tenenbaum and other work at that time, some important ideas about the perception of spatial structure and motion were beginning to be explored, that were later abandoned mostly in favour of work on recognition, tracking, and localisation, with little or no work on understanding of 3-D spatial structure. There has been a great deal of progress on specialised image processing tasks, driven by collections of benchmarks which have little or nothing to do with the ability to act in the environment, for instance benchmarks concerned with recognition of faces, or types of animals or vehicles, but without perceiving any spatial structure.

Further on, they come to the conclusion that "[...] as far as we can tell after extensive enquiry there are no AI vision systems that can perceive surface structure in such a way as to produce an understanding of the implications for actions." Finally they even give a reason for why the recent boom of learning agents by relating sensor and motor data in a machine learning framework is ill-posed: "Moreover, since the variety of types of surfaces and 3-D orientations of graspable, touchable, pushable, pullable surface fragments is astronomical any attempt to learn about such affordances by storing sensorimotor correlations will founder on a combinatorial explosion."

Later on, the authors describe that current vision approaches do not take the relation of objects to other objects and their parts ("multi-strand relationships") sufficiently into account. Concerning processes, these relationships change in parallel, meaning that several concurrent processes need to be perceived. "Some of the changes will be metrical (changing distances, orientations, shapes) others topological (changing between touching and being separated, between being inside and being outside, etc.)." Besides a lot of further requirements, one important statement is made about the cognitive ability of imagining. Humans can think about something and visualise processes that did not actually took place. In all that, they take constraints into account, such as gravity, structural properties and the like.

From this paper we can already see the multitude of both the representational content needed in a cognitive system and – on a meta-level – the necessary disciplines

[1] The cited paper herein is [BT78].

involved in investigating those capabilities. A lot of human-like competences are addressed and it is this spirit of abstract thinking that will guide us in the following.

Our focus is on perception, where the broad motivation for including top-down knowledge can be directly seen when trying to understand attention. Strong evidence for top-down guidance for attention comes from the work of J. Wolfe [WCF89, Wol94] where it is argued that search in humans is efficient due to the "guidance" of the feature selection top-down, letting "relevant" features pop out naturally as the corresponding output channels get activated stronger. Wolfe's approach builds on the well-known work of A. Treisman's Feature Integration Theory [TG80]. Finally, [AH97b] shows a related model of how humans learn selectivities along the visual pathway and show that improvements start at higher generalizing levels. Their Reverse Hierarchy Theory [HA02] even argues that "[...] explicit visual perception follows the visual hierarchy in reverse direction, from top to bottom [...]".

In psychology, it is well known that perception is in fact tightly bound to non-sensory information, such as one's own body. For example, the slant of hills appear steeper to people that are fatigued, are carrying heavy backpacks or have low physical fitness [PBGM95]. The same holds for the horizontal: Distance is perceived longer according to one's physical circumstances (carrying something heavy) and can even be influenced by previous visual-motor adaptation that reduces the anticipated optical flow [PSBE03]. The latter is related to the effect of (in this case: cross-modal) "priming", where current cognition can be strongly influenced by (shortly) previous happened cognitive events. For the case of purely visual priming, [AC06] presents a study where subjects first need to categorise pictorial stimuli into "emotionally positive" and "emotionally negative" ones. In the actual testing phase, those images are shown randomised and subjects must decide to which group they belong. However, before the actual stimulus, a different (priming) image is shown very shortly. Reaction time is significantly shorter in cases where the target image category is congruent with the priming image category. To put it in a nutshell, those examples show that a pure bottom-up approach of vision is wrong – at least for humans. And we would argue that it is ill-posed for autonomous robots as well.

On an even more theoretical level, we could cite here quite a lot of the history of philosophy as related work. Especially epistemological questions always touch (even though often not explicitly) questions regarding perception as well. An analysis of some philosophical positions of the twentieth century with regard to epistemology, perception and the question of resulting knowledge in humans can be found in [Sch08]. We will abstain from reviewing the philosophical approaches here and postpone the discussion of some of them to the respective chapters – especially when it comes to the notion of ontology and its connection to epistemology (Section 4.3), we will come back to some important philosophical considerations – and then we will also unveil the mystery why we put a quotation by Nicolai Hartmann on top of this chapter.

4.2 Problems Instead of Solutions

As [Slo06] makes clear, one problem that the investigation of truly autonomous intelligent agents faces is the fragmentation of AI into subfields during the last few decades. This seems to be partly due to the failure of highly ambitious goals and promises that have not been held. The title of this section refers to the invitation the author makes at the end of the cited paper: "The requirements-based roadmap [presented in this paper] requires far more work by far more people. They [i.e., researchers in AI] need to agree on problems, not solutions. The long term payoff could be very high." The need for laying the focus on problems (again) instead of on engineered solutions for very narrow and particular subgoals, is what especially holds for the subfield computer vision.

Part I dealt with typical examples of computer vision, where you have a clear problem statement and a neat solution. E.g., given a wire-frame model of a rigid object, a suitable camera is chosen, the lighting is checked and finally an algorithm is developed that can deal with exactly such a model in this and (very) similar situations – and nothing else. Mostly, the scenes in which the experiments take place are carefully chosen as well, including that clutter is reduced to the amount that the system can handle[2]. One example would be what we showed in Chapter 2: The lines that the system tries to find in order to fit the model must not be discontinuous – or not more than a certain threshold[3]. This implies that there is inherently an infinite number of situations where this vision system would just fail.

We could (and did) argue, that consequently we need to implement yet more vision sensors and combine them – as shown in Chapter 3. And see: This actually worked – until a certain point (again). Though we can now handle cluttered scenes and dynamic occlusions, we still have problems that cannot be handled with vision alone: One simple example would be that there are two cylinders of the same size in the scene, but one being a part of a column of the room, and only the other being the vase that the robot should bring: Without context information and a thoroughly embedded understanding of the scene and possible scenarios, we would fall back to a trial-and-error phase (at best).

So where is the deep underlying problem? Why can we not just stick the algorithms from Part I on top of a robot and it works?

In our opinion, all this is due to the fact that the model is totally separated from the actual higher-level tasks and knowledge of the complete system – because these tasks or knowledge simply do not exist. There is no notion of "object" in the system, no understanding what the system actually deals with. This points to one of the most burning issues: the notion of "object" itself. Starting from a bottom up approach, as is the case for perceptual grouping or in general when detecting pre-defined (model-based) features (such as lines or interest points), the gap that opens up to a semantically grounded understanding of the object that is handled is evident.

Summarised, this means that instead of tackling highly specialised issues, such as

[2] This explains why a huge amount of papers dealing with object recognition show pictures with the object under inspection lying on a perfectly white background and non-adjacent to other objects in the scene.

[3] See [Zil07] for an excellent argument against the usage of thresholds in computer vision.

finding the solution of a chess game or the recognition of faces, we need to first see the overall conceptual need – and in case of cognitive robotics (which shall serve as basic foundation for a robot companion), this conception includes the task that needs to be fulfilled, predictions that build on previously perceived objects and processes, higher-order generalisations of things happening that are similar yet not identical, the understanding what perceived concepts mean and what their meaning in a given situation could be, and lots more. There is one way how this ambitious goal of thinking higher-level with lower-level processing together should *not* be done – but what is unfortunately often the case: In recent years, the buzzword "cognitive vision" has become inflationary used and very often, this led to the approach to start from low-level vision processing algorithms and sticking *then* some more or less further "cognitive processing" on top. Or even worse, the (technical) algorithms were just explained as if they were cognitive. Those low-level processing used the same output, rationale and benchmarking that pure computer vision demanded, be it tracking, detection, segmentation or classification – just as in the approaches presented in Part I. The opposite way – and this is where the theory that this chapter tackles comes into play – is to *first* analyse the needs or possible approaches and mainstays of cognitive robotics using borrowings from diverse disciplines including psychology or philosophy and *then* to search for suitable tools to provide input (and output) for them.

We said that the enhancement of the "pure low-level vision layer" (by adding sensors and optimising algorithms) is obviously not enough. This is related to another hot topic that is usually discussed in computer vision for robotics: the drawbacks and advantages of specific vision sensors. Again, this makes sense for specific engineering applications (e.g., for a precise 3D model to be scanned, the choice will probably rather be in favour of a laser scanner than of stereo vision), but for their appliance in cognitive robotics much more important is the proper embedding of vision information into those cognitive functions that the *whole system* is in need of. Of course, different sensors have different properties and, for instance, you would not use a red laser light when dealing with blue and black objects. Clearly, a minimal situation-dependent quality is required. Still, if the information is not embedded in the system, even the best sensor would only deliver shallow data. The same holds for the algorithms. A lot of vision algorithms are totally fine, yet we are still focusing on special solutions and optimisations instead of first analysing the deep problems and implications for the whole robot and what the minimal set of cognitive functionality looks like. After that, we can tune the parts to fulfil the requirements. In short: Focus on problems, not solutions. Note that even in terms of the seemingly perfect human sensor system this is true: On the sensory level alone there are drawbacks and errors happening (just think about the huge amount of optical illusions where this becomes apparent) – yet humans are usually able to make sense of the world in any everyday situation and under any circumstances. One reason is that there are only objects in the world that do appear in some context – already their *where* and *when* of occurrence provides rich information.

Talking about the notion of object, of course, one could argue that a perfect geometric description of, say, a bicycle, can be equalised with the notion of the "object bicycle". But it can not. Though, of course, usual vision algorithms are in need of some kind of representation, again, their whereabouts are unclear – and it stays on this shallow level as long as the representation is not connected to higher order processing. A huge database will not solve this principled issue. Even if a bicycle detector system is provided with any bicycle that ever existed and that ever will exist (you can already see obvious problematic implications here), the deeper grounding of the concept bicycle stays problematic. As a simple example, imagine a bicycle-sign on the roadside. Humans immediately know that they cannot ride on it, that they cannot move it and that they even cannot really touch it as it is only a sign. The concept bicycle is invoked, but parallel to that reasoning is happening that takes information into account not only about the height and distance where the "bicycle" has been detected, but also what its occurrence means in this situation (in the context of being depicted on a sign), etc. We could say that this bicycle-sign is somewhat a fusion of the concept bicycle with the concept sign. We might use the same *word* for it as for a "real" bicycle and say: "Look at this funny blue bicycle", but this must not be confused that it is in principle a totally different concept we are referring to.

This is related to what Marr pointed already in 1982 [Mar82] and what we have shortly mentioned in the Introduction. According to Marr, there are four stages of visual perception. Starting from the retinal image, there is an image-based stage (where local image processing takes place, such as edge detection), then the surface-based stage (where intrinsic properties of visible surfaces are discovered), leading first to the object-based stage (where true 3D information is available, including hidden surfaces of the object) and finally to the category-based stage that includes affordances and further information connected with knowledge about this category (see [Pal99] for details). This sounds very much bottom up – and this is actually an important source of problems (besides the fact that Marr's view implies the existence of an objective reality – which in the light of constructivist, embodied and situated movements needs to be substituted by a subjective, task- and context-dependent view): Computer vision did not manage to reach the category-based stage most likely because of the lack of interactions from other information sources (e.g., by giving massive feedback from non-visual knowledge). Even generic object detection systems are using inherently shallow information about the object's affordances, possibilities and impossibilities, varying vs. constitutive properties, and so on. We need an interdisciplinary approach where perception must radically be seen as part of cognition. Not before integrating different information, we will be able to reach this stage and to design vision systems *for* cognitive robotics. This includes thoughts about the representation format used, which will be an important topic towards the end of this chapter.

4.2.1 Relativisation – A Return to Reality

Well, this all sounds very negative. Is computer vision senseless? Does the community not know where to go? No. All this critique does not aim at condemning the work done in computer vision as it is absolutely necessary and, indeed, astonishing results have been achieved. The only critique is that it is a discipline that developed on a path that more and more lost its connection to robotics and artificial intelligence. For the latter disciplines and particularly for cognitive robotics, it is only a *necessary precondition* yet not a *sufficient step* towards a true "understanding" of an image or a scene – an important precondition for autonomy. This is especially critical as a lot of what is heard of "cognitive vision" takes the crucially wrong attempt to put cognition on top of vision – instead of trying the find a minimal set of cognitive functions in which vision needs to be *embedded*. We could even make a very keen and probably controversial statement which puts these considerations in a nutshell: *Vision alone, i.e., dissociated from an agent (be it the human or the robot for the case of robotic vision) does not make sense; vision can only be seen as part of cognition.* This then allows for what we stressed in the Introduction: Vision is an inherently multi-dimensional process that includes a lot of information that low-level visual processing alone cannot provide. What is needed is the conceptual view on what vision can provide to cognitive robotics – taking borrowings from other disciplines.

However, we will later on argue that the use of psychology should *not* give the impression, that building up a blueprint of the human is the far goal. It rather helps us to possibly detect necessary functionalities. Furthermore, we can *not* promise to give an exhaustive cognitive theory of an autonomous agent in this chapter – that would be too much to be demanded. Even in psychology (or cognitive science) there is more than one "cognitive theory". Besides the "big theories" such as those of Bruner (on learning), Piaget (on development) and Vygotsky (on socio-cultural issues), there are "cognitive theories on the perception of movies" [Ohl89], "cognitive theories on the development of fear" [Jac81] and so on. The hunt for a clear comprehensive cognitive explanation of humans (in which some of these theories will take a role, some will not) is still open – and one that can by synthesised is even more up in the air. It is even the question whether such a comprehensive theory would make sense at all as it is unclear whether we can press all of human cognitive abilities into one coherent scheme. Anyway, our goal here is rather to give a view on those cognitive functions which we judge to be important for understanding visual processing and – here we hit the road pertaining to robot companion research – furthermore which we think are important for robots that serve the human on his/her side, i.e., who's ecological niche is well-defined (but nevertheless very broad). These "functional parts" are quite tightly bound to vision but should serve as possible building blocks in a later more extensive theory connecting to the higher-level functions. This investigation therefore additionally – and here we get back to the title of the section – tackles problems, not solutions, also on an implementation-independent layer. For too long, the paradigmatic question on whether symbolic or subsymbolic processing, on whether a more situated and embodied or rather a dynamical systems approach is the right way to go, has been discussed. All of the approaches have their application areas, all of them can outperform the other in some respect – what really counts is a functional layer and this has to be decided with respect to the actual application area[4].

Returning to the claim that vision is multi-dimensional, we agree with W. Burger[5] on the point that it is possible that we completely underestimate the complexity of computer vision. Furthermore, in this context, it is now clear that his second noteworthy statement that the algorithms of computer vision are totally fine and we just need to

[4] We will, however, later on argue why a symbolic approach is indispensable at *some* point, see Section 4.5.6.

[5] Talk at the 3^{rd} NFN "Cognitive Vision" workshop in Graz, Austria on May 9^{th}, 2008.

better use them, is not paradoxical to the first one, but rather points to exactly what we described above: Computer vision is more complex, because it is bound to processes that have nothing to do with vision at first sight. A pure bottom-up approach can never reach the level of richness in information that we usually associate with “vision”.

4.2.2 Roadmap

So what does the remainder of this chapter try to provide? Summarising the critique shown above, we will try to bring some issues together, to extract some cognitive functions that are “near enough” to the visual input yet “high enough” to allow a connection to further higher-level processing, such as task planning, scene “understanding” and the like. Note that the distinction between “high” and “low” already implies some hierarchical ordering that should only be carefully assumed. We are also aware that this is not a truly and radically interdisciplinary approach (which would include an interdisciplinary team to work on), but it shall trigger new ideas from artificial intelligence, robotics, cognitive science, psychology and philosophy.

Therefore we need to clarify the explicit and implicit assumptions that are taken and due to having different disciplines involved, there is need for a clarification of the notions in our context, i.e., the far goal of a truly autonomous home robot companion. Thinking about the notion “cognitive” alone is not enough, because we will soon get into the discussion of related terms like “intelligence” or even “consciousness”. This is of course not new to artificial intelligence or cognitive science and are issues that are also discussed in these disciplines, yet we are still mainly interested in the perceptual domain – in the sense that perception is one of the foundational features that should enable the robot to find its way around, to reason about what is and what is not in the environment and to finally extend its “ontology of the world” through perceptual input. This return to the perceptual domain will be staying central throughout these theoretical considerations.

After mentioning some of the functions that we regard as important, we will tackle the issue on how to glue them together, so that they do not keep being shallow cognitive terms that are yet again taken in order to have a better sounding title for a publication. We will argue for the use of an ontology or more ontologies. Implementation issues are postponed until the next chapter, where we will provide a novel contribution on an applied level.

As a foretaste to this practically oriented chapter, in which we will still be interested in the consolidation of the buzzwords “cognitive vision”, “situated robotics” and the like, we can already reveal that we will focus on the step from “proto-objects” (i.e., grouped features without semantic information, such as a cube or a cylinder) to the object level and the subsequent link to semantic information. Another important task that will be tackled is the attempt to analyse what kind of ontology is needed for an agent to handle objects, behaviours, processes, etc. The hard task will be to link theoretical and practical findings of computer vision to the higher-level “cognitive functions” that we propagandise here.

4.3 Body of Theories – Presuppositions and Mindset

Reading philosophical literature or works on cognitive science is often difficult if the main line of thought that the author takes is not stated explicitly. Clarifying the mainstays of one's own presuppositions and philosophical tradition is therefore a necessary step before digging into details. This section shall fulfil exactly this function – and shall help to shed light on our "view of robotkind".

In short, our basic attitudes can be condensed into the following statements. We will first give an overview how they relate to each other and go into details subsequently. We identify...

- ...that (tempting) *anthropomorphic paths* should be avoided as much as possible,
- ...that the philosophical tradition of *rationalism* is the most suited for our purpose,
- ...that *ontology and epistemology* cannot be treated separately, and
- ...that the *functional layer* is the right one to start with.

4.3.1 Overview and Connections

Our objective is a robot companion as defined in the Introduction – differentiating the whole approach from "robotics as industry". This means that we are not interested in a specific (usually clearly defined) goal to reach such as the visual inspection of parts on an assembly line but rather in more general issues like autonomous behaviour or "intelligent action". In our case, we focus on perception (also a cognitive activity), i.e., what vision needs to be like in order to provide the necessary information to enable such behaviour. Hence, as outlined in the Introduction, the following presuppositions are building on the meta-assumption that cognitive robotics (aiming at letting robots achieve complex goals by giving them capabilities that are inspired by humans or animals) is necessarily the foundational basis for a robot companion. We have additionally argued before that interdisciplinarity is the modus operandi of choice to tackle such complex issues. However, including psychology and philosophy into such work needs a clarification of our position towards anthropomorphic temptations. Consequently, this is the first issue we will tackle here explicitly.

The second point will delineate the broad philosophical framework and point out what constructivism and rationalism can tell us. This will help us to tackle questions regarding different views on and different knowledge about the world. Related to that is our third point, the question on the interrelation between ontology and epistemology – guided by this rationalist and constructivist approach. The bottom line is that we actually *cannot* distinguish between ontology and epistemology when working on robot companion research. This will also provide us with the foundations to discuss what to pre-provide and what to learn, what to be nature, what nurture and guide us to the conclusion of the central theme in this thesis: the functional approach. The discussion of the fourth and final point will not be given here but is postponed to Section 4.5.1 which tackles why we focus on a functional layer and what this means and implies for work on vision (and perception in general).

4.3.2 On Anthropomorphism

As could already be seen so far, a mixture of anthropomorphic terms has been introduced to engineering. We will detail *our* understanding of some of them in Section 4.4.

In general, it is imperative to be cautious where to draw the analogies from robots to humans. Home robots should not replace humans – this would imply ethical considerations that cannot be dismissed or answered easily. Furthermore, it is arguable why this should be done anyway. It might be interesting to study certain parts of development in order to draw *analogies* to humans, but a full-fledged copy should not be wishful – especially for robot companions we thus state that we are not working on a butler-robot that is human-like but that is able to fulfil certain tasks *for* the human. It shall be a *companion* in the best sense, not a copy of its operator. Here we are again: Tasks. We argue that in order to retrieve coffee or to locate the lost car keys, you need not be a conscious human – but, and that is the twist, you might need to have some "cognitive functionality" (like prediction or generalisation) that can be human-inspired.

This leads us to another important point: To be inspired by the functionalities humans apply does *not* refer to the implementation layer. It could be seen that it was very tempting in the last years and decades to use artificial neural networks because they are "biologically plausible". Whereas this may be interesting in cognitive science, where interaction between neuronal clusters or subparts are studied, one must not ignore that robotics is different: We want to get the task done, whatever works is fine. We also do not construct cars with legs because that would be "biologically plausible" – wheels just work fine and do get us from here to there. For the field of robot companion research, the goal is to construct an artefact that is able to serve purposefully next to a human, so the controller used does not need to be biologically plausible. This, again, leads us close to a *functionalist* approach.

One example is the function of emotions. Judged to be highly relevant for the equilibrium of the outside and the inside world of the human, we may want to implement the same for an artificial system. Thinking functionally, we may learn about the *purpose* of emotions, but *implement* something different to achieve the same functionality without necessarily having to "rebuild emotions" (e.g., by some chemical process) as such.

One argument against the view just presented is that it might be impossible to get full-fledged "intelligence" if we exclude biologically plausible phenomena due to a functional approach – the typical examples being consciousness or qualia (i.e., the subjective *quality* of a feeling; for a classical paper see [Nag74]). Monistic connectionists, for example, believe that these phenomena can *only* arise from emergence out of a specific arrangement of neurons. Although we do adopt monism in general, we discard this fixation on neurons and rather take a functionalist position again. This implies that questions concerning the need for consciousness can be postponed until and only if we find the necessary functions that consciousness provides. In any case, the "human implementation layer" (i.e., neuronal networks) is *not necessarily* the wishful blueprint for cognitive robotics.

At this point it seems important to point to the difference to developmental robotics. This highly interesting field is concerned with studying how cognition might be bootstrapped from the beginning, how further development takes place, and how this can lead to a state in which behaviours can be performed. In home robotics having a robot companion in mind, however, it is much more justified to pre-give a lot of the knowledge needed – quasi to let nature take a more dominant role than nurture. In principle here, too, we could argue that any information needed by the robot should be acquired by itself. However, from a practical point of view, namely the vision that at some point we are able to produce and use large numbers of companions out-of-the-box, it is highly unlikely that the user will have enough patience to let the robot learn and develop over several years (just like a child), eventually making mistakes and starting anew or with guidance of its "parents".

Of course, we do need mechanisms that allow for adaption to the *specific* operational area, yet the *general* invariant information parts should be predefined. This tackles one of the main difficulties of robot companion research: even with the predefinition of information, the system nevertheless needs to be flexible and adaptable – leading to the tricky part on how to pre-give knowledge that is not fixed to certain instances, but rather general enough to serve a whole domain. This is reflected in this thesis insofar as we are tackling abstraction (in conjunction with the notion of object concepts), yet not learning (generalising from instances). This does not mean that it is irrelevant to cognitive robotics! We are only rather focusing on how to get vision done even if specific concepts are assumed to be available. Still, inspirations can be given by developmental robotics, as there definitely is an overlap in the issues tackled.

Arguing against anthropomorphism could lead to the question why we use the "human sciences" at all when dealing with artefacts. The answer is simple: The use of these disciplines gives us the great advantage that we can exploit studies and insights from decades with the additional benefit that they provide us with viewpoints which cannot be taken by engineering alone (e.g., the subjective and analytic views of psychology or neuro-psychoanalysis). As we have pointed out earlier, pure engineering approaches are perfectly suited for finding a solution to a specific narrow problem – however, they seem ill-posed when it comes to a general level. To start from humans has a reason that is simple to say yet hard to implement: Humans accomplish a variety of tasks in a flexible, adaptive and ever-changing manner. We could, however, also start from animals, but that would again miss the chance to use insights and thought-provoking impulses from psychology and most parts of philosophy.

4.3.3 Philosophical Tradition

Broadly spoken, we follow a rationalist approach, influenced by constructivist ideas. Computer science (being a "synthesising science") is inherently rationalistic, even though connectionists might disagree at this point. Yet, one must not forget that the design of artificial networks is itself imposing a specific *structure* that on its part influences how stimuli are propagated. Of course, the network weights adapt themselves and are in this way constructing a specific interpretation of the data. Nevertheless, strictly spoken the system does not learn purely empirically (especially the "training set" must be chosen wisely). We do not want to dig too deep into the discussion with respect to the different paradigms that have dominated cognitive and computer science over the last decades and assume that the reader is sufficiently familiar with this topic. We need, however, mention that we believe a strict rationalist approach will necessarily entail the use of symbolic processing for robotic computer vision *at some point*. The reason is that symbols (and their logical computation) are necessary for specific capabilities – especially when it comes to explicit considerations such as imagining possible outcomes of actions, planning for processes taking place in the not too near future and similar higher-level "thoughts". All of these need *explicit representations* that can be worked with.

Back to the rationalist approach, a concise and for our aims suited account of rationalism as opposed to empiricism is given by R. Case [Cas99, p.38]:

> In reaction to British empiricists, philosophers such as Kant [...] suggested that knowledge is acquired by a process in which order is imposed by the human mind on the data that the senses provide, not merely detected in these data. Examples of concepts that played this foundational role in Kant's system were space, time, causality, and number. Without some preexisting concept in each of these categories, Kant argued that it would be impossible to make any sense of the data of sensory experience: to see events as taking place in space, for example, as unfolding through time, or as exerting a causal influence on each other. For this reason he believed that these categories must exist in some a priori form rather than being induced from experience.

Needless to say, this approach poses some very crucial questions that have kept developmental psychologists and developmental roboticists alike busy for many years, in particular: What are those concepts that need to be provided in order to bootstrap a successful learning process? In our (perceptual) domain, the crucial question herein is: Which concepts do we need to predetermine for the purpose of *ordering* the perceptual stimuli?

One could argue that Quine's objection – "[...] all inculcation of meaning of words must rest ultimately on sensory evidence" [Qui69] – is an argument against pre-giving anything, as words are concepts of a language, and language is learnt empirically. However, the crucial notion herein is the inculcation of *meaning*, so we would counter that we only need to be sure that everything we (as designers) pre-give the system needs to be clearly defined in terms of what it *means* to the system (meaning is never objective but always subjectively bound to the experiencing agent) – or to put it in a different notion: to *ground* the symbols [Har90], another grand and controversial problem[6]. This is the hard part of robot companion research – not so much the discussion on whether to predetermine more or less (this is very important in developmental robotics), but rather *how* to predetermine the concepts in order to link it to meaning in the system. These concepts will constitute the ontology, will need to include an explicit form of meaning, and will consequently be much more than sensorimotor correlations.

[6] Note that we are, however, not of Harnad's opinion that symbol grounding can only be done bottom-up. We rather adhere to the argument of [Slo02] that explains the possibility to construct meaning top-down starting from abstract forms und refining later when confronted with additional evidence. We will elaborate more on this later.

How does constructivism now fit into this picture? One of the mainstays of radical constructivism – as proposed particularly by the works of H. Maturana and F. Varela – is that we never work with an objective reality but rather with our subjectively perceived reality. To put it in Ernst von Glasersfeld's words [vG92]:

> Here there is a direct conflict with a tenet of the traditional scientific dogma, namely the belief that scientific descriptions and explanations should, and indeed can, approximate the structure of an objective reality, a reality supposed to exist as such, irrespective of any observer. [...] Objectivity is a subject's delusion that observing can be done without him. Invoking objectivity is abrogating responsibility, hence its popularity.

[Zie01] lists some of the basic principles of radical constructivism: knowledge is not passively received but actively built up, the tendency of the system is guided towards viability and – probably most important – cognition serves the subject's organisation of the experiential world, not the discovery of an ontological reality[7]. In its radicalism, constructivism introduced notions like "operational closure" (an autopoeitic system is closed and there is no information transfer to and from it), "self-reference" (the system *only* refers to itself during the process of self-preservation) and structural determinism (meaning that only those changes can be made that are determined by organisation and structure of the system), but also milder terms such as the "structural coupling" with the surrounding media (which allows for an embedding and sort-of communication of the system with a system of higher order) [Sch87, Mat87]. For us, the constructivist approach primarily points to the crucial fact that every autonomous system is only dealing with its own representation of reality. This implies that whatever information we deal with, be it hard-coded or built up by the system, must be suited to and embedded in the system itself. However, in the next Section we will see that for the synthesising sciences and here, especially in computer science dealing with autonomous systems, the border between the dismissed "ontological reality" and the agent's own epistemology must be assumed to be non-existent. Concerning conceptual development, an important ancestor of constructivism is Jean Piaget, whose theory on "assimilation" and "accommodation" provides an explanation on how conceptual changes are performed by combining new experiences with cognitive structures that already exist in the subject[8].

If we decide to pre-give some concepts to the system (e.g., of how objects look like) we must be very careful to bind them to the overall system's ontology. With that, the system should then be enabled to use these concepts for further knowledge derivation. In connection with perception, these concepts are built up using sensory stimulation; [Pal99, p.10] uses the term "model" and explains why constructing a model is not leading to a complete arbitrariness: "[...] *the observer is constructing a model of what environmental situation might have produced the observed pattern of sensory*

[7] [Sch87] talks here of "epistemological solipsism" as opposed to the (misunderstood) criticism of "ontological solipsism".

[8] Herein, assimilation refers to the reaction of situations by using existing knowledge (schemas); accommodation on the other hand means the alteration of present knowledge and information.

stimulation. [...] Invoking the concept of models does *not* imply that perception is 'pure fiction.' If it were, it would not fulfil the evolutionary demand for accurate information about the environment." Later on [ibid., p.12 f], he also explains (with the example of a concrete model built from sensation) why building models makes sense: "A perceptual model of the three-dimensional environment does not need to be modified much as we move around because the only thing that changes is our viewpoint relative to a largely stable landscape of objects and surfaces. [...] Without a perceptual model that somehow transcends momentary stimulus information, vision would not be able to guide our actions appropriately." As can be seen, the use of such a model is mainly motivated by evolutionary demand.

We can now formulate our combination of rationalist and constructivist viewpoints: With constructivism we share that all concepts must be grounded in the system and must bear the information needed to construct a view of reality feasible for "survival" and further knowledge acquisition. They must be bound to the experiential world of the agent. Rationalism, in turn, is not necessarily conflicting with that when it says that there are some innate concepts that help ordering the perceptual data. For a robot companion, this means that we can provide the data that the agent needs to know in order to work autonomously (this is opposed to the pure spirit of constructivism), but in a manner that is subjectively useful for the agent (and this is opposed to the more objective conceptual view of rationalism).

A first attempt to think along these lines can already be found in Aristotle's works, as [Rus99, p.257] points out:

> He [Aristotle] did not believe that *epagoge* [i.e., induction] could lead to true knowledge; all it could achieve was the guiding of the developing mind toward the kind of experiences that could further its development. [...] First, far from believing that developed knowledge (*episteme*) is something gained entirely by induction from sensory experience, Aristotle viewed it as a set of necessary truths, arguing that scientific knowledge, and knowledge more generally, is concerned with that which is necessarily the case, and with *epagoge* playing a guiding rather than foundational role in the individual's coming to this view. The truths of science (I shall take that to include the folk science that the child acquires) are not mere listings of phenomena but rather express a system of laws such that if one accepts certain founding principles further truths can be *deduced.*

4.3.4 The Inseparability of Ontology and Epistemology in Robotics

In the following, we use the words "ontology" and "epistemology" mainly in the philosophical senses, where "ontology" refers to the concepts of the objective truths ("reality" or "ontology of the world", i.e., the totality of *what there is* [Hof04]) and "epistemology", on the other hand, is the theory about the process of acquiring these concepts, the cognition and perception of experiences, which is hence strictly subjectively built up. This results in a personal world-view of the agent, for which we will also use the term ontology, but we will call it the "personal ontology" for reasons of clarity.

Palmer makes clear that "Vision is [...] a *heuristic process* in which inferences are made about the most likely environmental condition that could have produced a given image" [Pal99, p.23]. Is this in contradiction to the rationalist approach delineated in the previous section? No. It seems that there are (more or less) "stable" concepts which not only get fed (and constructed) by vision, but which themselves then feed what vision sees, i.e., the heuristics are *guided* by the concepts. This leads to a very crucial point, namely the question of the interplay between an "objective" ontology of the world and the "subjective", constructed, acquired concepts (personal ontology) of the agent: "In any event, the belief is widely held that philosophical truths are what Kant called 'analytic,' meaning true a priori and in virtue of the meaning of the terms employed, in contrast to the synthetic truths of psychology that tell us something about the world" [Rus99, p.264]. This means that these differences between psychological and philosophical views can often be pinned down to the following: Ontology is fundamentally different to epistemology. Whereas this view is of great relevance to the *human sciences*, we would argue for a different perspective for *robotics*, namely that epistemological truths of the artificial cognitive system are essentially the same as or at least very tightly bound to the ontology of the human designer – this is an advantage or the crux of the (only possible) third-person-design-perspective that we do (and have to!) encounter in robotics. This fact is often neglected but will strongly influence our understanding of the specific nature of those concepts and has deep implications on *what can be acquired* by the system *at all*. In order to motivate our view, we will start with a look at developmental psychology, more precisely at an approach that coincides with such a convergence of ontology and epistemology and then review a relevant philosophical position, namely Nicolai Hartmann's epistemology.

In developmental psychology, [Cas99, p.50] reports that

> [...] contemporary theorists in each school now agree (1) that the notion of a systemwide cognitive structure should be replaced by a notion of structures that are more domain specific, (2) that children's cognitive structures should not be modeled as systems of logical operations, but as systems for making meaning, each with its own distinctive conceptual and/or symbolic content [...]

This twist to the search for *meaning* is important and underlines what we have said about the intentionality of tasks and the subjectively constructed and tuned concepts. Furthermore, it is now obvious that there is a specific preformation of the human sensors due to evolutionary demand as well as the "mode" on how to acquire concepts [Kei99, p.169]:

> There are often several different ways of usefully organizing a group of things. A group of animals can be seen as pets and nonpets, as edible and inedible, as predators and prey, and so on. My claim here is not that there is one right way to pick out informational patterns. Rather, the claim is

> that the patterns are not merely arbitrary, especially in those cases where causal interactions cause certain property clusters to be stable.

In this quotation we see the implicit adoption of Kant's view that causality is an a priori concept which we use to *order* perceptual data. The role of *logic* (be it explicit or implicit), we would further argue, is exactly in using those a priori concepts for these purposes and to establish the links between them. Furthermore, it implies that there is an important interrelation on how humans acquire concepts (epistemology) and the nature of the concepts (ontology). This is underlined, when the author writes further on [ibid.]: "While some effects along those lines [of the situated cognition movement] certainly do occur, the emphasis here is on the simple idea that one cannot remain completely agnostic with respect to the structure of the world and hope to have an adequate account of cognitive development." Finally, it is obvious that existing "stable clusters" influence further information acquisition, which is made more explicit by [Ham90][9]: "Experience can make us see that certain things are so. We may not be able to see them in that way unless we have the concepts which are presupposed in so seeing them."

We have just seen that despite the compelling arguments of constructivism, we have to readmit external reality (the "ontology of the world") into the conception of the personal ontology in order to come to an intuitive grasp on how development in human works. For computer science, we furthermore *only* have the external reality – or rather: our (human's) account of what the external reality looks like – at hand when designing knowledge bases for artificial agents. We must therefore combine epistemological and ontological questions and discuss how the ontology of the agent coincides with or diverges from the objective – or, again, more precisely: the designer's – ontology. The point we want to make here is that constructivism still is a concept we need to think about, although we need to lessen its epistemological solipsism. Furthermore, this allows us to accept the pre-definition of concepts, i.e., to have a set of ontological entities as representations in the agent which are *not* constructed by learning of the agent itself.

A final quotation by Keil before we step onto philosophical grounds shows that also developmental psychologists see the necessity of dampening constructivist approaches [Kei99, p.166]: "All of these contrasts also highlight the need to take a realist stance to the world around us. In essence, cognitive development can no more study the acquisition of knowledge by merely looking at the machinery in our head than visual neurophysiology can study the retina by merely looking at retinal anatomy and not considering the nature of light." He makes clear that this kind of *realist stance* affects the conception of all areas of cognition [ibid.]:

> Almost 50 years ago James and Eleanor Gibson started on programs of research showing that, for perception, one needed to characterize the structure of what is perceived and how those informational patterns might be detected and used in an organism's ecology. It is now evident that similar analyses must be undertaken not just for what is perceived, but also for what is thought about.

[9] Cited after [Rus99].

Looking at the history of philosophy, one name pops out that is directly in line with such an understanding of the interplay of epistemology and ontology: Nicolai Hartmann. His attempt – named "critical realism" – was not only to purge the Kantian position from metaphysical touches (especially regarding the notion of the "thing-in-itself" which we will encounter later on), but – as a critic of phenomenology – also to generally reunite epistemology and metaphysics. His conviction was that cognition is nothing that is creating objects in the human mind, but rather a capturing of something that does exist independently and no matter whether it is perceived or not. This was also Kant's opinion, however, Hartmann more radically tried to strike a balance between realism and idealism. This is in line with constructivist approaches, where the traditional dichotomy of realism and idealism is broken up by neither focusing only on the subject (idealism) nor only on the object (realism), but instead on the *relation* between subject and object (see, e.g., [VTR93]). Hartmann, as well, emphasised on subject and object being elements of an ontological interrelation that has *no* causal character (because the "image of the object" within the perceiving subject is not a spatio-temporal object, which would be the prerequisite for being in a causal relation, cp. [Har65, p.324]).

The quotation we put on top of this chapter is from the very beginning of one of Hartmann's main works, the title of which already shows his wish to reunite personal knowledge and ontological reality: "Grundzüge einer Metaphysik der Erkenntnis"[10] [Har65]. The content of the quotation might be translated as follows: The gnoseological (i.e., epistemological) problem involves the ontological one and cannot be treated without it. But it is not the same. Hence, the elementary analysis of the thing must precede the analysis of its perception.

Hartmann was of the opinion that modern ontology needs to be a "critical ontology". This means that it must stem from the experience – pointing to an important twist: Ontology is no longer only concerned with some objective view on reality (which is, more precisely, actually impossible to achieve as we are always judging from a human perspective), but with the interplay (in Hartmann's words: a "being relation" [ibid., p.182]) between the experiencing subject and the things out there. We will, however, never know how much our own representation of the world coincides with the "true" being of the world. The point is that this is actually not important. We can consider it a fact that there is a coincidence up to a certain point, because we could not live well in the world otherwise (in evolutionary epistemology, we would say that this would otherwise impede our functional fitness, cp. [Wuk90]).

To transfer this to robotics, consider the interplay of the different "ontologies" (personal and objective) in Figure 4.1: Whereas both human and robot have a specific representation of the world and both are (ideally) able to adapt it due to the influences given by the world's "reactions", the "being relation" between the designer of the artefact and the artefact itself is not only that they perceive each other, but, more importantly, that the design of *what* the system is able to represent is influenced. Note that this is *necessary* if we want the artefact to do what we as designers want it to do. This issue is related to what has become known as the "frame-of-reference problem" in artificial intelligence of which [PS01, p.112] lists three main aspects: First, there is the

[10] The German title could be translated as "Fundamentals of metaphysics of knowledge".

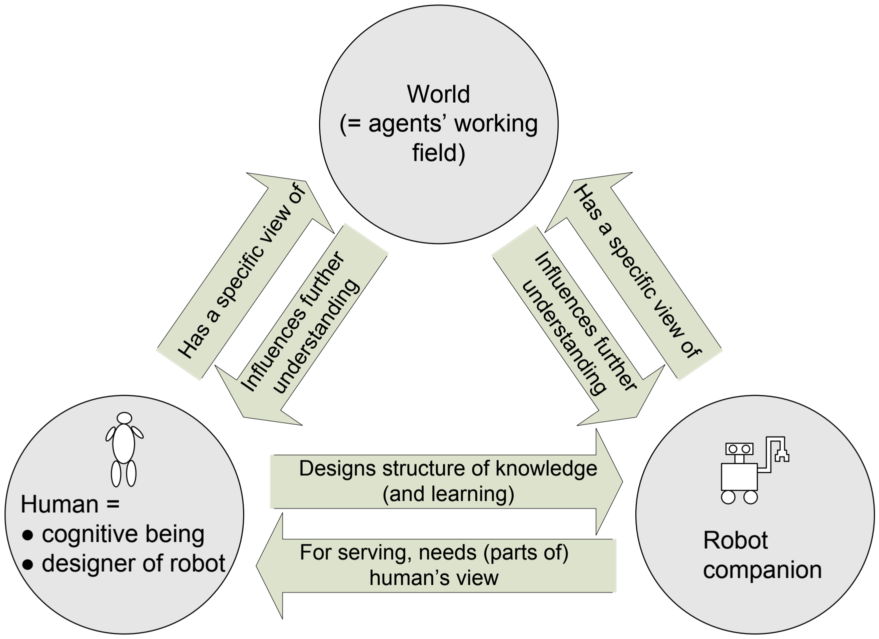

World
(= agents' working field)
Has a specific view of
Influences further understanding
Has a specific view of
Influences further understanding
Human =
● cognitive being
● designer of robot
Designs structure of knowledge (and learning)
For serving, needs (parts of) human's view
Robot companion

Figure 4.1: Dependencies of world-views and roles.

perspective issue that forces us to distinguish between the perspective of an observer and the one of the agent itself (this is what we call first- vs. third-person-perspective). Second, there is the *behaviour-versus-mechanism issue* that states that a behaviour is always the result of system-environment interaction and cannot be explained only by internal mechanisms (in Figure 4.1, this tackles the two-way interaction between World and Robot). Finally, the frame-of-reference-problem also includes a *complexity issue* which just states that the degree of complexity observed need not coincide with the degree of the complexity of the underlying internal mechanism.

In our opinion, the frame-of-reference-problem gets a fourth aspect when it comes to a robot companion: In contrast to the "complete agents" that the cited work [PS01] is concerned with, for a robot companion there is direct need of a deliberate (at least partly) compliance of world-understanding with the understanding of the human agent. It is the reference frame (working environment) which is shared in this case and which thus affords a "shared understanding" of it between the two agents. We have tried to point this out in the figure by the arrows between human and robot companion.

An objective reality is implicitly always taken for granted in computer science, of course (thus, we will not discuss truly solipsistic philosophies). This is reflected by the fact that we construct the sensors in a specific manner. What needs to be done, is, however, to consider in what ways we need to adapt them and first and foremost their outputs in order to level it for a specific understanding of reality in the cognitive system. This means, we need to model possibilities in order for the system to get its ontology out of its experience. The point now is that we would argue that it is perfectly all right if we start from concepts that we as humans have built up (in *our* ontology with *our* cognition), because we want the robot to understand *our* world and *our* view of the world. This is, of course, an implication of the specific *niche* that a robot takes: It is located right beside the human user's side. In order to fulfil a task *for* the human, it needs to understand (at least up to a certain extent) how the human operator "sees" the world. The difficulty is, of course, that the epistemology of the system needs to change with the gradual enrichment of its ontology. This leads to a tricky balancing act as we would need to know in advance how the interplay between perception modalities, pre-given concepts and newly acquired concepts has to look like. We might prime it at the beginning with certain concepts that we impose (which roboticists who have a different goal than a robot companion might reject as being too much of an influence), but there should be the possibility to change the overall understanding (i.e., the personal ontology) with growing experience, much as what holds for organic agents: "In his [Aristotle's] terms, one might say that human infants are born with innate predispositions to structure their perceptual inputs along certain lines, but as they gain more acquaintance with objects (e.g., see them falling unsupported, find them resisting their actions) they are caused to abandon primitive conceptions for more elaborated ones" [Rus99, p.258].

4.4 Common Notions

After having outlined some of our main presuppositions, we will now explain our understanding of the notions "intelligence", "cognition" and "consciousness" which have become very fashionable in computer science, too. This is important as these terms are controversial both because they are used in a variety of different contexts as well as because they generally lack a clear definition. Especially artificial intelligence, of course, has discussed the meaning of the term "intelligence" for decades. Here, we will carefully examine how "intelligence", "cognition" and "consciousness" can be understood from a pragmatic and – for computer vision/cognitive systems engineers – useful perspective. Thus, we will not review the meanings of these notions in depth as it has been done in the humanities, but rather keep it simple for a manageable synthesising view. To this end, we will tackle positions of which we can get impulses and leave out discussions that might be relevant from a philosophical point of view, yet not from an engineering one. As will be outlined in more detail in the following, the bottom line is that all of those notions must be seen in a gradual sense, just as Jeff Hawkins states in [HB04, p.180] for intelligence: "All mammals, from rats to cats to humans, have a neocortex. They are all intelligent, but to differing degrees."

4.4.1 Consciousness

Let us start with one of the most controversial notions in connection with artificial cognitive systems: "consciousness". The question may be raised why we tackle consciousness at all in the context of robot companions, because we have argued before that we will stay on a functional layer without the need to implement all eventualities of humans. However, we might as well say that even consciousness fulfils specific functions – so we need to have a look on whether we have a requirement for it.

We do not argue for a robot to necessarily be self-aware or to have a feeling of subjectivity. These aspects are usually used for distinguishing inferior bio-organisms from higher developed mammals. Instead, we are arguing that dealing with our environment in a meaningful way simply needs some sort of "taking part". Hence, we are only using the term consciousness in order to motivate our approach to the necessity of intentionality. This implies that most of the involved discussions on consciousness led in philosophy are of minor interest for us. More precisely, we will *not* tackle (the very interesting) discussions on "qualia" (which started with the seminal work of [Nag74]), Ned Block's distinction between "phenomenal consciousness" and "access consciousness" [Blo97] (roughly the same distinction as between the hard and the easy problems of consciousness, see below) or David Chalmers' famous Zombie thought experiment in which an observer just cannot know whether a behaviour is due to a conscious event or not [Cha96]. All these considerations would be needed if we are actually interested in implementing the full range of aspects related to consciousness. However, as said, whether our robot is fully self-aware or not is not of interest for us in home robotics where we just want behaviour to be shown, i.e., a task to be done. With other words, we are interested in the position of "weak AI" (the robot should act *as if* it was intelligent) and not "strong AI" (it is *actually thinking*). The reason is twofold: On the one hand, we believe that we need to first tackle the problems of weak AI as there are still a lot of unsolved issues before we can move on to strong AI. On the other hand, for the tasks of a robot companion that we seek to achieve, the requirements of weak AI suffice, as we will see on the basis of Chalmers' list further below.

One – for engineering purposes – relevant philosophical consideration, however, is the one of John Searle, who reduces the problem of consciousness to a matter of perspective [Sea97]. This view can be interpreted that consciousness is just the first-person-perspective of an experience, the "What-is-it-like-for-Me" of a specific event. For us, this would remove the need of investigating whether there is consciousness or not (because we as observers could just not mind if an action is due to a conscious event or not). Taking this position would be, for sure, the easiest way for us and would indeed even support our view that we are only interested in third-person-matters. However, discarding other approaches would disable us from spotting underlying functions of consciousness that might be of importance. So let us continue our search.

To use Chalmers' notions [Cha95], robot companion research is not interested in the "hard problem of consciousness", i.e., the subjective aspect of experience (qualia), but it needs to be interested in the "easy problems of consciousness", which is in line with what we called functional aspects. Chalmers lists some examples [ibid.][11]:

"The easy problems of consciousness include those of explaining the following phenomena:

1. the ability to discriminate, categorize, and react to environmental stimuli;
2. the integration of information by a cognitive system;
3. the reportability of mental states;
4. the ability of a system to access its own internal states;
5. the focus of attention;

[11] For the sake of better readability, bullets in the original text have been changed to an enumeration.

6. the deliberate control of behavior;
7. the difference between wakefulness and sleep."

Some of those *need* to be tackled by robotics (e.g., 1,2,4,5,6), some could even in the context of humans probably be subsumed under different notions than consciousness (1,2,5) and some pose hard challenges for a robot companion (2,4,6,7). Overall, however, they seem to be manageable to a certain extent.

Looking at older philosophical positions that are less interested in explaining how consciousness actually comes into place and works (i.e., before such discussions were named "philosophy of mind"), there is an interesting combination of the notions "conscious" and "cognitive" to be found in the epistemological theory of objects (in German: Gegenstandstheorie). The theory of objects was an ancestor of Husserl's phenomenology and for its founder Alexius Meinong, *consciousness is always levelled at something that could be an external or internal object* [Sch98]. External objects can be thought of as real, existing entities (i.e., objects in the everyday sense); with internal objects he refers to inner procedures. Therefore, the real existence of objects is not a necessary criterion for something that consciousness focuses on. Of course, the emphasis of conscious as being conscious *towards* something (which already was Franz Brentano's principal assumption), can be found similarly in a second aspect that is frequently investigated in the context of cognitive systems: attention. It must be stated that object is meant in the broadest – not necessarily materialistic – sense here. The notion of an object from the usual engineering point of view is thus totally different from the philosophical one. Furthermore, it is to be underlined that "intentionality" is meant exactly in this sense: as the guidance of our consciousness, of which attention is a necessary but not sufficient part. We will later include intentionality to the cognitive functions that we judge as being necessary for cognitive vision. We should additionally mention here, that one of Meinong's central claims was that those objects that we are always conscious towards, i.e., that our intentionality is guided towards, can either be real or ideal, the latter not being concrete instances of things (i.e., objects in the materialistic sense) but rather anything else – which comprises impossible objects or abstract concepts of objects [Pup13]. This already introduces a hierarchy of objects, from simple real entities of things to higher-level abstracted, nonexistent ideal objects (in the philosophical sense).

Consciousness is, of course, not only tackled in philosophy but in other disciplines as well (such as cognitive science or psychology). We would like to especially tackle a quite recent field, namely "neuro-psychoanalysis" which tries to link psychoanalytic theory with recent neuroscientific findings. Its main proponent is Mark Solms, for whom consciousness is one of the basic properties of the mind[12]. For him, the mind disappears as such if there is no consciousness. Thereby, consciousness is inherently evaluative and has the purpose of letting us be aware of feeling pleasure or unpleasure [Sol08]. So it is concerned with the mediation of inner and outer world in the sense of making the system react to tensions between these two. Furthermore, "[...] consciousness may be defined by its *content* (what we are conscious of) as well as by its *level* (how conscious we are)" [Yov08] (referring to [ST04]). The first view is covered by the intentionality

[12] In [SV08a], we contrasted Solms' psychoanalytic view with our functional approach in detail.

principle which has already been mentioned. The second view is insofar interesting as we could – for the attempt of engineering a cognitive system – adopt this *gradual view* on consciousness and deliberately exclude subjectivity and qualia (i.e., the hard problem of consciousness). Cognitive systems (be it organisms or artefacts) can this way be *more or less conscious* – total consciousness finally comprising self-awareness and meta-thinking. The latter is not needed in a functional artificial system – excluding, so to speak, qualia and allowing Zombies, to put it in Chalmers' notion. Such a system would not have subjectivity as we humans know it ("we" implies that none of the readers is a Zombie...). Maybe qualia arises anyhow as a "deus ex machina", rehabilitating anthropomorphism – or maybe not. For a full-fledged mind, we agree with Solms [Sol08] that it is not a matter of indifference whether there is consciousness or not – but a full-fledged mind is not what should be in the fore for a cognitive robot companion.

If our bottom line of those considerations about the understanding of consciousness is that our task at hand is to provide the *functionality* of consciousness, we might take away that:

- "Consciousness" can serve as a notion associated with the principle of *intentionality*, needed for implementing task-directedness (the guidance towards an object – in the philosophical sense)

- This includes the mediation between Inside and Outside and hence the functionality of motivation.

4.4.2 Intelligence

The notion "intelligence" has been of central interest over the last decades both in psychology and in artificial intelligence. We are not aiming at extending this (often indeed circular) – definitional round dance, but rather make clear in what way intelligence needs to be understood and discussed in the context of robot companions. Therefore we will restrain the State of the Art to some relevant aspects for our view.

In [PS01, p.6 f], an overview over both scientific and common sense definitions is given, ranging from tackling problem solving via memory to motor abilities. Intelligence can moreover be defined gradually (like "Einstein was more intelligent than the average person"). Strictly spoken, this would imply the existence of what psychologists termed "general intelligence g" of which Sternberg makes clear that it cannot exist [Ben99, p.10]: "Dismissing a century's efforts, Sternberg argues that g does not exist and instead proposes a theory of 'successful intelligence' [...]". In the course of argumentation, Sternberg [Ste99] cites studies by which it becomes clear that intelligence is not "[...] just a single thing that can be measured by a conventional static test of intelligence [...]" and extracts three main aspects that are comprised by intelligence: analytical, creative and practical aspects.

The search for one single all-embracing principle named intelligence that covers all of those aspects has been ill-posed as the infamous history of artificial intelligence reflects[13]. So if there is no such unitary thing that we could call intelligence, what do

[13] E.g., compare the actual (non-generalisable) achievements with the famous sentence by Newell

IQ-tests in psychology measure? The American psychologist Edwin Boring gave a (now famous) answer to that already in 1923: "[...] intelligence as a measurable capacity must at the start be defined as the capacity to do well in an intelligence test. Intelligence is what the tests test" [Bor23]. With other words, you get what you ask for and this kind of directed and selective measurement is what we can use in a positive turn to apply to cognitive perception in computer science. It can already be suspected that this leads to a strictly behaviouristic view on intelligence.

In line with the argument that there is no such thing as general intelligence goes the view that "[...] extremely young children are best viewed as 'universal novices,' while adults are best viewed as individuals who have become expert in the wide range of problems that daily life (or school) presents" [Cas99, p.48]. This could be shown by studies with experts in chess, medicine or physics that found that "[...] the distinguishing feature of experts was the vast network of specific knowledge that they possessed, not a more powerful set of general heuristics or strategies" [ibid.]. The author comes to the conclusion that this is also the reason why "[...] expert systems on a computer were more successful when they built a huge repertoire of specific knowledge, and a powerful way of representing that knowledge, than when they tried endowing the system with more powerful problem solving strategies" [ibid.].

If we adopt this view that adults are "only" experts in a large variety of specific problem areas rather than have some kind of logic that implements general intelligence, then we can also see why situatedness has become a popular view with respect to living cognitive systems: The system is herein adapted to the right kind of niche in which it is living. An interesting study pertaining to this fact is cited by [Ste99] where the same kind of mathematical problem (relating two variables, e.g., weight and price of a consumer product) could be much better performed by housewives in an applied everyday situation (e.g., a supermarket) than by the same subjects on an abstract level in a testing situation. This implies that contextual information has massive influence on the "intelligence" shown.

We can now again ask what this entire discussion implies for our objective. One – in our opinion – very important implication is that intelligence can only be judged from a third person perspective. Intelligence is nothing that is "inside the agent" but rather something that can only be observed from outside, i.e., behaviouristically. It can, for example, be assessed when the agent is *doing something* or when it is *interacting* with other agents or objects. Furthermore it is gradual. This means that in a given situation there might be many solutions, some of which seem "more intelligent" than others. We can overstress this view and even argue that if we humans judge ourselves, we do it to a certain extent in a behaviouristic (here: in a self-judging) way. We might consider a statement such as "Here, this person seems to be more intelligent than I am" or "In this situation, I behaved unintelligently" – we are always judging our behaviour, our performance, and furthermore we do that *in relation to others*. This is different to consciousness in this respect that we are conscious of ourselves even without this kind of comparison, and even more: only with respect to *ourselves*. Consciousness and intelligence are consequently two sides in the same game, the former analysing a

and Simon: "A physical symbol system has the necessary and sufficient means for general intelligent action" [NS76].

given situation from the first person perspective, and the latter from the third person perspective – both are gradual and gradually judged. The point now is that engineers – emphasising on the functional fit (as radical constructivism calls a concept very similar to "viability") of the artificial system to our, human's life-world – are constructing those machines only for the behavioural perspective (at least up to now) as they want the system to behave intelligently (which is totally fine for robot companions). In Jackson's famous textbook on artificial intelligence, we can find a related view [Jac85, p.5]: "To summarize the definition in one phrase, one might say that intelligence is the ability 'to act rightly in a given situation'." Secondly, we have seen that a wishful kind of intelligent behaviour might rather be achieved by careful investigation in which domain a cognitive system will find its niche, who is judging it (the user) and what kind of knowledge and reasoning capabilities are necessary in order to fulfil its expectations.

Concluding, intelligence is what engineers really want to achieve in "cognitive technical systems". Consciousness might be a way to achieve it – but the true goal is to build machines that *behave* intelligently – the behaviouristic touch is obvious. To come back to Chalmers' notion again, Zombies (i.e., non-conscious creatures) may behave exactly the same way as (conscious) humans do. We can likewise look at two kinds of robots: conscious (self-aware) ones and non-conscious ones. If both achieve the same performance (judged by the external third-person perspective), it is obvious that consciousness (to the last extent) is not necessarily important.

Our take away of this section finally is:

- "Implementing intelligence" is an ill-posed goal, as intelligence is an effect that is *shown* and that can only be studied *behaviouristically* from a *third-person-perspective.*

4.4.3 Cognition

After having outlined our understanding of "intelligence" and even "consciousness", we will now present our conception of "cognition" as this notion is obviously of interest for "cognitive robotics" and "cognitive vision". Again, we are especially interested in how this term relates to the others and in the implications for a robot companion. Unfortunately, cognition is itself a term as controversial as the others. There are a lot of definitions that span diverse areas, and it makes no sense to enumerate all the different views. The following examples are therefore not meant as an exhaustive definitional review, but should rather show typical examples.

For instance, [Kei99, p.165] locates the following ways of framing problems with respect to cognitive development: "[...] domain specifity versus generality, hybrid versus homogeneous learning architectures, implicit versus explicit forms of knowledge, and abstract versus concrete forms of thought." Another example of such a broad field of investigation that is concerned with the notion cognition would be [Mey08], in which cognition is defined as the collective name of all processes and structures that are related to perceiving and recognising. Note that in this definition, cognition is shifted to the subjective sphere whereas in others, the focus is behaviouristic again: For example, in [Mat87] we find that cognition is assessed by an observer due to potent action or effective behaviour in this area. This, again, points to the fact that cognition needs to be bound to some kind of *action* in a specific *context*.

Finally, the definition of "cognitive vision" shows this broad field of issues tackled, too [Ver04]:

> *A Definition of Cognitive Vision*
>
> A cognitive vision system can achieve the four levels of generic computer vision functionality of detection, localization, recognition, and understanding.
>
> It can engage in purposive goal-directed behaviour, adapting to unforeseen changes of the visual environment, and it can anticipate the occurrence of objects or events.
>
> It achieves these capabilities through learning semantic knowledge (*i.e.* contextualized understanding of form, function, and behaviour); through the retention of knowledge about the environment, about itself, and about its relationship with the environment; and through deliberation about objects and events in the environment (including itself).

Note that in this quotation we find the difficult (and unsolved) necessity of *understanding*. Furthermore, we can see the imperative connection to *behaviour*, to *adaptation* and *anticipation*. Moreover, we find that important parts are *semantics* and *context*. In this, we see that cognition comprises all the hard parts – both for the engineering sciences and the humanities. Thus we can add to our tentative list of cognitive functions which only holds "intentionality" so far: "prediction" (another word for anticipation; for us connoting temporally closer events, therefore better suited for perceptual purposes), "abstraction" as well as "generalisation" (both necessary for flexible adaptation) and "symbol binding" (cardinal for semantics and thus understanding). For now, we can put in a nutshell how the notion *cognition* relates and contrasts with *consciousness* and *intelligence*:

- "Cognition" is not necessarily a conscious process (subliminal learning is very likely to be a "cognitive process"). Hence, cognition and consciousness need to be separated.

- "Intelligence" has been stated as being *only* applicable to the third person perspective, cognition, however, can be *judged* behaviouristically, but it is a process of the system.

We would define cognition as the superordinate concept that allows for the functional fit of the system. Of course, cognition, consciousness and intelligence are related. If a system (biological or artificial) behaves functionally fitted to its life-world, a third person would call this behaviour "intelligent" – and in the biological case, the activities can additionally be judged "consciously" by the system itself. This kind of "definition" of cognition stays shallow, of course, and can only be considered as a working hypothesis. It remains to be seen whether further research findings from the diverse disciplines involved in cognitive science will lead to a more concise description.

Our take-aways from this section are finally:

- We can talk about “cognition” if we see that the system is flexible, “understanding” and adaptive.
- Cognition can also be used as collective term for what is *shown* (third-person-perspective) as intelligence and which could include (first-person-perspective) consciousness.
- Needed parts of cognition include adaptability (generalisation, abstraction), semantics (symbol binding) and understanding (entailing prediction).

4.5 Emphasis on Functions

Until here, ever now and then we have suggested that a functional view on cognition is the only feasible way in which to proceed if we want to systematically tackle the complex issue of vision for a robot companion. We will now summarise the motivations and reasons to do that and at last name the (subset of) functions that should be started with. Note that we are not claiming to give an exhaustive theory of how cognition works in living subjects. The only claim we make is that this is one way how “cognition” in a future robot companion could be constituted.

4.5.1 Why Choose a *Functional* Approach

We have already argued that one possible and likely way that nature found to solve the complex cognitive efficiency might be due to a large number of domain-specific expertises that we acquire through our lifetime. It is equally likely that the representational mechanisms behind that are quite diverse and can *not* be subsumed under a common principle, as the different paradigms that have all had their en-vogue-time suggested (e.g., symbolic and subsymbolic, dynamic systems- and situated approaches, also cp. quotation on p. 59). However, some kind of general cognitive *structure* might still be in place, which we would see as exactly pointing to some functions that seem feasible to assure a good functional fit. In any case, the emphasis should lie on the fact that everything we perceive and act on is an object that has certain *meaning to us*. We could file all the traditions into these high-level goals and certainly we could find examples where some approaches are more successful than the others – the point is that the question for such a complex far goal as enabling an artefact to carry out some complex task is on how these different implementation paradigms can interact, where and when the transition from one “way of thinking” to the other takes place (similar to Minsky’s conception in [Min06], although he focuses on emotions).

With the just mentioned traditions, we refer to the historical strands in cognitive science. [Cla01] lists three major paradigms: symbolic (such as in “Good Old Fashioned AI”), subsymbolic (“connectionist”) and the dynamic systems approach (actually going back as far as to cybernetics in the 1940s). All of these approaches have in common that they followed the goal of understanding intelligence in order to construct artefacts that apply this intelligence and perform some desired work. It was probably due to this dual goal that religious wars have been fought whether one "intelligence framework" is better than the other. What, in our opinion, has been missed out is what usually the case is: *aurea mediocritas*, the happy medium: All have their right and all have their place.

An example: Thinking in terms of "connectionist intelligence", *one* possible implementation of prediction can be achieved via recurrence in the network (cp., e.g., the early works of [Hop82], [Jor86] and [Elm90]). This stays, however, on quite a simple level and with restricted temporal distances between action and prediction. Thinking in terms of human prediction of what might happen in a year (we could call this anticipation), we will most likely rather prefer the symbolic approach, where an event constitutes one kind of symbol and every "consciously thought-of" eventuality another one.

Such conjoining issues can lead to a straight-forward and conservative view with respect to perception: on a "low" level, a connectionist approach reflects what computer vision usually does, namely statistically relating input – a purely symbolic approach seems like unnecessary and implausible overhead. But what is later *done* with the output of these lower levels seems to be better explained with a symbolic approach. For instance, we might use perceived *(proto-)objects* for understanding a scene, thinking about possible agents interacting with the objects, and so on. All "around this", the embodied dynamical approach of cognitive science would explain how causal interactions happen, how one thing leads to another, creating a complex web of actions, perceptions and thoughts this way.

It is not entirely clear why the different positions have been arguing in a principled manner. It rather seems that they are arguing on different *layers*. Note that the just described explanation where to localise subsymbolic, symbolic and dynamic approaches is just *one* possibility and does not mean that there could not be different modes at work on any layer. The burning question for us will be later on, on which layer to start with our artificial visual perception[14].

What has been said until now results in the conception that the issues involved in cognitive home robotics should be discussed on a functional layer. This is not totally new, of course, instead there is especially one tradition which also did not focus on implementation issues (as also did not dynamical systems- and embodiment approaches as they are a bit fuzzier as well) but rather on similar functional ones: the position of *Functionalism*, with which we therefore need to confront our view[15]. Coined by Hilary Putnam in the 1960s, functionalism is one of the classical positions of the mind-body debate in the philosophy of mind. Its mainstay is that *mental states* can be defined as

[14] Interestingly, a similar plea against such religious wars comes from a developmental psychologist. In [Kei99, p.170], F. Keil argues that "All too often debates in psychology swirl around absolute dichotomies. [...] Such oppositions may help promote debates, but they can obscure the possibilities of more mixed models. [...] A young child is considered associationist, or concrete, or exemplar-based in her representations, while an older child is considered rule-governed, abstract, or principle-based in her representations. There are major developmental differences to be sure, but they may not rest on such dichotomies. In particular we can ask if a hybrid architecture might not be more reasonable in many cases."

[15] The following digest of Functionalism has been written with the help of [Pal99, p.623].

functional states, and these mental states are supposed to be causally connected to other mental states, the environment and one's own behaviour. Our focus, however, is that the *capabilities* that a cognitive system needs to *show* can be achieved by implementing certain functions (note the behaviouristic touch). We are not so much interested in the existence of mental states as such. What we do adopt, however, is the opinion of "multiple realisability", but in a behaviouristic turn, which means that different physical implementations can *show* the same behaviours because of implementing the same *cognitive functions*. We would, however, loosen the constraint that the same function is served by causally connecting inputs and outputs in the *same* way in all these different implementations. We would rather say that it is exactly the challenge to find those connections and processing methods which allow – despite of *different* sensors and *different* thinking layers – for the same "intelligent behaviour".

Of course, it can be safely assumed that mental states and cognitive functions cannot totally be separated or that it is just a matter of definition of "mental state", however, we propose that we should work on implementing the latter before tackling the former (if ever) and can nevertheless reach functionality that a robot companion is in need of. To put it in a nutshell: As we are not interested in a cognitive theory about humans but just borrowing findings in order to push development on a technical basis, we can safely accept the objections to Functionalism (e.g., that *qualia* is probably not fully explainable by mental functions) and nevertheless pursue an understanding and adoption of cognitive functions for robotics[16].

The functional level is especially the right one for an interdisciplinary project where views and stances are quite different at first sight, cp. [Sol08]. As already indicated, it remains an open question and has been quite a debate in conjunction with Functionalism whether we can model "everything" functionally. What we need to ask in cognitive robotics is what *we need to model* and whether *these* parts are open for a functional approach. In Section 4.3.2 we encountered the example of feelings as being on the one hand a "qualitatively felt mental state" (catchword qualia), but on the other hand have "just" the function to mediate between the endogenous needs of the organism and the external objects of those needs [ibid.]. Rigorously spoken, the latter is relevant for robot companions, the former is not. We said that this is related to consciousness, which we considered in Section 4.4.1 in more detail. There we saw that one feature of consciousness according to [ibid.] is that it allows to *know* (by feeling it) when there is a mismatch in our outside-inside-balance (e.g., feeling dreadful because of something that might happen, etc.) – this sub-function would be relevant for a robot companion as well; it allows us to achieve reasonable, weighted and assessed prediction – a capability that we will define as one of the basic functions needed for cognitive perception.

There is one big advantage that we can name in favour of the functional approach: It allows us to systematically enhance the capabilities of the agent. Without the need to first know how emergence on such a large scale could work, we can try to implement gradually more functions – the hard task is to find a common integrating framework and means of representation. This is mandatory in a synthesising science like computer vision. It is ill-posed to believe that the more complex we construct a system (e.g., by

[16] The classical work on Functionalism is [Put60]. For a concise summary including typical objections, see [Cla01].

a massive neural net), the more likely we will have success. We must be very careful in expecting what will happen. It is definitely not likely that the more abstract levels we get and the more complex stuff gets, meaning will just arise. To think of a difficult problem (like how the mind works) in terms of attributing the solution to complexity is seldom correct and hardly useful when synthesising.

Our emphasis is still on perception, and hence we focus on functions that are related to perceptual capabilities as well as the implied consequences for representation. How much such a selective progress is itself in contradiction to a radically embedded and constructivist view keeps being an open question. The simplest answer is that we designers of artefacts always face this problem because we just do not know better. We need to start from small parts – as long as we keep the big picture in mind, this should be fine. The functions which we will define in this section later on will for sure not form an exhaustive list, but they will help us to show that a thorough ontology is needed, which is more than just a database of known objects, but rather constitute the possibility to be the interface to higher cognition and on which – in the best case – we can gradually connect more cognitive functions.

We already defended our position that taking human cognition as *model* does not imply to take it as *blueprint* and therefore rejected the confinement to "biologically plausible" implementations (see Section 4.3.2 on anthropomorphism). This, however, confronts us with a common critique the mentioning of which we postponed to here in order to first clarify our understanding of cognition and of our functional attempt. The critique goes like this [Joh99, p.147]: "The dissociation of cognitive development from developmental biology over the past few decades has, in my view, been counterproductive due to the investment of time and resources into the exploration of hypotheses and models that while perhaps philosophically attractive were biologically implausible." The same could be said for our approach: Why follow a human model or parts of it when not also borrowing from the structure of the substrate in which humans are implemented, i.e., neural networks? For us, working towards a functioning home robot, the goal is *practicality*. The strong binding of robotics on biology is not necessarily the right way to achieve this. Consequently, we also object to statements like "A controller for cognitive robotics need to be an artificial neural network" – from a roboticist's view, this biological plausibility is not needed (we are not trying to explain intelligence as it is). This is for sure an arguable statement, but our conviction is that for our goal "robot companion", we can study cognition on this abstract, functional layer, and moreover, we can do this to a certain degree piece-wise. The weak assumption here is that the research frontier can be pushed further although possibly human performance is a (necessary) mixture of all parts. Speaking of the human, we have to look at respective disciplines (psychology, philosophy), but we need, as said, be very cautious with anthropomorphic implications.

Summary: Functional Layer

We can now list the aspects that led to our conclusion to focus on the functional layer. In short, our hope is that with reflecting on necessary cognitive functions, we might be able to see the wood and not only the trees when standing right in the middle – impossible for pure vision.

- "Cognitive perception" (if we allow such a kind of pleonasm) is more than putting cognition on top of pure bottom-up processing; there are capabilities that vision serves and those feed back on what vision needs to deliver. These capabilities are the cognitive functions, in which perception is "embedded" (catchword *vision as part of cognition instead of cognition on top of vision*). This holds especially for the case of robot companions, where we called this embeddedness *situated perception*.

- Functions should provide our system with higher-level *functionality* by allowing to drive lower-level inputs and outputs. With other words, the requirements of the functions tell us designers what the vision sensor should deliver, as they drive what should be seen.

- Taking such a higher-level viewpoint on vision allows us to take a sort-of analytical position which is in general inaccessible for the engineering sciences. This is related to the emphasis on the behaviouristic view on robotics.

- On a meta-level, the functional layer is best suited to find interdisciplinary points of contact.

- All what has been said does not mean to emulate the human, i.e., not to sketch a theory of human cognition.

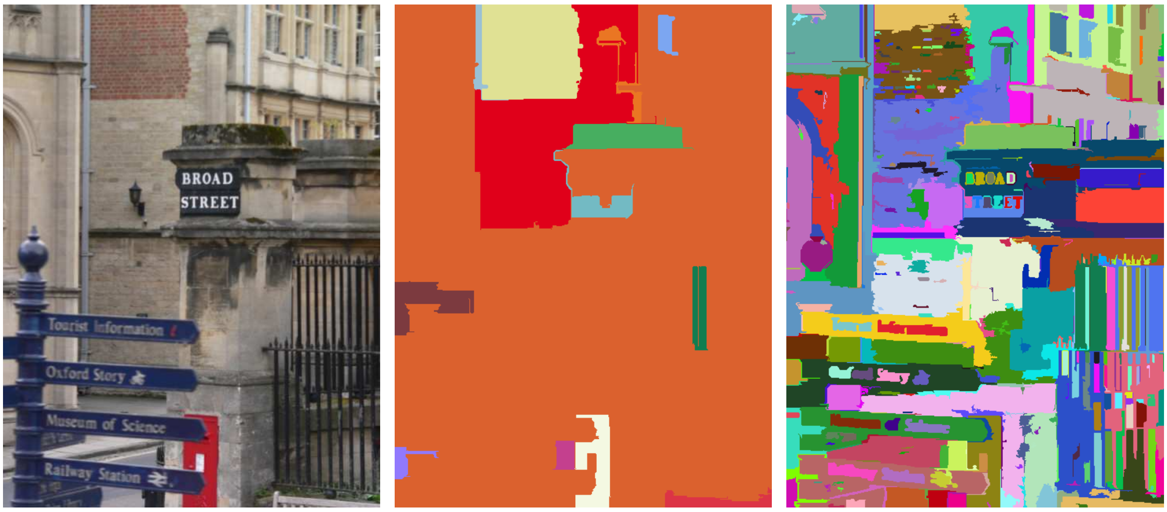
- Especially, using human "inspirations" does not (necessarily) refer to the implementation layer, i.e., neural networks.

- We will focus in the following only on functions for *vision* (which can partly be generalised to *perception*). Hence, the list will not be exhaustive with respect to the overall system.

- The functional approach allows for a change in the level of description, e.g., to locate subfunctions, superfunctions, etc.

This leads us back to the appeal we made in Section 4.2: Problems instead of solutions! The implementation layer can be discussed if we know the crucial topics that would enable *situated perception for a future robot companion*. Otherwise, we are not pursuing what is actually at stake: exactly this goal.

Long enough now, we have talked about cognitive functions in general needed for robotic vision in a home scenario. It is time to nail ourselves down to what functions we are talking about. In the following we will try to clarify why we propose *intentionality*, *abstraction*, *generalisation*, *symbol binding* and *prediction* (see requirements list in Figure 1.2 of the Introduction) as crucial functions involved in the interplay between perceived raw data and further reasoning and action of an agent. Those are the functions that we could already extract from our discussion on the notions in Section 4.4. These functions are needed for acting, i.e., they are needed for *processes*, be it *inside* the robot or in the world *because of* the robot – i.e., the duty of a robot companion. Once again we need to emphasise that we are only interested in functions that relate to perception – there will be more cognitive functions needed when it comes to a complete agent.



(a) Broad Street in Oxford... (b) ...segmented coarsely... (c) ...and fine-grained.

Figure 4.2: A matter of intentionality: which one is the "correct" segmentation?

4.5.2 Intentionality

As paradigmatic example for our quest through cognitive functionality, let us imagine a typical robot companion scenario: A robot is urged to retrieve a specific object – say, a coffee mug – from another room. The robot has not seen *this* mug before, however, it knows what a mug *is*, i.e., it knows its *meaning to a human*.

Obviously, the search (and hopefully: retrieval) of the mug must first of all be motivated, i.e., the robot needs a task to fulfil. This is what we refer to as intentionality: The "sense of direction" towards a complex goal. We argue that this needs to be bound to perception as it will influence not only the detection of the cup, but also all the intermediate steps: the search for the door, the location of walls in order to find the corridor, the extraction of those planes relevant for the task, i.e., the most likely locations such as a cupboard or the kitchen table.

The importance of intentionality for "intelligent behaviour" is also underlined in [Bor99, p.26]:

> Fresh evidence indicates that infants fill in missing parts of figures, integrate information over space and time, form categories and concepts, abstract prototypes, match across modalities, distinguish the behavior of animates from inanimates, discriminate small numerosities, imitate extensively, and retain memories over days and months [...]. All these understandings in turn make it possible for the infant to act with intentionality, a critical feature of intelligent behavior.

The fact that a high-level issue such as task-orientation (intentionality) is important for *perceptual* questions, can easily be seen with reference to its influence on typical computer vision tasks. They get a different twist because of this action-orientation. Consider, for example, the case of segmenting an image (Figure 4.2). Segmentation is one of the core problems of computer vision. However, without knowing what information is relevant for the current task, segmentation is an ill-posed problem. Figure 4.2(b) and 4.2(c) show two examples of segmenting image 4.2(a)[17]. Both segmentations are not perfect, however, in case we are searching for the partly occluded red mailbox at the bottom of the image, the coarse segmentation is totally fine, whereas in the fine-grained case, it will be hard to find it due to high clutter. On the other hand, sometimes such a fine-grained segmentation is necessary, for example when we need more information of the road sign (e.g., to further apply object character recognition so we know that we are on Broad Street). In short: Depending on the *task*, the parameters will change drastically.

We have also met intentionality in connection with consciousness when we said that the former is an important part of the latter. We also said that consciousness has another important duty in humans: the mediation between the inside and the outside world. We are not claiming that a robot should be conscious in the everyday sense (cp. Section 4.4), but we can state that this second job of consciousness might be implemented as some sort of drive which is on its part again connected to intentionality. Consider, for example, that there is a motivational system at work that enhances "pleasure" when a task has been fulfilled properly and "displeasure" if a request of the human operator is ignored. This way, the mediation between the "feeling" when starting and the "feeling" when completing the task, is tightly bound to the task (intention) itself.

4.5.3 Prediction

Prediction is a the second cognitive function that we consider crucially important for situated perception as it gives the agent information what is likely to be perceived in the next step. Thus, it helps to break the high-level task knowledge down to the current situation. This can be achieved by integrating contextual knowledge (where the robot currently is, which room is likely to have what kind of objects in it, and so forth), which in turn helps crucially to prune the search space. This means that sense can be made in the image faster when the robot is able to compare only a subset of known concepts, e.g., kitchenware in a kitchen, instead of the whole bunch (maybe including cars, houses, gardening tools,...). This is what we refer to as prediction with respect to vision: the anticipation of what is likely to be seen next. Philosophically, especially Popper's epistemological falsificationism radically accentuates the role of prediction: "There is simply no new knowledge without some kind of earlier knowledge, some kind of expectation, upon which it is a modification. And such modifications occur especially when earlier knowledge runs into trouble – for example, when an expectation is disappointed, when it gives rise to a problem" [Pop96].

In [HB04], the central theme of the whole book is prediction. Hawkins is very much focused on the neural level, however, the central topic is that the human brain is constantly predicting what will happen next, be it during preparing pancakes in the kitchen or when putting the foot down on the ground for the next step while walking. Most of these things are done "unconsciously". The reason for *knowing* that we predict

[17] Both segmentations have been performed using [FH04] with differing parameters.

is explained by the author "[...] because if any of those common motions had had a different result from the expected one, I would have noticed it" [ibid., p.91]. In the same book, Hawkins cites Rodolfo Llinas from the School of Medicine at NYU, who also suggests prediction as an important function: "The capacity to predict the outcome of future events – critical to successful movement – is most likely, the ultimate and most common of all global brain functions."[18] Strongly related to this function is the impact of *feedback* (for a biological brain, these are the feedback axons from the neurons back to "lower" cortical layers, e.g., from V2 to V1). [HB04, p.156] enumerates some of the implications, e.g., the possibility for thinking about things that do not take (or have not yet taken) place:

> Finally, in addition to projecting to lower cortical regions, layer 6 cells can send their output back into layer 4 cells of their own column. When they do, our predictions become the input. This is what we do when daydreaming or thinking. It allows us to see the consequences of our own predictions. We do this many hours a day as we plan the future, rehearse speeches, and worry about events to come. Longtime cortical modeler Stephen Grossberg calls this "folded feedback." I prefer "imagining."

This capability is strongly linked to how we would see prediction as necessary function for our robot companion – not with regard to implementation (neurons), but that there should be massive feedback on the where and what there is likely to see (from the scene and context descriptions), in order to allow for concerted and meaningful interpretation. Feedback, however, can even be more. In our domain of "perceiving what the environment provides", there should be "predictions" from higher layer to lower ones. Marr (cp. Section 4.2) pointed to different stages in visual processing in an upwards fashion: from simple image features to high-level scene understanding. The prior explanation of prediction would be the feedback from scene description to the object level. Most probably, however, that there should be feedback connections all in between, i.e., between features (line, edge, corner), proto-objects (cubes, cylinders), objects (this cup, that hat), and scene description (my cup next to the hat on the table in the kitchen). We will later see what this implies for the symbolic approach (see Section 4.5.6).

Both, intentionality and prediction, are related to "attention". In fact, we would argue that for the perceptual domain, intentionality is what the task drives the robot to head for and hence more global; prediction is more local in the sense that it defines where objects are searched in the current image and which objects are hypothesised to be existent in the current scene. Seen this way, attention is what these two functions together lead to; a combination of both that guides the field of view and the regions taken under closer inspection. Whether we judge attention to be a separate function or the combination of other two (maybe sub-) functions, is theoretically a matter of taste. We prefer to tackle intentionality and prediction as they serve a similar role for the visual domain.

[18] [Lli01], cited after [HB04, p.90].

4.5.4 Abstraction

Abstraction is the third cognitive function that we want to address particularly with regard to vision capabilities. At *abstraction*, the bottom-up and top-down processing meet. Having generated a prediction of what is likely to be seen in the image, the robot might now be able to take the feature input from the camera in the region of interest (e.g., the kitchen table) and try to find those objects that the scene description proposes. However, some or all object instances might look different with respect to size, colour, shape, and visible surfaces to those that the robot has learnt previously. So as to detect a coffee mug which the robot has not seen before, it needs to have a representation of the *object concept* of a coffee mug (one could also imprecisely say: the category of mugs) – i.e., the constituting substrate of what a mug "is like". For humans, we take this capability for granted – we can look at a new car model and immediately "see" and "know" that it is a car. Most likely (although possibly not always) we take different modes in account for doing this – of perception (smell, taste,...) as well as other cues (context, location, time,...). We propose that this is done by relating features together in a guided manner (because of prediction and intentionality) – this kind of abstraction lets us see relations between features that "could" be a mug or a coffee can, to continue our example. Maybe it could actually also be a silo – but in the context of a kitchen, this is unlikely. In any case, we need some kind of mechanism which abstracts from actual appearance to a *concept* that holds for all mugs.

We already encountered the philosopher Meinong when talking about the relation between consciousness and intentionality. The same philosopher also talked about *ideal objects* which can be abstract entities. This means that the combination of being guided towards something, the prediction what "could" be seen in the image and the abstraction of feature relations into a possible object of a category, forms a kind of perception which is able to see things that have not been seen before yet delivers "meaningful" output. Again, this might seem trivial for a cognitive scientist, but is not for a computer vision expert. Here, almost all research is dealing with "real objects", with "observation of the world as it is". However, cognitive robotics will most likely be in need of concepts that go far beyond the observed materialistic world.

To a certain degree, abstraction is related to what has become known as "categorisation". We are cautiously using this term as it is quite overloaded with a specific bottom-up view in computer vision. In cognitive psychology, however, the information processing approach to categorisation goes as follows (as described by [PS01, p.379]): After forming a structural description of the object that needs to be classified, existing category representations *similar* to this structure are searched and the most similar one is selected. Finally, *inferences* are drawn about the entity and the categorisation information is stored. This might lead to a concept like "sit-able". We postpone the discussion on the different categorisation theories in psychology to our later section on concepts (4.6.3).

Regarding information processing theory, our notion of *abstraction* would mainly fall into the first step of the chain, namely the forming of a structural description from the observation (with structure denoting the "constituting substrate", i.e., not necessarily only spatial information). Concerning the theories on concept formation, we will later adopt a sort-of prototype theory, although our use of a computational ontology will force us to use the *method* of classical theory, i.e., an all-or-nothing decision when categorising.

Figure 1.5 in the Introduction showed the main idea of abstraction and its implications for classifying instances. A first idea would be that a robot, confronted with the various images shown in the figure, should be able to detect "every" arch. This is in fact the task of "generic object recognition" in computer vision with respect to the appearance (form, structure, interest points) of the object category alone. For the field of cognitive robotics, however, we immediately see that the question digs deeper. The seemingly clear goal gets difficult when the burning question becomes: What constitutes an arch as being an arch. Only the form? Is it its function? Or maybe its size? When do we humans "see" an arch? This issue will be discussed next; for now, it is important to realise that any abstraction mechanism needs to take these "constituting substrates" into account in order to allow for task-dependent and situated categorisation. Considering only the *form* (structure) of the arches in the figure, they all have some kind of combination of "columns" and "top-bar" – and in this way are related to the structural schema that we put in the middle of the figure.

4.5.5 Generalisation

Generalisation is the next cognitive function that we want to address in connection to visual perception. With this notion we refer to the ability to *form* object concepts from actual instances. Hence, it is also concerned with connecting conceptual descriptions with actual instances, just the other way round. There is much literature in psychology regarding human concept learning, which we will partly review in Section 4.6. For the *process* of generalising from instances, we adhere to Aristotle's approach. Referring to his connection to the traditions of rationalism and empiricism and also the later Piagetian theory, Russell describes it as follows [Rus99, p.256]:

> For Aristotle, rationalism and empiricism (as they came to be called) are each mistaken. [...] He denied that there could be knowledge that did not presuppose certain forms of perceptual experience. Empiricism, on the other hand, assumes it is possible for a creature *with no knowledge at all* to acquire some knowledge, despite the fact that the very idea of knowledge acquisition would seem to imply the possession of prior knowledge to get the process of the ground. Aristotle's attempt to resolve this dilemma can be viewed, according to David Hamlyn [...], as a kind of "genetic epistemology" in the sense that it attempted the kind of account that Piaget attempted, while not itself being a recognizably Piagetian theory.
>
> What is innate for Aristotle is not mental structure but a potential (*dunamis*) for acquiring knowledge. [...] He proposed the mechanism of *epagoge* (meaning literally to "lead on" and normally translated as "induction"). By the process of *epagoge* the mind comes to sort particular instances into general principles aided by the fact that some experiences retain their character through the flux of sensation.

If we take on this view, it would mean that before the robot companion can actually work for humans, it would need to acquire those concepts of objects, processes and own duties by sorting out irrelevant "accidentals" (to stick to Aristotle's terms) and only keeping the ουσια, the essential substance. This would need to be done by a combination of observation, experimentation and guided learning – much as how a child learns (in Piagetian terms by changing his/her schemas through assimilation and accommodation). In other words, for having the robot working in its niche properly, it would be in need for some kind of learning mechanism that detects salient features (not only in the visual channel(s)!) from instances, abstracts this in order to build more general knowledge (we could call it "things-in-themselves") which helps for classifying and drives abstraction when confronted with instances later on. We will *not* tackle this learning issue, although we are convinced that this is one of the most exciting questions involved in the discussion of binding higher level cognitive functions to low level perceptual input. Partly we omit this because it is unclear how to do this in a truly generic way, and partly because for our focus on visual perception, we would fall back to a purely structural and appearance-based description. Hence, we will circumvent this learning in the practical part of the thesis by pre-giving a kind of high-level structural description and concentrate how actual images can be abstracted so to detect instances of those higher-level concepts.

4.5.6 Symbol Binding

We will now tackle the last cognitive function that we consider central for visual perception of a robot companion. In short, symbol binding is concerned with connecting observations with situated knowledge of the robot. We argued before that one limitation for using classical computer vision approaches for a robot companion is due to the fact that the system does not "know" or "understand" what it deals with – vision is often seen stand-alone and detached. Instead, we propose a view which is in line with situated and embodied approaches and in which current visual perception is judged to be necessarily tightly bound to context and task. In the following we tackle this question less from a learning point of view (how the symbols get "grounded" or "bound"), but rather more from a conceptual point of view which asks what "the meaning of an object" denotes for the system.

Connectionist (or "subsymbolic") accounts could (to a certain extent) well explain how the agent is grounded in its environment (due to direct correlations of sensor input and motor output). Symbolic computing, however, has the doubtful reputation of actually opening up the problem of "symbol grounding". This has been of central interest from the beginnings of artificial intelligence on. A reference work of symbolism by one of its most famous proponents, Jerry Fodor, is [Fod98] in which he explains that "[...] the meaning of concepts is fixed atomistically and it is fixed 'informationally' by causal inputs from the environment" [Rus99, p.266]. Movements against symbolism have focused on the fact that a totally disembodied symbol crunching machine (as in "GOFAI"[19]) cannot explain fully how meaning gets "inside the head" and have – not

[19] GOFAI means "Good old fashioned artificial intelligence" and is nowadays widely used to refer to the symbolic approach to intelligence.

least because of this fact – led to a more grounded understanding what semantics are about by focusing on embodiment and situatedness [Cla01, PS01]. Sometimes, especially in the aftermath of connectionism, symbols as such have totally been dismissed and focus has been laid on direct sensorimotor-pattern interpretation.

We would plead for a view which agrees firstly on the existence of symbols for reasons that refer to the complexity of tasks and to the evolutionary necessity, as we will explain shortly, and secondly on the substitution of "symbol grounding" with "symbol binding", denoting that the meaning of concepts can and do change due to task, context, accumulated knowledge and so on. Related to this view is Sloman's approach [Slo02] which denotes meaning as being widely determined by *structural relations between concepts*, with the additional *help* from sensory information to reduce uncertainty. He calls it "symbol tethering", and underlines that symbol grounding is just another word for concept empiricism (which already Kant refuted, but still keeps persistent within the computer science community). Additionally, he stresses that different species have either lots of concepts from birth (precocial animals), where others have less (altricial animals) and links this with differing evolutionary demands [SC05]. In any case, the meanings of concepts are not "built up" by observing instances, but rather are shaped in the interplay of innate structures (concepts) and empirical observations.

We will now delineate, why we adhere to symbolic computation approach at all. The approach to the phenomenon "cognition" has – in our opinion quite correctly – been shifted in the last years from *pure* symbol manipulation to more extensive views involving a dynamic, situated and embodied perception and action complex that aims at surviving in and adapting to its environment. However, in our view this does not imply that symbols are outdated as such – underneath the overall conception of functional fit, subprocesses involving symbol manipulation might still be in place.

Of course, Brooks' radical critique on representation [Bro91] does have its point. Although it seems a bit exaggerated[20], it is true that pure symbol manipulation is not *necessarily* what robotic intelligence needs to be about – it might be true that instead of "GOFAI's symbol crunching", aspects like embodiment (and situatedness) are better suited for some research projects. However, if we take it radically and *exclude* explicit representation, then we will – at least to a certain extent – always have the limit of a purely reactive agent. This is fine, as long as we are dealing with systems whose task is, e.g., to collect soda cans [Con90] – i.e., behaviour-based robots. But if we are aiming at a far more complex goal (e.g., including reasoning about possible objects and situations or weighing possible contextual situations in order to decide where to search for a specific item, not to speak of human-robot interaction via a "symbolic language"), symbolic representation is inevitable. With other words, it is true that we might not need representation on every layer – but at some (higher) point, we do (we will discuss this layering issue in Section 4.6.4).

This idea of having an explicit form of representation for some capabilities is also related to considerations on abstract thinking, of course. From developmental psychology it has become clear that we cannot even see human development in a purely correlative manner – due to [Kei99] abstract terms are used correctly by children at an age in

[20] "Representation has been the central issue in artificial intelligence work over the last 15 years only because it has provided an interface between otherwise isolated modules and conference papers."

which formerly they were judged to represent the world only in a concrete manner. He puts it in a nutshell with the phrase [ibid., p.170 f]:

> Whether it be in computer science, linguistics, or psychology, there is a recurrent need for a system that not only tabulates the frequencies of properties and their intercorrelations but also represents rules, principles, and mechanisms. When a person taxonomizes animals, or vehicles, or tools, that person uses a mixture of principled reasons for sorting categories as well as brute force comparisons of some feature sets.

This indicates once again that not *one* "way of thinking" is the right one, but we need to be open for accepting different paradigms within one system. Recall the "recognition" of the object at hand that we used in one of the vision chapters (Chapter 3). There, it was, too, a brute-force combination of all known interest points in order to estimate the correct pose of the object in the current image. On *this* layer, a purely numeric measurement was very well suited. Likewise, in our texture approach in Chapter 2, we detected the border of the object under inspection by applying statistics. In both cases, such an "in-the-code" representation of properties of the object makes sense. However, to use the object in a broader context, to exploit additional information about it, and to expand the knowledge we have about it, ontological rules and symbolic relation of parts are better suited.

Moreover, there seems to have been an evolutionary need for the use of symbols in addition to (simpler) sensorimotor contingencies, cp. [SC05, Slo07], and, as [Slo01] argues, that it seems that evolution has discovered at some point (quite late, however) that a lot of environmental demands can best be handled by using abstract symbols and reasoning over them. So despite the fact that in the course of AI-history, symbolic computation was questionably thought of as being the *only* way of intelligent thinking, we might be able to use symbols for a layer where they can still be considered to play an important role.

There might remain the question on using connectionist accounts. The point is that we can *not* assume that some natural clustering from purely low-level network modelling would give us the same advantages. Here we also touch upon the discussion on correlation vs. causality – with respect to language, [Rus99, p.268] points out:

> Then there is the question of what exactly is being modeled in the connectionist language work. Fodor (1997), I think rightly, points out that what the networks are doing is modeling *correlates* of grammatical classes, not the classes themselves. For example, being preceded by the words "a" and "the" and preceding words that often refer to actions and that takes "s" and "ed" suffixies *correlates* with being a noun. But for a net to pick up on these regulaties is not that same as its representing the class "noun" simply because those features are not definitional of that category but are, rather, statistical signals of it in English.

This argument nicely shows that subsymbolic networks do not have the possibility to switch to another "mode of thinking". Additionally, this also touches the philosophical question on necessary truths that living intelligent agents make use of (Humean correlation is qualitatively distinct from Kantian causation). By the way, the Qualitative Spatial Reasoning community, too, emphasises the necessity to bridge from quantitative to qualitative information [CH01]: "Spatial reasoning, in our every day interaction with the physical world, in most cases is driven by qualitative abstractions rather than complete a priori quantitative knowledge."

This implies for us that the "symbols" will need to be abstract enough to allow for further high-level (symbolic) computing and low-level enough to allow for a tethering to concrete task-affordances and instances of objects in the environment.

Coming back to our focus on a robot companion, we might say – as a first sketch – that symbol binding in conjunction with perception denotes the capability of the robot to know in every step what the objects detected *mean.* As we saw before, meaning is subjectively bound not only to the perceiving subject, but also to the current task and situation the robot is in. Therefore, the "attachment" of some meaning needs to be done situated from the perspective of the robot – a strictly constructivist approach would even *only* allow for such a bottom-up and strong subjective account. However, as our objective is a robot companion, there is need for some kind of representation of meaning in terms of the human operator as well – cp. the dependencies in Figure 4.1 and the related discussion in Section 4.3.4. Returning to our example, this means that the "coffee mug" is a simple object that has some affordance from the perspective of the robot, such as that it can be grasped, pushed, pulled, crashed and the like. An intelligent robot companion, though, needs to additionally know what it means for a human, i.e., that it can be used to carry liquid in it which humans need to drink in order to survive (cp. Figure 4.3). A complex task like "Bring me some water!" will involve such an object – an object that from the perspective of the robot would maybe not carry this information in the first place. One could argue that the fact that it is bowl-like should be enough for the robot to choose it as water-carrier, so that the additional knowledge that humans prefer to drink out of such an object is needless – but this misses the point. A truly understanding robot needs to be able to (at least in a minimal sense) take the *perspective* of the human – this is the challenge. It is our conviction that this change in perspective, this linking of different meanings referring to the same object or object concept can best be done explicitly with a symbolic approach.

Let us reconsider Figure 1.5 in the Introduction in which we saw different arches and asked what the constituting substrate is that makes each of them an arch. We can now see that this, again, points to task-dependency: Finding everything that has the "structure" of an arch is a different task than an assignment such as "go underneath the arch" would require. For the latter example, arches on by-passers' T-shirts depicting Guardi's Capriccio (top right in the Figure) or Prague's bridge (bottom left) should not be true positives. This would ignore the *meaning* of arch in the current *task context.* Consider the case for humans: Maybe, looking at the bridge (bottom left) we would not see an arch, but rather a bridge with a tower. Not before we think about the "function" of the hole in the tower, we might come clear about the fact that it constitutes an arch. Thus, despite the fact that we humans seem to work with a consistent concept "arch" so easily, it is not totally clear what the underlying substrate (philosophically: the thing-in-itself) of the arch looks like. Probably – and this is the reason why we postponed

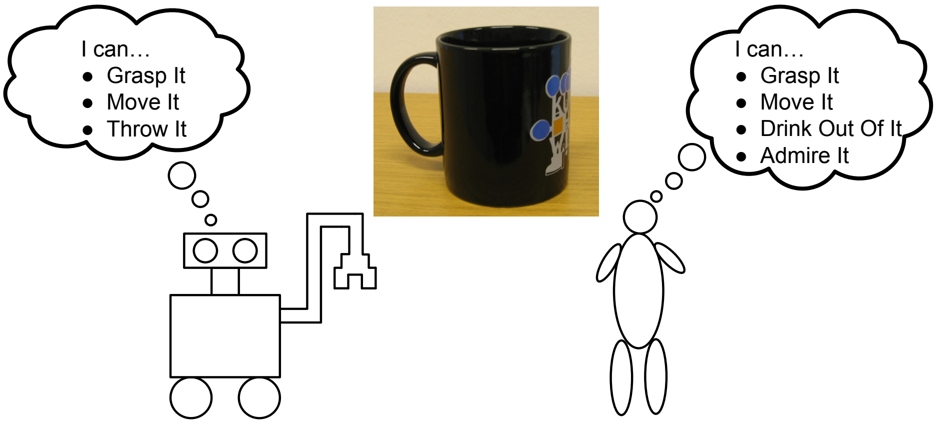


Figure 4.3: Differences in *meaning* due to different world-views.

this discussion from Section "Abstraction" to here – this is all related to the *meaning* of arch (in a given situation); with other words, to what a given percept "binds". So additionally to the previously mentioned meaning from the human's perspective, we need to be able to bind different meanings to an object according to different contexts.

Some possible (interconnected) factors or reasons why an arch is an arch are listed in Figure 1.5 as well. Some are more generic than others (e.g., structure vs. colour), and it is very likely the interplay of different "activations" that let us humans "see" an arch in each of these images. This points to the crucial question on how to represent all this information in a flexible way - in a way that allows connecting different tasks, contexts and object concept descriptions for truly generic perception that is needed in autonomous cognitive systems in a home environment. One possibility is to use a logical, interconnected and easily expandable database (which we will actually do in practical Chapter 5). For humans, however, there are most likely quite different "modes of thought" that can be switched depending on the current situation. Another possibility is that there could be "bits and pieces" that are arranged depending on task and setting.

4.6 Glueing Functions Together: The Need for an Ontology

Every now and then we said something about representation, about concepts and about the linkage between the different cognitive functions proposed for computer vision in cognitive robotics. We will now bring these bits together and approach the question on how the functions are connected and what the overall ontology ("knowledge base") of the agent needs to be capable of in order to achieve these competencies. Therefore, we will discuss the notion of "concept", especially of "object concept" and propose a "layering view" for the processing of visual information. All in all, this will give a foretaste and rationale for the ontology used in the last chapter of this thesis.

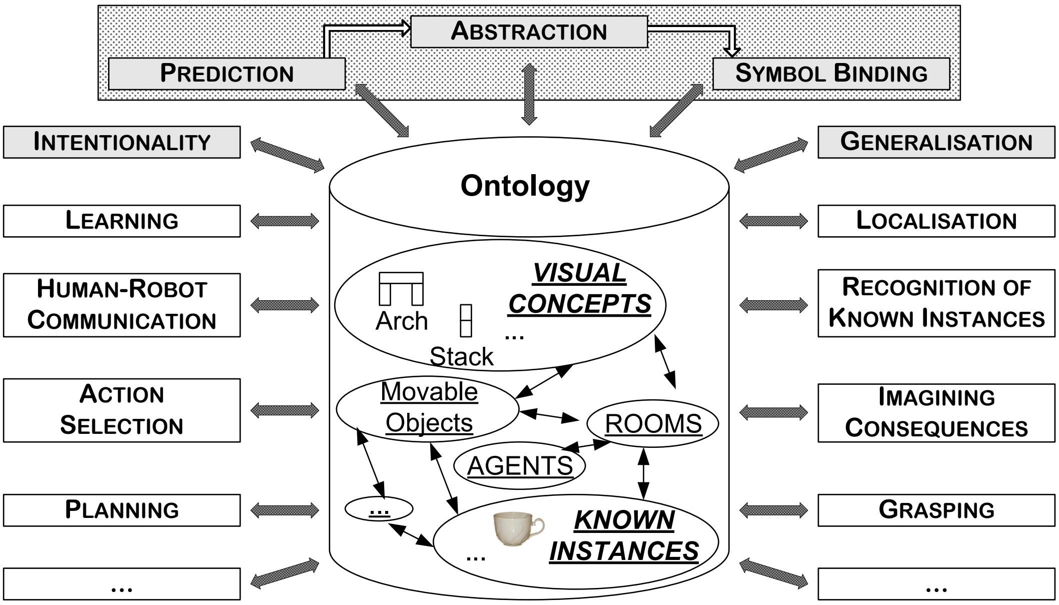


Figure 4.4: Cognitive functions working with the same ontology: the subset of *visual* functions that we focus on in this work is marked in grey with special focus on those concerned with visual abstraction (on top). The ontology itself holds diverse concepts (some examples are shown), that are interconnected in various ways.

4.6.1 Connecting the Functions

In the last section we delineated our view on functions needed for situated perception of a cognitive robot companion. Those functions should, of course, not be separated from each other but should rather be tightly interconnected to other visual and non-visual functions. Recall our list from the Introduction, in which we itemised several requirements for a robot companion, serving the most general goals of autonomously understanding and acting at the user's side in a home environment. We propose that it is imperative that all these miscellaneous functions should work on a concerted shared ontology holding rich knowledge. Only this way, situated perception and action, especially the integration of vision as part of cognition, can be guaranteed.

Figure 4.4 shows our understanding of the interplay between the different cognitive functions. Those that we explicitly tackled in the last sections as being necessary for situated perception are highlighted and those three of them that are especially concerned with perception are arranged on the top of the Figure. *Intentionality* is somehow the precondition for them as it gives task- and thus context-information. *Generalisation* actually "fills" the ontology for all the remaining vision functions with abstract descriptions of some object concepts. As said, we will not deal with this specific sort of learning in this thesis; underlying is our belief that without knowing how to *use* an ontology, it is hard to design mechanisms how to *build* it up. Rather, we will focus on the cognitive function of *abstraction* as this is the function which is most tightly bound to the question of the combination of the high-level ontology and perceptual data and thus takes a prominent role for our focus. Perceptual abstraction will especially be connected with the questions of *prediction* (which delivers expectations on what is in the image) and *symbol binding* (how to give abstracted information meaning). This connection is underlined by the two respective arrows; of course, these are not the only direct connections between functions. Furthermore, we put other (more concrete) visual functionalities like "Recognition of Known Instances" or "Localisation" in the Figure in order to indicate that there are other, more concrete functions concerned with visual input as well. They are, however, located on a different layer insofar as they are direct functionalities, not cognitive functions. This means that prediction, for example, would "trigger" the recognition module if there is need for an instance-detection. Abstraction, on the contrary, is more generic and deals with making sense of input that has not been observed before. Hence, our chosen subset of five functions is a selection for a specific kind of generic, situated perception.

The use of such buzzwords as "planning", "action selection" or "grasping" hints at the fact that there are a lot of functions that a future robot companion will need to implement, yet that are themselves hard scientific topics. As we do not intend to provide a full-blown cognitive theory here, these are somehow "placeholders" for likely cognitive functions and concrete functionalities that should be provided. There will be some more (especially those that do not deal with perceptual information in the first place), as indicated by the three dots on either side of the Figure at the bottom.

All functions are connected with one and the same *ontology* which needs to provide the data all functions work with. We can now understand why this ontology needs to hold a variety of different information which is "adequate" for an understanding of "what there is". Particularly with respect to our goal, it must contain information of some sort which...

- ...*intentionality* can use as guidance with respect to the task followed and for knowing what the next steps require,
- ...*prediction* can use to "hypothesise" what is likely to be seen in the next image,
- ...*abstraction* can use to find abstract descriptions in concrete instances,
- ...*semantic binding* can use to retrieve situational meaning – with other words, here perceptual data (e.g., proto-objects) become symbols (objects), and
- ...*generalisation* can use to build up these abstract descriptions.

Pre-announcing, abstraction points to a specific content of the ontology, namely object concepts which hold abstract descriptions of what object categories look like. This is shown by the elliptical (super-)category "Visual Concepts" inside the ontology. Exemplary other concept groups are "Rooms", "Agents", or "Known Instances", the latter being needed by classical object recognition techniques (in the Figure: "Recognition of Known Instances"). There could be more high-level categories such as "Movable Objects". This already indicates one big requirement for the ontological representation: There must be the possibility to represent various types of relations (such as that an agent can *use* a cup, and a cup *is-a* movable object; additionally a cup *can-be-located* in a specific room).

Needless to say that the information-*types* can be quite different for the respective functions. For example, whereas prediction, in our understanding, needs rather "perception-near" data, such as how bits and pieces are connected or even what those bits' appearance is, for example, a traffic light will force attention to red and green circular shaped blobs, other functions would need information about scenes, maybe even about the human operator's preferences. Therefore, planning the next action or weighing options, will mostly use non-visual information.

Let us get back to our focus on perceptual functions, but deliberately take an example of the auditory channel:

For instance, the interplay of the cognitive functions suggested would look like the following in the case of the auditory channel: Having the task to recognise a symphony, the ontology provides necessary information for *intentionality*, i.e., about the focus of attention to hearing, maybe the bodily movement towards the speakers (= rather non-auditory information), then *prediction* would supply information about the attention towards the next phrase, maybe activating some kind of harmonic understanding (this is sort-of auditory information). After *abstracting* the heard chimes (because of different set of instruments playing the tones, or a different key in which it is performed), *semantic binding* would try to match the heard sequence with known representations in the ontology (both rather auditory information) and retrieve the information about the composer and name of the piece, so that *higher-level cognition* (again task-bound) can be activated to, e.g., give the answer or write it down (non-auditory information).

We saw that the type of information needed for the different functions could and will be quite distinct. This begs the question why to use *one* knowledge repository and not several and even more how consistency can be retained, why and how to represent data explicitly and not "in the code". The latter is what is usually done in computer vision. E.g., highly specialised techniques are able to "re-cognise" objects by means of feature points. Often, the information about what is recognised is stored somewhere implicit in the code or in high-dimensional "feature vectors" whose similarity to a given image patch is calculated by a distance measure[21]. Without explicitly integrating this kind of information in an ontology where it is bound to other information (what the object's name is, for what it can be used, in which scenes it might appear), however, this impedes the versatile usage of this valuable information for other parts of an overall goal. E.g., predicting what to see, connecting one object's information with another's for coming up with a scene description, the retrieval of information about the function of an object, or the hypothesis formation about what the occurrence of one object could mean for the task at hand – all this can much easier be used, integrated and extended if we find a common formalism to represent information. Again, inspiration might be drawn from a different discipline, i.e., neurophysiology, in which it became clear that there is integration from various sorts of "neuronal clusters" inside one's brain, be it for cross-modal integration or for top-down/bottom-up integration [KGB07]. There is no strict separation between "bottom-up" image processing and "top-down" (ungrounded) knowledge.

All this makes clear that there are a lot of questions on what the content, the con-

[21] Recall that this is what we have done for the appearance-based object tracking approach described in Chapter 3.

struction and the usage of the ontology needs to look like. In the following, we will try to give an overview of considerations regarding concepts, concept formation and categorisation on the one hand and the notion of ontology on the other. These topics are related, as we can view the totality of concepts as constituting one's "personal ontology". Categorisation or concept formation is then the means of arranging observations in a manner that they are consistent with or even enlarging our ontology. Likewise, we can refer to "ontology" philosophically as the science of "what there is" and consequently define the relevant concepts that are "out there" (i.e., the "world ontology"). Thus, we will review related work both with respect to concepts and concept formation and pertaining to ontology in one section. Again, we will have a look on philosophy and psychology alike.

4.6.2 On *Ontology* and *Concepts*

Most generally, "ontology" is a notion that stems from philosophy and denotes the branch of metaphysics that is concerned with questions on what "is", what the nature of being is and what existence means. The shortest description would be that ontology is the study "of what there is". We are not reviewing the long-lasting history of philosophical inquiry of ontology and shorten it to the fact that ontological questions include not only the nature of being, but also whether existence can be categorised, whether there are "universals" (i.e., the problem about the existence of abstract entities such as "mankind"), or whether and how we can distinguish between *essential* properties of objects as opposed to *accidental* ones.

We have earlier made the distinction between *personal ontology* and *world ontology* whereby the former is actually not what philosophy is concerned with when talking about "ontology". Rather, this would be what *epistemology* deals with. However, the word ontology has been widely introduced to denote any kind of knowledge repository, e.g., the "ontology of an agent", so we will also use this term in order to refer to *all the knowledge that an agent can use*. Furthermore, in our conception, *ontology* is the collective term for all the *concepts* that the agent uses to deal with its environment. These concepts comprise static information as well as process information ("theories") about the world and its constituents.

In philosophy, concepts have been discussed from various points of view. [ML06] give the following suggestion for classifying the discussion points: the *ontology* of concepts, the *structure* of concepts, *empiricism and nativism* about concepts, concepts and *natural language*, and *concepts and conceptual analysis*. For our concern, the ontology of concepts is for obvious reasons our main focus. The authors stress that the basic assumption taken in the discussion about concepts as psychological entities is the adoption of the "representational theory of mind", which takes a structural approach, meaning that concepts can be composed of simpler representational entities. Early philosophers such as Locke and Hume can be allied with this view such as their "ideas" are taken to be mental images and representational entities consequently. Modern philosophers do not think that all representations are images; however, the representational theory as such is still in discussion. We already met one of the most famous proponents, Jerry Fodor, whose "language of thought hypothesis" [Fod75] states the working of the mind language-like, i.e., that all thought can be mapped to a symbol-like structure.

We will not further elaborate on the decades-lasting discussion on whether a representational theory as such does make sense or not, but rather point to the fact that it is still not clear on how these representations would be "implemented" in a neural net such as the brain. As we pointed out, we are not interested in elongating the discussion about how human conceptual understanding works but rather in how we can get interesting clues that point to a possible usage in a cognitive robot, thus, we will just accept the representational view. For our goal of situational detection of object categories in images, we need a special kind of "concepts", which we call *object concept* and present next.

4.6.3 Object Concepts

With object concept, we refer to a description how objects (existing, materialistic ones, not philosophical objects) "look like" on an abstract level. Therefore, we are interested in what each occurrence of this object class has in common with the other instances. This is a tricky question and has been discussed in philosophy mainly as "thing-in-itself", though, as we will see in the following, not necessarily on a materialistic level. We are aware that the reduction of the discussion to really existing objects is a simplification that we take in order to come to a grip on what to use for visual perception in a machine (we have first presented these ideas in [SVFB07]).

We will now have a closer look on what has been said about concepts in general and what this implies about object concepts in particular. Interestingly, we can actually go back as far as to Plato (427-347 BC), whose theory of "ideas" (spread across various dialogues) actually gave a first account on the fact that humans know something that goes beyond perceived instances. His view was that we saw those *ideas* (eternal prototypes) in heaven before we are born and we can recall them in a sort-of unconscious way when confronted with an instance (a mere "shadow") that partakes in this idea. However, Plato did not actually refer to materialistic objects when talking about ideas, but rather meant "entities" such as mathematical or ethical "Forms" (a synonym for "ideas") [Pla04][22]. The philosophical work of his pupil Aristotle (384-322 BC) is a bit more interesting for us, as he purged the Platonic position from metaphysical touches insofar, as the conception of ideas became a system of *categories* with which we classify sense-data. He started from ουσια, the substance (e.g., being human) and then went further on to quantity, quality, relation, place, time, position, state, action and affection[23]. Additionally, by defining that, e.g., all humans have the same "human-ness", he introduced the notion of "accidentals", meaning that besides the same essence, other properties form a specific instance (e.g., Socrates). Hence, an arch is an arch because of its taking part in the ουσια of arch-ness. Additional accidentals might be that it is red, high, or dirty. However, ουσια still keeps being a quite metaphysical term as it is defined as the "actual being" [Ari99] or "what is underlying in something", a typical

[22] As with any Platonic dialogue, it is not entirely clear which parts of the content stem from Plato himself and which from his teacher Socrates.

[23] The first four categories are the most important. The last four differ in Categories [Ari06] from the ones in Metaphysics [Ari99].

example mentioned by Aristotle is the *soul* in humans [Ari06].

Performing a two-millennia-jump, our next stop is Immanuel Kant (1724-1804), who did not only point to the fact that the perception of the world around us is biased by the perceiving subject (an early account of situatedness and constructivism, if you will), but who also first widely introduced the notion of the "thing-in-itself" that is now very close to our understanding of an object concept. With the term thing-in-itself, Kant described those actual existing categories of the world that we can *not* directly perceive; and as the thing-in-itself is inherently not ascertainable, we cannot know its *existence* either. We can only guess and assume it when perceiving the "things-for-us", which are the actual phenomena. Our (subjective) reality is consequently made up of appearances that come from but are not the same as the things-in-themselves. The latter only affect our senses [Kan99]. For a modelling approach such as computer science is heading for, the fact that object concepts are inherently *not* tangible is disappointing, of course. Let us have a look at a successor, whom we encountered already when analysing the quotation at the beginning of this chapter: Nicolai Hartmann (1882-1950). What is most important in our context of object concepts is his conception that we are able to perceive important traits of the thing-in-itself and not only of the thing-for-us. By this he means that during the established "being relation" of percept and perceiver, we notice certain constituting aspects of the objects and by that we are able to categorise them. Additionally, for Hartmann this is the reason why we are able to talk about things that we did not actually perceive. We would argue that this further implies some kind of abstraction – an abstraction towards those important substantial properties.

With this understanding of object concepts, we need to tackle the question on how these are formed, i.e., how they arise in the cognizer's mind. In the spirit of this thesis, although we are not tackling learning that much, this is important as it will influence how an artificial system needs to acquire concepts needed for intelligent behaviour. Philosophically, we could – in principle – start with a review of all the work on empiricism vs. nativism, which we will skip as this is not our major concern. As a very short summary, [Bor99] notes that the discussion goes back as far as to Plato's *Laws*, and reviews shortly the main historical line including Locke's ("tabula rasa") and James' empiricist positions in contrast to Kant's and Descartes' rationalist views. It suffices to say that the same discussions have been carried out in psychology under the term nature/nurture debate. As usual, recently most theorists claim that both positions have their right – there should be some innate stuff that guides development [Spe98]. With other words, the question breaks down to what is the necessary innate *structure* which guides the acquisition of *content* during developmental phases.

According to [Fel03], humans induce the simplest category consistent with a given set of example objects when learning concepts and categories from examples. In a more recent paper [Fel06], the author underlines the plausibility of a representation by means of *qualitative structure*, where complex patterns can be broken down into simpler (more atomic) patterns, having fewer features, i.e., lower degree. The author models this kind of compositionality and shows in a mathematical formalism humans' bias towards low degree of complexity. However, the author is aware that the reduction to Boolean concepts does not account for the whole capability of human concept learning but underlines the importance of qualitative structure.

We need to mention the famous PhD-thesis of Patrick Winston [Win70] as we are using arches as examples, too. He also used this paramount example, but was mainly concerned with the problem of *learning* descriptions of complex objects. The usage of conceptual representations in terms of databases, expert systems or ontologies (in the engineer's sense) will be tackled later. A more recent work that covers concept learning is [Vie06]. Her focus is on necessary design features of the architecture in order to allow for acquiring concepts by an autonomous agent, specifically with respect to affordance concepts.

A very important issue is the discussion of how concepts are represented in the human's memory, which we can maybe adopt as general paradigm on how to handle categories (concepts) of objects. In psychology, there are three major views on the assumed representation of the concepts in one's memory. According to [PS01, p.380 f], we can distinguish:

- Classical theory: An object is judged to be a member of a group if and only if it possesses a set of defining features that this category constitutes. This way, the decision is all-or-nothing. E.g., an arch is an arch if and only if it possesses exactly two columns and one top-bar.
- Prototype theory: Going back to Rosch's work [Ros73, Ros77], people store prototypes of each category ("arch-ness") with which new percepts are compared. Class-membership is therefore fuzzier as it is only a matter of weaker/stronger affiliation to that category. This prototype is not a representation of a real existing thing but incorporates all properties that are typical for the respective object class. Additionally, there is a preferred level of abstraction of *most basic categories* which "[...] carry the most information, possess the highest category cue validity, and are, thus, the most differentiated from one another" [RMG$^+$04].
- Exemplar theory: According to this theory, the cognizer stores representations of typical examples that are grouped by category and again, a similarity measure decides the class-membership of a new percept. This implies that there is no "additional" abstract category knowledge. An arch is thus an arch if it is similar enough to an arch that has already been seen before.

It is interesting, that often you can only find the latter two as possible approaches to abstraction in psychological literature (e.g., [Lef06, p.202]), which seems plausible as the first one is rather the extreme all-or-nothing form of prototype theory. For computer science, however, as long as we adhere to logical rules when doing classification, this classical theory is somehow the *method* of how to decide which observation falls into which category. It is also notable, that according to [ibid.], again, there is empirical evidence that the truth lies somewhere in the middle: humans probably store both very abstract prototypes and specific examples in order to represent concepts. This is related to what [Kei99, p.170] writes: "The study of concepts in adults has repeatedly come across a need for two kinds of architectures [...] considered as associationist versus symbolic representation. [...] concepts as mirrors of probabilistic patterns in the world and concepts as mirrors of causal, logical, and mathematical patterns (the 'concepts as theories') view. While both themes can point to supporting phenomena, neither seems sufficient on its own for handling how we use most adult concepts."

4.6.4 The Question of the Right Layer

As has become clear until here is that there is need for an ontology with which all the functions can work. What can we, as designers of artefacts which show cognition, learn about the discussion on ontology, concepts and especially object concepts just presented? Recall that an object concept in our understanding denotes a static (for the time being) description of what is the constituting substrate of a category. Therefore, if we stick to a purely spatial description, our prime example "arch" could be defined as having two columns and a top-bar. Another possibility would be to define arch as something that stays on the ground-plane and is a "hole" with "something around" (a surely quite fuzzy description). This section is concerned with questions on what actually goes into the ontology: the lines that make up the columns, the arch as such or parts of the arch (column, top-bar)?

First of all, as shown in [Fel06], we can take away the necessity of a prototypical concept representation (as opposed to the exemplar one). This difference to typical computer vision re-cognition work (such as the learning of hundreds of images of an object so to detect further similar ones) is due to our orientation towards an autonomously acting robot companion showing situated perception. The next question concerns the "layer" that is chosen to represent explicit information bits.

We have already indicated that we propose to use object concepts as the right layer to start implementing the ontology of the agent in order to achieve visual abstraction. The ontology needs to be capable of representing various levels of detail, i.e., "lower-level" features (such as lines and points) as well as "higher-level" features, such as scene descriptions. The following considerations aim at explaining our view on how the agent's "layers of cognition" might look like – with an emphasis on perception again. The general conception is quite close to what have become known in psychology as "dual processing theory" of the mind [Eva03, Eva08]. This theory describes human cognitive processing as a two-way system of reasoning, one being evolutionary old and concerned with both innate input modules and domain-specific knowledge ("unconscious processing"), the other evolutionary recent and allows for abstract reasoning and hypothetical thinking. Both are necessary for human cognition, and their distinctive properties include the facts that the first one is fast (reactive) and the second one is constrained by working memory capacity, thus not always applicable.

In the following, we are using this conceptual view, but try an interpretation what this implies for the perceptual reasoning capabilities of an artificial agent. We are approaching this issue by postulating four assertions, each being based on the former ones.

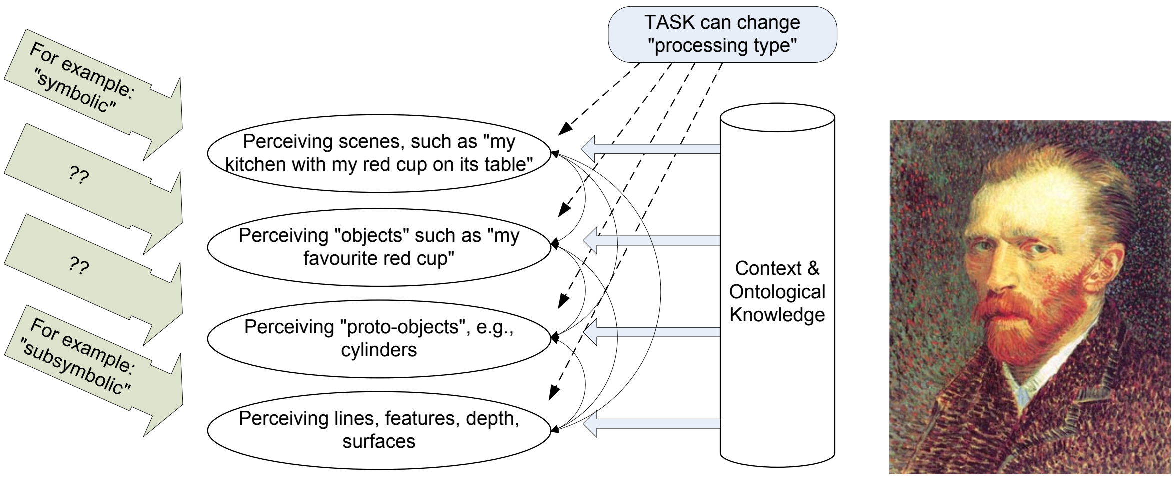
Assertion 1: There are multiple layers of "cognition".
With this claim we want to express that for any kind of cognitive mechanism (function), there might be several layers – not only in a hierarchical sense. They range from simple actions like relating two atomic features to each other in a causal sense (like in perceptual grouping), to more complex ones like thinking about what to buy as Christmas

present for my little cousin knowing that she still believes in Christ Child. The first one is mainly done unconsciously, the second consciously. However, the complexity of the features involved, do not causally determine whether consciousness is involved or not – or what the "grade of consciousness" involved is. Think of tasks as encountered in psychological testing: Here, you need to concentrate on very simple features (lines, points) and find out how they relate to each other – a conscious process.

Assertion 2: Depending on situation and task, there are different "ways of thinking" taking place.
Fast, reactive behaviour is done unconsciously in a way that resembles "knowledge in the code" as used in computer vision tools. The system executes specific procedures or a set of condition-action rules without "knowing why". Thinking in functional terms, as we do, it does not matter which parts are implemented in a subsymbolic (maybe "emergent") manner, and which in explicit symbolic terms. However, this is what we could *rather* classify or *explain* as "subsymbolic" processing. It is, for example, the arrangement of features that give rise to the perception of a Gestalt (such as in perceptual grouping) – in humans this is also done mostly unconscious, somehow "learnt" and "ingrained" in one's vision system. Abstract thinking, on the other hand, is consciously done and maybe thus better be *envisaged* as being reasoning on a symbolic layer. This difference in processing can also be called "ways of thinking" (in the style of [Min06]).

Assertion 3: Humans are able to switch from one way of thinking to the other if the situation demands this.
This is a crucial point. We propose that there is not *one* possibility to process and represent data – according to "layer" or complexity. Rather there are many different ways of thinking that we can use according to what is the task at hand. It may be that there is a preferred means of processing for a specific task and there may be even a task that can *only* be performed in one way of thinking, but in general, humans are able to switch. This should be possible for an intelligent robot companion as well. Especially this implies the last and most important claim:

Assertion 4: Therefore, items having been perceived "unconsciously" (e.g., in a subsymbolic or any probabilistic manner) might get conscious (symbols) for further computation if necessary.
Being able to change the mode of processing would enable the agent to "think in different layers". There are definitely a lot of cognitive tasks that need not be done in a way that can easily be explained using symbols. For example, our everyday perception of coherent object movement might really be simply done ingrained and unconsciously. Driving a car for the first time will demand a high degree of attention with a lot of influence from textbook-sentences ("symbols"), whereas after having years of experience a lot of previous worrisome percepts will just be handled without "noticing" – it *became* ingrained. For our hypothetical robot, this would mean that the perception of a proto-object could have been done Gestalt-driven; however, this Gestalt can later be taken as symbol and further analysed to consist of this or that edge, etc. The ontology of the agent needs thus to be able to account for that and to instantiate symbols on a lower level as well.


For example: "symbolic"
??
??
For example: "subsymbolic"
TASK can change "processing type"
Perceiving scenes, such as "my kitchen with my red cup on its table"
Perceiving "objects" such as "my favourite red cup"
Perceiving "proto-objects", e.g., cylinders
Perceiving lines, features, depth, surfaces
Context & Ontological Knowledge

(a) Layers of visual processing.

(b) Van Gogh's Self Portrait from 1887.

Figure 4.5: The preferred "way of processing" (subsymbolic, symbolic) might be switched due to the task. All layers are influenced by current knowledge and context.

Figure 4.5(a) tries to sketch this conception of dual-processing for perceptual questions. With the help of a simple example we will explain this view a bit more. First of all, having our definition of cognition in mind as being the superordinate principle that governs "intelligent behaviour" (cp. Section 4.4), it is safe to say that on each of these layers, cognition takes place. Looking at Van Gogh's famous self-portrait from 1887 (Figure 4.5(b)[24]), we might first only see a man with a red beard. If we do not know that this is Van Gogh, we might indeed only have the "symbol" man with red beard generated in our head, which is the totality of the brush strokes of the image. We claim that not each stroke got a symbol by itself but rather that the entirety generated the symbol that can further be used for higher-level thoughts about the last Van Gogh exhibition or plans for future cultural trips. The point here is that the "low-level processing" – which enables us to see the whole – works without even noticing.

However, getting to see more details of the picture, we might "instantiate" one or more brushstrokes as symbols in the sense that they get items of high-level thoughts: We might wonder about the colour transition from collar to jacket or the directions of strokes on the skin, which enables thoughts on the era of post-impressionism, etc. What happened was the transition of "entities" that have previously remained in the unconscious processing part (strokes) now got our attention and led to more detailed understanding. Note that the functioning of the "symbol man with the beard" remains intact. With other words, we switched the details into our symbolic processing without (necessarily) disturbing the whole[25]. Needless to say that the "amount of symbols generated" is further dependent on the individual's history: Being an expert in the history

[24] Source: `http://de.wikipedia.org` (12/2008). Public Domain image; the original is currently at the Art Institute in Chicago, USA.

[25] A side remark: This non-disturbance is not necessarily the case. There are situations where existing higher-level knowledge primes our further visual experience, such as the famous dalmatian hidden in a seemingly random black-and-white cloud.

of art for sure influences what is usually seen as "way of perceiving" – which we could call here "dissemination of activation throughout the layers". Without wanting to further address the question of consciousness, we can, however, see that all this involves a pragmatic explanation for the *need* of consciousness: Something gets consciously in our minds whenever an arrangement becomes a symbol that needs to be worked with on an abstract layer. At this point, this symbol can be used for (often language-based) reasoning and more generally "thoughts about that thing". This is for sure a simplification as we must be cautious to equalise subsymbolic processing with unconscious thoughts. However, it is much more probable to figure explicit thoughts in terms of symbols than of neural activations.

With this view in mind, it is imperative that the ontology needs to be able to instantiate "any" object and object part (object not necessarily in the materialistic sense) and link it to other objects. This link needs to be flexible enough to describe relations like "is part of" or "entails" as well as more fuzzy ones like "is one possibility of" or "might contain".

So far the theoretical considerations, what about the requirements with respect to *vision*? Here, again, we need to have possibilities to establish various kinds of detail-levels (such as "face" or "brushstroke"), however, this could be quite a lot and it might not be necessary. We would rather argue to *start* on an *object concept* level, this means to describe parts of objects or objects that could cover more than specific instances. The transition from low-level features to higher-level reasoning is best done at this layer. This way, the ontology contains information which is "low-level visual" enough to allow for matching of sensory input to it and "high-level cognitive" enough to constitute the link to symbolic AI, including planning, reasoning, retrieving of higher-level information etc. Our paradigmatic example will be "arch", which we encountered already in Figure 1.5. We will see that, of course, also the building parts (column, top-bar) are object concepts that are represented in the database, but we will not go further down and split those object concepts to lines, corners or specific features. In any case, once we are able to bind observed perceptual data to such a structure in the ontology ("concept"), we have established a *symbol* which is open to any computation well-known to symbolic AI.

Coming back to our need for storing information that an artificial agent needs in order to act purposefully on the human's side, we can now see that the ontology needs to have the capability to be enriched by various sorts of links between parts, wholes, scenes and back to the feature level. This is quite a requirement and we will see that there are tools that do in principle support such a diversity of demands. What we start with to store are the *object concepts* and they in turn – here we have come full circle to our philosophical starting point – reflect some sort of Hartmann's notion of the *thing-in-itself*: The constituting and common substrate that defines an arch as being an arch can be perceived as important traits of the thing-in-itself arch by abstracting from the thing-for-us, i.e., the actual instance. In pure structural terms, it is the presence of two pillars and a top-bar. This is for sure only a quite basal example, but it allows us to point to all the main issues.

One of the core issues is the question on the transition from low-level pixel data to high-level concepts. Managing to do this also means to be able to decouple image processing from higher-level cognition insofar that there is still a causal relation between it, yet we need not code every possible instantiation in the vision software – imperative for versatile behaviour. Although we often refer to "reasoning" in connection with dealing with the concepts stored in the ontology, we do not imply that all reasoning takes place as correlating concepts in the ontology. However, for situated perception it might be the starting point to connect to non-visual cognitive functions – as broadly discussed in this chapter. At some point we need to have such a symbolic processing, otherwise vision keeps being occupied with ungrounded bits – once these bits are connected by a common concept, we are able to "tether" more information to it.

4.7 From Theory to Practice

For our practical considerations, we need to find a suitable representation for the object concepts on the one hand, and have to tackle the transition from vision output to the object concepts on the other hand. This way we will find a possibility to transfer the theory presented in this chapter to a first practical implementation. In the next chapter we will do this by showing that abstract descriptions of object concepts can be linked to "low-level" vision tools.

This section of our theory chapter is thus first concerned with possible data representations useful for our purpose and examines the notion "ontology" in computer science (Section 4.7.1). Then, we discuss how abstract concepts and concrete perceptual data go together (Section 4.7.2). The question on how much needs to be pre-defined by the designer (Section 4.7.3) concludes these last theoretical considerations.

4.7.1 "Ontology" in Computer Science

Let us start with trying to think about which tools might be relevant and useful for the implementation of our theories to practice. The goal is to find a representation scheme that allows us to implement our philosophical theory, which we delineated extensively in this chapter, on ontology and epistemology, on the functional approach deploying intentionality, prediction, abstraction, generalisation, and symbol binding, as well as on concepts and "ontology" as such. First of all, due to our considerations on realist and evolutionary epistemological views we may assume an ontological basis of "what is out there" in order to understand the epistemological requirements. This means that we need some kind of "knowledge storage" which further allows for the mentioned capabilities. We can see that *any* kind of information storage that an agent is equipped with is often called its "ontology", however, this fact stems from its philosophical roots and needs further specification for our concrete task.

Research on concepts in artificial intelligence has mainly focused on questions of concept *formation* due to the wish to be able to model how this capability might work in humans. Exceptions are expert systems that make extensive use of explicit forms of knowledge *representation*. Especially in the last years, the notion of ontology became en vogue also in artificial intelligence due to the semantic web movement in which an explicit and human-like language representation of concepts had been needed.

Interestingly, these "ontology tools" in computer science also use the term "concept" to refer to single entities that are in relation to other entities. Particularly, in the engineering sense the reference definition of "ontology" is given by Tom Gruber [Gru93]:

> An *ontology* is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an ontology is a systematic account of Existence. For knowledge-based systems, what "exists" is exactly that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational *vocabulary* with which a knowledge-based program represents knowledge. Thus, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names are meant to denote, and formal axioms that constrain the interpretation and well-formed use of these terms.

As can be seen from this definition, a tool that is capable of doing these things would be perfectly suited to implement our theoretical considerations. It is notable that an ontology must "make sense" as it defines the vocabulary the system can work with. This means, that it must be firstly grounded in a sense that it provides everything that "evolution" requires (cp. evolutionary epistemological approaches, such as [Vol81]) and secondly, it must be tuned to the cognizer's "lifeworld" – a term introduced by the philosopher Husserl [Hus54, Hus64] and adopted for artificial intelligence by Agre et al. [AH97a]: "A lifeworld, then, is not just a physical environment, but the patterned ways in which a physical environment is functionally meaningful within some activity." For the case of a robot companion, useful refers, of course, to a shared interaction area between human and machine.

As first shot, one might think that these requirements can simply be coded in a common relational database which accumulates knowledge in tables. However, we soon find that it gives us no possibility to store the data following strict structural constraint, e.g., for allowing automatic consistency checking. With other words, the university of discourse might be inconsistent. If we look further, we see that more recently, techniques have evolved that try to overcome this problem, such as the *vocabulary*, the *taxonomy* and the *thesaurus*. A very good and concise description can be found in [PU03], which we use to give the following digest: Whereas a controlled vocabulary does not include any semantic content (it is just a set of terms), taxonomies and thesauri specify meaning by using relations between the terms. The former use strict hierarchical links only, the latter allow for additional (non-hierarchical) connections. The word ontology, finally, has been used for all of these flavours, but nowadays the strict use of "ontology" refers to an implementation of a *thesaurus* in a *dedicated ontology language* (such as the "Web Ontology Language", *OWL*). In fact, nowadays there are dedicated ontology languages and corresponding tools, which we will use in Chapter 5, where we will also review work that has already been done using such a structured account.

A computational ontology in this latter sense follows a strict formal structure which can be exploited by applying reasoning techniques. We are interested in using this kind of knowledge representation for our purpose. After all, this section and the next chapter are not least interested in whether a formal representation of information in an ontology language is suited for the situated perception in robots we are seeking, i.e., for representing qualitative information that vision can work with and which helps implementing the cognitive functions discussed in this chapter.

The main advantages of using an ontology (such as *OWL*) can be listed as follows (some of these points will become more obvious in our examples in Chapter 5):

1. Automatic reasoning: Reasoners can do consistency checking and deduce implicit knowledge.

2. Rules: Related to the first point, is the possibility to check the occurrence of specific arrangements of individuals by applying rules.

3. Human language definitions: The use of natural language helps during development and testing.

4. Storing semantic knowledge: The free choice of the *content* (as opposed to the *structure*) of the ontology allows storing, e.g., affordance-based, function-based, or task-related information along with the concepts.

5. Abstraction and interconnection of concepts: Using the relations mentioned, not only hierarchical information (*is-a*) but any kind of interrelation can be described (e.g., *can-be-located-in*, *likes*,...)

6. Separating knowledge from computing: With the help of ontologies, we are able to separate "common sense knowledge" (although, however, tuned towards our application) from vision processing (up to a certain degree).

7. Speed: Last but not least, ontology reasoners are very fast and scale well. They are highly specialised to deal with a large number of concepts.

Of course, this implies that we are inherently following a "symbolic computation" approach (indicated by terms like "explicit", "describable relationships", "vocabulary", "set of representational terms"), as became also clear in our conception about a layered view on cognition in Section 4.6.4. We defended its general necessity in Section 4.5.6 and will tackle the question of representing abstract concepts next.

4.7.2 Abstract Concepts

So far it has become clear that we would like to have a means to store "object concepts" in an ontology of the agent that can be used for situated perception (by applying the functions described earlier). This "object concept" layer describes objects on the basis of their constituting substrate and – for cognitive robotics in general – these concepts can be linked to various other information valuable for the system. Another central term besides “object concept” will be *abstraction*. This is one of the cognitive functions that we worked out earlier and it is a function that can be applied on a visual level. We will see how abstraction can help to get from visual input to object knowledge. “Conceptual” or “abstract” knowledge is quite a tricky term and we draw the inspiration for describing useful knowledge on that layer (as opposed to concrete instances) again from other disciplines: “We don’t remember or recall things with complete fidelity – not because the cortex and its neurons are sloppy of error-prone but because the brain remembers the important relationships in the world, independent of the details” [HB04, p.75]. Vision is usually, especially in works on object recognition very detail-near. We need to get more general (at least for robotics) and this can (full circle to our introductory remark on multi-dimensionality) only be done with integrating task knowledge, context, and so on. This implies that our ontology will not only hold *object concepts* in order to perform visual abstraction (as indicated), but additionally other kinds of abstract knowledge with which they can connect.

Using abstract concepts, we might not be able to change the fact that an engineer’s bias is always present (resulting in all the difficulties that we sketched in the triangle of robot, human and world of Figure 4.1), but we can try to specify a representation means for the system that is much more powerful, flexible and adaptable. The main rationale is that the fact that even if there is a bias, this need not necessarily imply that there is not a layer on which we can have a representation of things “out there” that separates between the *meaning to the human* and the *meaning to the robot*. With other words, the storage of additional – not necessarily vision-bound – information is central when trying to model cognitive capabilities of the overall system.

Here we can again cite the developmental psychologist Keil that gives an (at first sight) unintuitive account on why to use abstract (and not concrete) knowledge might not be totally misplaced [Kei99, p.178 f]: “In many cases it seems that development might best proceed in the opposite direction, namely, from the abstract to the concrete. Such cases seem especially vivid when we think about explanatory knowledge. [...] If infants do indeed have an intuitive folk physics and psychology, the forms of those theories must surely be in abstract terms, waiting for experience to fill in the details.” We can take this quite surprising stance and carry it to the extreme if we assume – similar to Kant – that there are abstract categories that we make use of, just as Chomsky states: “The true theory according to Chomsky involves underlying abstract categories that cannot be learned, induced, or constructed on the basis of experience and must therefore be innately specified in the mind itself” [Nel99, p.187]. For now, we need to argue for the pre-definition of this qualitative and abstract knowledge in order to focus on perceptual issues in the next (vision-bound) chapter.

4.7.3 Innate Specification

The abstract object concepts that we defined to be a good means to connect visual perception via abstraction to a symbolic computation can either be acquired by the system itself (catchword generalisation) or pre-defined by the designer of the system. A constructivist “learning” approach is indubitable a plausible, flexible, and extendible way of acquiring concepts by the system, and in general, this is the way to go in the long run. Up to a certain point, however, the predefinition of concepts is inevitable and as we are not focusing on learning methods in this work, we will pre-define quite a lot in the next chapter (namely, for example, how the object concept of an arch looks like by using pre-given spatial relations). We insist on the fact that this does *not* mean that we defend and follow a nativist approach as such.

To explain this point, we can stress the – now classical – example of face recognition: [Joh99, p.153] describes the main approaches[26]:

> One of the topics in which there has been considerable debate between nativist and empiricist perspectives concerns face recognition abilities in young infants. On the empiricist side, [...] it was assumed that face processing was an acquired skill developed through prolonged exposure to faces. [...]
>
> On the nativist side, two lines of evidence are commonly invoked to support the idea of an "innate cortical module" for face processing. The first is that several studies have shown evidence that newborn human infants preferentially respond to facelike patterns. [...] The second source of evidence comes from recent functional neuroimaging studies, which show that particular regions of the cortex, such as the "fusiform gyrus face area," are specifically activated following the presentation of faces [...]
>
> By considering evidence from biology, Morton and Johnson (1991) presented an alternative, constructivist account of the development of face processing in which they argued that there are at least two interacting brain systems in operation. One of these systems [...] underlies the tendency for newborn infants to orient towards faces, while the other system was hypothesized to acquire information about faces through exposure to them. Importantly, the first system was argued to bias the input (preferentially toward faces) to the second system, thus ensuring that it specialized for learning about faces.

Having our paramount example of detecting an arbitrary arch in mind, we can formulate (admittedly exaggerated): In order to acquire the ability to detect an arch, a cognitive robot can follow an empiricist, a nativist or a constructivist account: On the empiricist side, the system might be confronted with a large number of arches and is then able to redetect everything that looks similar to what has been seen earlier[27]. The nativist conception would stand in for a pre-definition of what arches look like – this is actually our starting point. However, the important constructivist amelioration to this approach is to define the *relations* as well as the detection of the *parts* used as innately specified and to allow the system to deduce how specifically sized parts or proto-objects can be related using this innate knowledge and instantiate a new concept

[26] The cited paper herein is [MJ91].

[27] This is the machine learning approach to object recognition. Those systems usually rely on a huge database of pre-learnt instances or a so-called "bag" or "codebook of features". Needless to say that scaling is a problem as these approaches usually compare the image at hand with *all* objects in their database – every object having many features, each consisting of, e.g., 128 dimensions (as in the case for SIFT [Low04]).

if this is something useful and sometimes encountered. The difference to usual feature extraction would be that these relations are kinds of *qualitative* information used for the description rather than quantitative feature vectors. This is very tightly related to what Kant called a "synthetic a priori concept" (with a priori, however, not implying to be innate, cp. [Slo08b]) and where the difference of this kind of causality lies as opposed to Humean correlation. For this latter (constructivist) account, the vision system looks out for *relations of parts*, using already learnt concepts and rules ("first system") and links specific relations found to concepts ("second system"). This latter capacity could be called "generalisation".

In any case, for a robot companion, the object concepts have to be related to what the user needs the robot for. This is one of the main reasons why we for now tackle the abstraction of visual input. Besides that, it is obvious that there might be some object instances that the robot has seen before and maybe even uses often. In this case, it would of course be better suited and more plausible that it actually learnt an *instance representation* (still, however, tightly bound to additional semantic information). We will see in the following chapter how we can account for that by also storing in the ontology this kind of knowledge about how the robot knows about things. This exploits a big advantage of a technical ontology, namely that any kind of relation might be defined and that we thus are even able to use already learnt instances in an *intentionality-guided* manner.

4.8 Wrap-Up: Ontology for Situated Perception

In this chapter we have described theoretical work in the field of cognitive robotics with a specific focus on visual perception. The goal is to clarify how computer vision techniques can be adapted for robot companions. There are a lot of considerations necessary for robot companion research, which have been overlooked due to the focus on specific solutions instead of on general problems. Building on the general assumption from the Introduction that robot companions are a subset of cognitive robots that themselves are a subset of cognitive systems, we have located and defended four mainstays in the course of argumentation:

1. *Anthropomorphism* must be tackled very carefully and instead of searching for "biologically plausible" models we should focus on inspirations how specific well-tuned functions interact in cognitive perception.

2. A *rationalist* and *constructivist* approach is most suited (and anyhow inevitable) when working in a "synthesising science" such as robotics.

3. *Ontological* ("the world as it is") and *epistemological* ("the world as we represent it") questions are indivisible for a constructed robot that should deal in its proper niche: next to the human.

4. The *functional* approach allows us to start with considerations on cognition of a robot without needing to tackle implementation issues.

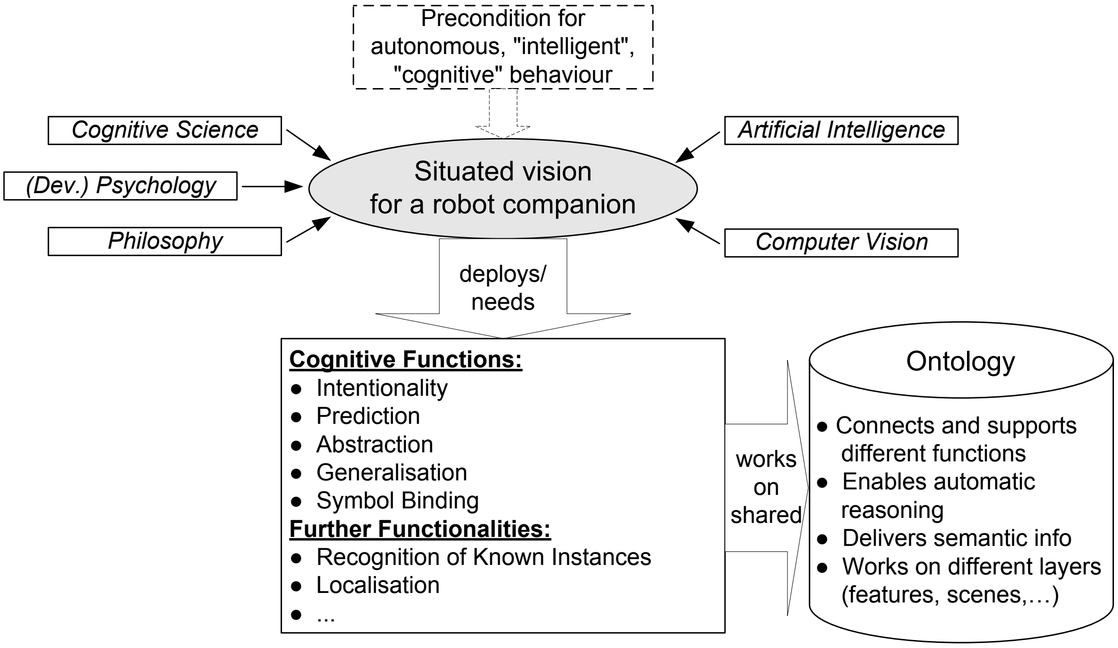


Figure 4.6: Summary of approach: Disciplines, functions and ontology.

Besides that, we have clarified our understanding of “consciousness”, “intelligence” and “cognition” and tried to sketch a pragmatic explanation. We have then elaborated further on our emphasis on the functional layer and located five functions that are “near enough” to be of interest for situated vision of a robot companion and “high enough” to enable the stressed necessity of connecting to non-visual and higher-level cognition:

- *Intentionality* (as task-directedness)
- *Prediction* (as anticipation of what will be perceived)
- *Abstraction* (as decoupling of the accidentals from the foundational substrate)
- *Symbol binding* (for giving an observation a meaning *for the system*)
- *Generalisation* (for generating abstract description from perceived examples)

As the glue to hold these and other cognitive functions together and consequently as “interface” between them, we have defended the use of an *ontology*, first in a general sense, and later on with specific focus on the means to represent *object concepts* that we have located in what philosophy has called “things-in-themselves”. In connection to this, we have outlined our view on different modes of processing that a cognitive system should include.

On a meta-level we can sketch our approach as shown in Figure 4.6. As can be seen, we assume that situated vision is a necessary precondition for the autonomous behaviour of a robot companion that we aim to achieve (and which might be called “cognitive”) – recall the example tasks we listed in the Introduction, ranging from retrieving an everyday item from another room to helping with the shopping. We argued that an interdisciplinary targeted research, involving philosophy, developmental psychology, cognitive science, artificial intelligence, computer vision, and robotics is needed for this objective.

In the course of our analysis, we distilled five functions for situated perception and saw that also a lot of other capabilities can be found which are necessary, both vision-related and non-vision-related ones. Additionally, there will be need for some concrete capabilities, like the recognition of object instances that have been seen previously by using appearance information (features). This approach implies the need for a shared ontology with which all capabilities work and which now gives us directly the tasks for a practical implementation.

We already analysed in the last section that a State-of-the-Art ontology language from computer science is able to provide the requirements for our goal. Specifically, we need to develop an ontology which is able to connect and support the different functionalities, i.e., it needs to guide the bridging from quantitative pixel data to qualitative "symbols" which can further be used to compute additional information. Then, automatic reasoning for consistency checking and deduction of further knowledge should be integrated with these symbols generated as well as the querying for additional semantic information. Finally, we have the task to make sure that the system is able to work on different layers, in the sense depicted by Figure 4.5 – from classical computer vision techniques (recognition of objects) to dealing with abstracted knowledge (classifying an instance as being of a specific object concept) and situated reasoning.

Our understanding is correlative to our perception. We must seek to see.

La Lumière
Robert Delaunay
[Del57]

Chapter 5

Situated Robotic Vision – Implementation

In the previous chapter we brought together some important issues for the field of robot companion research and the role of perception herein in particular. In this chapter, we will now combine these theoretical thoughts with techniques from computer science and vision and show means how they can interact. Therefore, it is less theoretically but more practically oriented. This demands some notes on the ontology language and tools used as well as on the design of the ontology as such. One section will also concern suggestions how the step can be performed from low-level vision (partly even in a more primitive form than presented in Part I of this thesis) to a kind of representation that is needed for higher-level tasks – such as tackled in common problem fields in artificial intelligence: planning, reasoning, learning. We will deliberately start with very simple vision techniques and focus on this transition.

A comprehensive example will show how vision tools of different complexity can interact and be used according to the current situation the robot is in. The goal is to show that notably the non-visual information stored in the ontology can be exploited by perception-near techniques – deploying the "cognitive functions" presented earlier. Thus, this is the chapter which brings together all the different issues tackled in this thesis, ranging from the use of features such as interest points in computer vision via theoretical considerations on cognitive functions to a State-of-the-Art ontology. By the way, in this chapter, whenever we mention *ontology* we mean it in the computer science sense, i.e., in the strict sense that refers to ontology tools and languages, such as *RACER* [HM01] or *OWL* [Wor04]. The definition to which we point by that has already been given on p. 97.

This chapter is structured as follows: After giving a review of related work (Section 5.1), we will present the ontology language and tools that we use (Section 5.2). Afterwards, we start off with a simple example that shows how an arch can be detected without any knowledge of its actual appearance but by rather using the spatial relations between constituting parts as thing-in-itself of the concept of arch (Section 5.3), and finally present our vision of solving a complex task by using explicit non-visual information stored in an ontology along with different vision techniques (Section 5.4). The discussion (Section 5.5) addresses some typical objections to symbolic and logical computation in computer vision.

5.1 Related Work

This section is concerned with giving an overview of the State of the Art. First, we give a short overview on usual representation techniques in (classical) computer vision (Section 5.1.1). Second, we review more extensively work on ontologies in computer vision. We will not reflect on expert systems or the use of ontologies *as such* neither list high-level ontologies that have been developed for robotics (such as [CCPR02]), but rather focus on those specifically developed for visual applications and thus more related to perception (Section 5.1.2). Finally, we summarise the State of the Art and point out the novelty of our implementation (Section 5.1.3).

5.1.1 Representation Techniques in Computer Vision

Here, we will give an overview on how the data used by classical computer vision (i.e., in non-expert system situations) is usually represented.

The simplest kind of representing "how an object looks like" for further usage by a computer vision tool, is to use a quantitative "description" of an object, exactly the way we did in Chapter 2 in terms of a wire-frame model. There, a text file holds the coordinates (in a model-based reference frame) of the corner points of a simple geometric object (such as a tea box). The program uses this information for fitting found features (edges) to this model. The appearance-based tracking approach of Chapter 3 used a slightly more complex, yet still purely quantitative description of the handled object, namely Superquadric parameters which can be used both to infer the borders of the object and the coordinates of feature points found on the object's surface.

Although these model-based approaches also constitute a kind-of abstract description (e.g., as nothing is said about the object's texture or colour when using only the shape-information of the distance between corner points), it is – due to the quantitative data – highly *tuned* for the *specific task*, so that a further connection to non-visual capabilities is not straight-forward. The main drawbacks can be listed as follows:

1. There is no account for the fact that there are objects of the same *sort*, which "just" vary significantly in appearance (to which, e.g., same or similar actions can be applied).

2. There is no clearly structured approach to represent the knowledge (e.g., in a State-of-the-Art formalism), so there is no possibility to infer additional qualitative, semantic, functional or task-related information.

3. Related to this, there is no possibility to apply automatic reasoning.

4. In the worst case, the designer of the system is the only one who can adapt it due to a mixture of technical implementation details and *implicit* semantic information.

For specific applications and especially for subtasks, such as handling a specific object in a specific situation by visual servoing, a model-based representation is, of course, necessary and well suited. For a flexible and adaptive robot companion, however, we additionally need a more qualitative representation technique. More focus on qualitative instead of quantitative information is provided by semantic networks and graph-based approaches – both providing some of the advantages that we seek: e.g., extensibility and explicit representation. What is still missing, however, are automatic inferencing techniques like those used in expert systems, where "new" knowledge can be deduced due to already known information parts.

These possibilities are provided by modern ontologies, such as the *Frames* language. We have already given the formal definition of *ontology* in the computer science sense on p. 97. We will now review approaches in computer vision that exploit the mentioned advantages of ontologies (mostly using *Frames*).

5.1.2 Ontologies in computer vision

Concerning the ancestors of recent ontologies, [CL97] discuss the usage of expert systems for image understanding tasks on an abstract level, denoting that future image understanding systems should have two main components, one holding general purpose knowledge and the other being liable of acquiring domain-specific knowledge. They then provide a survey over expert systems that implement (parts of) this sort of distinction.

A now classical position paper is [NS96], in which the authors stress the advantages of formal knowledge representation for image interpretation. The separation of the (formal) classifier from actual image processing code allows for much richer descriptions by simultaneously providing automatic inferencing possibilities and consistency checks. They also mention a big drawback of stiff object classifiers, namely that this does not permit a hypothesise-and-test scenario which the authors judge as being indispensable for complex vision tasks. Referring to Kanade's model [Kan78] for image understanding as presented in 1978 (!), they stress that efficient image interpretation is concerned with both bottom-up (hypothesis formation from partial knowledge) and top-down processing (hypothesis verification). They defend the usage of an explicit knowledge database by mentioning that one must not forget the other advantages, such as precise conceptual definitions and well-defined semantics. They then provide first ideas to extend current (1996) systems with these capabilities. In short, they propose to extract the hypothesis formation from the logically still purely deductive engine.

[MNW99], too, demonstrate the advantages of using description logics for computer vision. Although by that time, the web ontology language (*OWL*) has not been specified, its foundation, description logic (DL), a particular subset of first-order predicate logic (PL-1), did exist and is discussed by the authors. They outline that the inferencing technique is of possible usefulness for computer vision, and the usage of DL provides the additional advantage of being decidable. Additionally, there are services such as subsumption check, consistency check, classification and abstraction. All in all, this provides provably correct and reusable software.

A kind of expert system shell for retrieving images from a database due to specific image features is given by [SDM02]. A description logic is used for semantic indexing, where syntax is already provided at the level of segmented regions under the assumption that complex objects are composed out of the basic ones. Reasoning services such as retrieval, classification and subsumption are used for exact and approximate matching. The system is meant to be a human-machine interface for retrieving images through queries (by example or sketch).

[Tho02] gives a survey on knowledge-based techniques for image processing and distinguishes two main types: the use of image *processing* libraries and the automation of image *understanding*. The extraction of semantics as main problem in artificial intelligence is tackled where knowledge representation is described as "[...] a set of syntactic and semantic conventions to describe a piece of knowledge". She then explains *production rules* (working in strict *modus ponendo ponens*, being fragmented knowledge and lacking efficiency in problem solving) as opposed to *Frames*: *Frames* refer to the prototype-theory of concept representation and are basically a set of attributes with each attribute having several slots describing it. The properties that can be described could be internal (size, age,...), structural (sub-parts,...), relational (above,...) or role-like ones (father, server,...). Finally, she also explains that *hybrid systems* might account for heuristics (coded as rules) and object representations (coded as frames), which allows for constructs such as using those rules as frames themselves.

[MTB03] introduce the use of a "visual concept ontology" (consisting of spatial, relational, colour and texture concepts) which constitutes a middle layer between high-level domain knowledge and low-level image processing. They emphasise the usefulness of *ontological commitment*, meaning that ontological engineering is useful for the cognitive vision community due to providing a shared reference. Classification itself is quite inflexible which is due to the logical structure of the knowledge base and the association of numerical descriptors to observed features. In their 2004 paper [MTB04], the same authors emphasise the big advantage of using high-level domain knowledge in order to be independent of the application as well as to be able to reuse it for other purposes. Their domain concepts are composed of subparts using spatial concepts without temporal information. With the additional use of the RCC-8 spatial calculi [RCC92] and context concepts, each concept is composed of and restricted to four descriptions. In [HT03], the authors take the same basic approach, using a visual concept ontology with numerical descriptors but use three dedicated knowledge bases, one to do semantic image interpretation, another for anchoring symbolic data to image processing and the last one for intelligent image processing. They experiment with natural images which aggravates the description as no simple geometric model can be used. As ontology language, they use *Frames*, too. Again, they rely on quite perfect segmentation.

[MTH04] use the same approach for the purpose of image retrieval. They claim to solve the Symbol Grounding problem by linking numerical descriptions to visual concepts. They note that the hard part with an approach like this consists in the construction of the knowledge database, and to establish the link between visual processing and domain knowledge. The latter is facilitated by using expert knowledge only for the high-level domain knowledge and a machine learning approach for learning the mapping between visual features and this domain knowledge (by using numerical descriptors). A lot of preprocessing for this learning is, however, done manually (e.g., segmentation). The search when retrieving images is done in an ontological tree of depth 8, their complete ontology consists of 103 visual concepts (such as “granulated texture” or “circular surface”).

In order to overcome the domain-dependency of lots of usual computer vision approaches as pointed out by [CL97], and the total domain-independence of perceptual grouping tools, [ZTB04] try to combine the advantages of both for the domain of Thessalian graves images. This leads to a top-down guidance of segmentation in order to get from low-level image processing to semantic object knowledge.

Similar to our aim of thinking higher-level cognition and visual processing together, a classical work is [BBC93], who focus on the question of attention when it comes to interpret a scene. They, too, use naïve physics and try to explain arbitrarily complex stacked block structures with their system called “BUSTER”. Furthermore, they argue that the task of vision is not the construction of a scene model explaining the image data by finding rules of image formation, but rather that visual understanding is a matter of finding the “causal semantics”, i.e., the discovery of a causal explanation of the scene.

More recently, in a very relevant paper [NW03], the authors argue that although model-based scene interpretation is well-known, logic-based approaches are seldom encountered, which is why they investigate the explanation of a scene by modelling scene descriptions in a description logic. Their concepts comprise high-level descriptions such as *3D-body* or *polygonal-region* as well as more concrete ones such as *scene-saucer* or *view-A*. They distinguish between several “cognitive situations” which can be one of the following: context-free interpretation, exploiting spatial context, exploiting temporal context, exploiting domain context, exploiting focus of attention or intention-guided interpretation. These cognitive situations determine the repertoire of interpretation steps.

5.1.3 Summary: State of the Art

As can be seen, ontologies have already been introduced to computer vision tasks, such as *classification* in a narrow domain. Those systems mainly have a goal like the expert systems of the late Seventies and early Eighties of the 20^{th} century. Hence, most of the approaches found in the literature take the extraction of specific defining parts for granted. We, on the contrary, have a robot companion in mind, thus, we need to integrate classification and semantic binding into an overall representation framework, which we chose to be by means of such an ontological language, by simultaneously not relying on “perfect” visual preprocessing.

Furthermore, to our best knowledge, no system exists that uses the capabilities of the “web ontology language” (OWL), a description logic, or one of its similar flavours, such as $RACER$, which we will use as tool for the following implementation. The systems mentioned in the related work mostly use the *Frames-language*, which uses a different notation and supports less inferencing techniques. It is, in fact, a direct successor of expert systems' technique.

In the remainder of this chapter, we will tackle two specific novel approaches, one is the use of a State of the Art ontology for performing visual abstraction from deliberately elementary vision output (see Section 5.3), before we will try to integrate visual information into an ontology that is concerned with the overall robot companion and show a preliminary implementation (Section 5.4). The first goal, visual abstraction, aims of course at a similar goal as so-called "generic object detection" (for an extensive survey, see [Pin05]). However, the big difference is that we are less concerned with learning appearances in order to advance recognition rate of object classes, but rather try to bridge the gap to *qualitative* information. Such work will thus not be reviewed here. We are interested in representational issues, and will stick to ontological approaches. Note that this is due to our focus on the qualitative change to non-visual functions and semantic information binding.

5.2 Language and Tools

The ontology system that we are using for our showcase implementation is *RACER*, developed by the University of Hamburg [HM01]. It is either possible to use *OWL* [Wor04] in *RACER* or to use the generic formalism of *RACER* – both have the same degree of expressivity for us and we will thus discuss the capabilities of ontologies in a general manner[1]. For the sake of compactness, we will not give a complete overview on what can be represented using this formalism. A comprehensive overview can be found in [HKRS08]. [WNR$^+$06] give a direct comparison of *Frames* and *OWL*. As editor to build *OWL* Ontologies, *Protégé* by Stanford University of recommended. The Semantic Web framework *Jena*[2] was used to instantiate and query the ontology within our system.

Ontologies have means to represent classes, properties, individuals and – most important for us – relations between the entities. As outlined in [SWM04], there are some difficult issues that need to be considered when defining an ontology. One is the *level of representation* which will decide whether a class might be considered an instance of something else and the *subclass vs. instance* problem denoting that due to the context an entity might be a subclass, whereas in another it might be an instance. The prime example is an ontology of wines, where "CabernetSauvignonGrape" is considered an individual (instance) of "Grape" and not a subclass of it. The authors draw attention to the fact that this decision has to be driven by the intended application.

As outlined in the previous chapter, we are first aiming at using such an ontology to model *object concepts* that are of relevance in a robot companion scenario. With object concept we refer to a class of objects that can be represented first and foremost by their constituting substrate, i.e., what different instances of them have in common – this

[1] When talking about *OWL*, we refer strictly spoken to *OWL-DL*, which is one of three flavours of *OWL*. *OWL-DL* – with regard to its complexity – is settled between *OWL-Lite* and *OWL-Full*. *OWL-DL* is the maximal subset of *OWL-Full* which still remains to be *decidable* which is of uttermost importance for automatic reasoners.

[2] Jena is available at `http://jena.sourceforge.net` (12/2008).

might range from spatial descriptions to functional ones. The showcase implementation that we begin with is very basic and shall point to two things:

First, it shows *one* way of how abstraction (from quantitative appearance to qualitative description) can be performed, using a basic structural model. There are far more complex vision algorithms that are able to perform this specific task (the detection of an arch) much better; however, the advancement of recognition rate is not our goal. We are aiming at showing that with a structured qualitative description language we are able to use very basic vision techniques and still come up with (semantically meaningful) "re-cognition" of important concepts.

Second, this example explains how concepts can be described in a clear and expandable structure in a way that vision algorithms can work with them. Again, there have been various attempts in the past decades to use logic in computer vision and we do not claim that this work is new. The novelty is to show how our theoretical approach via cognitive functions and the *OWL*-approach can be combined with respect to robotics – first in the next section on a very basal level and further on, in Section 5.4, on a more comprehensive scale. This means that the kind of representation used in the next section will, in a more extensive ontology, only be *one* potential means. In fact, the main strength of an explicit ontology is exactly the possibility of designing knowledge on various abstraction levels and connecting these different descriptions, ranging from rather vision-near spatial relations for 2D images to rather vision-unrelated scene information. The complex showcase example of Section 5.4 will try to give this bigger picture afterwards.

5.3 A Vision Example Showing Abstraction

In this section we will present a very basic "textbook" example that shows how the step from visual information to conceptual knowledge *might* be performed. Note that we deliberately use a very simple vision technique in order to show that the transition from quantitative perceptual data to (logic-based) high-level information can be done by using this kind of abstraction technique (one of many possibilities).

Our goal for this simple showcase is to find an arch in a given image. To this end, we try to model the "constituting substrate" of an arch in the ontology. This work is thus two-fold: On the one hand, we need to model the general layout of the object concept ontology, and on the other hand, we need a suitable vision tool that extracts the parts. Additionally, we will say something about the framework that invokes the vision techniques and queries the ontology.

For this example, we decided to define the constituting substrate of an arch spatially, i.e., not function-based, but by relations of necessary parts. The simplest way is the claim that there need to be two columns and a top-bar, with the additional information that these parts have specific spatial relations to each other. However, as [KG05] make clear: "The Web Ontology Language has not been designed for representing spatial information, which is often required for applications such as Spatial Databases and Geographical Information Systems. As a consequence, many existing OWL ontologies have little success in encoding spatial information."

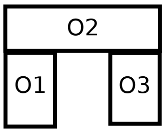
The suggestion of the authors is the integration of the RCC-8 calculus [RCC92, CBGG97] to reasoners and *OWL-DL* formalism. However, this would still not provide us with relations that are crucial for perception-near computing, such as "above" and "left adjacent"[3]. We investigated two different approaches to this problem. This is a design choice which influences how queries on the ontology will be performed.

Section 5.3.1 shows the implementation with a *restriction*-approach, afterwards Section 5.3.2 with a *rule*- approach. Then, we demonstrate the necessary vision preprocessing (Section 5.3.3) before we again split the explanation of the interaction between ontology and vision into the restriction-approach (Section 5.3.4) and the rule-approach (Section 5.3.5). We conclude with summarising drawbacks and advantages of the two ways of representing knowledge in the ontology (Section 5.3.6).

5.3.1 Ontology-Modelling with Restrictions

The first possibility to design and describe the necessary concepts is to define their "layout" through *restrictions*, as we did in [SV07]. However, here one of the main drawbacks of representing relations needed for vision tasks in an ontology language such as *OWL* becomes apparent. *OWL* builds on *RDF* (*Resource Description Framework*), which implies that information can only be stored in *triples*. This means that concepts are described with restrictions by using *one property* which relates *one individual of this concept* to *one individual of another concept* (i.e., only *binary* relations are allowed). In the case of the concept *Arch*, the description of its structure in triples could look like this (the sketch shows what O1, O2 and O3 is):

```
CompositeObject2D
hasParticipatingObjects exactly 3 Thing
hasSpatial2DRelationO1O2 some yOnTopAdjacent
hasSpatial2DRelationO1O3 some xRight
hasSpatial2DRelationO2O3 some yBelowAdjacent
```



This obviously quite difficult if not awkward description means the following:

- First row: *Arch* is a subclass of "CompositeObject2D", which is an arbitrarily chosen name denoting any object concept that is composed of some parts.
- Second row: *Arch* is built up of 3 (necessarily distinct) objects, in our case these will be two columns and one top-bar. In *OWL*, "Thing" means *any* concept in the ontology, i.e., one column of this arch might already be a column or an arch itself. This is advantageous as it allows arches composed of more than three parts as long as the roles taken are correct.
- Third to fifth row: These now define how the arch is built up from the three objects, with the object identifiers used shown in the sketch on the right. E.g.,

[3] The RCC-8 calculus discriminates the following relations between two regions: disconnected (DC), externally connected (EC), equal (EQ), partially overlapping (PO), tangential proper part (TPP), tangential proper part inverse (TPPi), non-tangential proper part (NTPP) and non-tangential proper part invers (NTPPi). Typical applications are located in the field of geography.

hasSpatialRelationO1O2 is an object property requiring that the arch needs to have at least one (keyword *some*) instance of object that is of the class *yOnTopAdjacent*.

As can be seen, in order to reason with the help of the ontology whether there is an arch in an image, we need to define *classes* in the ontology as strange as "yOnTopAdjacent" or "xRight". This is not only counterintuitive but actually also conflicting with ontology design principles, after which classes should denote (logical or real) objects or relationships, but not properties. Furthermore, classes are sets that contain individuals. In the definition of *Arch*, "hasSpatial2DRelationO1/O2/O3" denote so-called *object properties*. In fact, it would be much better to define a relation that says that an arch has 3 parts and to have concepts defined for them (*Column*, *TopBar*). Then we would only need to define relations needed, e.g., "hasRightColumn", "hasLeftColumn" and "hasTopBar" and would be able to really relate individuals to each other. However, this would imply that we know in advance which part in the image to be inserted into the ontology as individual "top-bar" or as individual "column". We do not know this (if we did, reasoning would be senseless). Additionally, we would have no information about how RightColumn and LeftColumn relate to each other (which we do in fact have in the example when we say "hasRight" as opposed to "hasRightAdjacent"), therefore, we chose to use such an unintuitive representation of "concepts" as reading from the ontology is easier later on for reasoning (see there). We know, however, that this way, we are actually *not* using the ontological representation in a way how it should be used, i.e., we are in fact giving away many of the advantages of using an ontology and fall back to any arbitrary explicit representation of information.

Another big problem with this approach is apparent: it implies that the description of an object concept gets quite complex the more relations are needed to specify it. The complexity is given by the binomial coefficient as this resembles a combination without repetition. Thus, an object concept consisting of, say, 10 objects, would need up to

$$\binom{n}{k} = \binom{10}{2} = \frac{10!}{2!(10-2)!} = 45 \tag{5.1}$$

relations to be described. This is the very definition of unwanted combinatorial explosion, especially when it comes to reasoning, which we will consider shortly.

To put it in a nutshell: In principle, it is possible to describe object concepts via spatial relations defined as such in the ontology. The overhead does, however, bear no relation to using a common relational database.

5.3.2 Ontology-Modelling with Rules

In order to circumvent these counterintuitive limitations imposed by the restrictions-approach, we decided to use the *rule-mechanism* that ontologies provide, as we did in [SV08b]. One advantage of using ontologies is the possibility to use the reasoner for searching the dataset in the ontology for compliance to a given rule. In the case of our arch example, a rule which is searching for the occurrence of an arch-like structure looks like this[4]:

```
1:   (firerule
2:           (and
3:             ($?c1 top) ($?c2 top) ($?t top)
4:             ($?c1 $?t UnstableSupports) ($?c2 $?t UnstableSupports)
5:             (neg ($?c1 $?c2 isPartOf)) (neg ($?c2 $?c1 isPartOf))
6:             (neg ($?c1 $?c2 isLeftOf)) (neg ($?c1 $?c2 isRightOf))
7:           )
8:      (instance (new-ind Arch $?t $?c1 $?c2) Arch)
9:   )
```

In the following we will explain the meaning of the different rows of this statement (a number indicates the corresponding line number):

`1:` The statement "firerule" indicates that this is a rule.

`2:` This rule has some preconditions which all need to be fulfilled (*and*).

`3:` First of all, there must be entities in the database that can take the place of the *variables* $c1$,$c2$ and t. Those entities are from the top-level class *top* (comparable to *OWL*'s *Thing*), so that again any object can serve as column or top-bar for this arch.

`4:` Here, the function of $c1$ and $c2$ in relation to t is defined, i.e., there must be a relation stored along with the instances in the ontology that states that the instance of $c1$ is "unstable supporting" the instance of t as is the instance of $c2$.

`5,6:` Some negations are stated in order to avoid problems with subsuming concepts.

`8:` The final line constitutes the "consequence" of the rule and indicates that a new instance of *Arch* is stored in the ontology with the name "Arch-t-$c1$-$c2$" (having t, $c1$ and $c2$ being replaced by the respective names of the variable-fillers).

The actual rule used in the examples will be slightly more complex, because we also labelled the arch-parts according to their role if the rule is fulfilled in order to speed up and simplify subsequent queries (i.e., whatever $c1$ really is, it then becomes instantiated as taking the *role* of a column *for* this new arch). What needs to be underlined is that in this case, there is no "direct" information about what constitutes an arch as arch stored in the class *Arch* as such – only "indirectly" as arches will be instantiated if this rule is firing, i.e., if an arch is detected in the image that complies with the rule's body.

In both cases shown (using a description through restrictions or using rules), as you can see, there is no quantitative description of arch in the sense of either appearance information or size parameters. This is what we refer to as "object concept". It is the purely qualitative description of what constitutes an arch.

[4] Different rule languages use different notations – the main principle stays, however, the same. In this case, the formalism for the rule language nRQL (*new Racer Query Language*) is given.

5.3.3 Vision Processing

The basic image processing tools used for extracting possible candidates that might take a role in our object concept arch is the same no matter whether using the restrictions-approach or the rule-approach. As said, the tools presented here are not meant to advance the State of the Art in visual processing – this is not the goal of this thesis. They are deliberately chosen to be quite simple and the focus shall be laid on the mapping from quantitative to qualitative data. Moreover, we are using two complementary modes of vision – namely a region-based segmentation algorithm and an edge-based Gestalt grouping tool – and show that both can help us for our aims. Of course, when it comes to using our conceptual approach on a real robot, there will possibly more vision techniques needed for different situations, lighting conditions and other parameters that we do not put our emphasis on here – which is on the step from proto-object to (semantically laden) object.

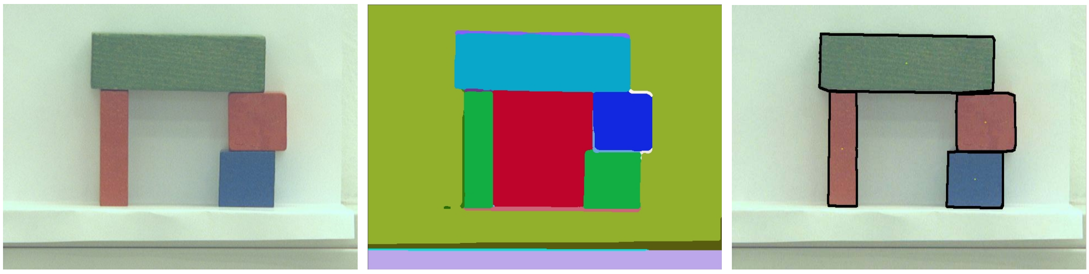
For "rather homogeneous" surfaces, the segmentation approach of [FH04] is well suited to extract parts of the image that are possible components of the sought-after object concept. We said earlier that generic segmentation "as such" is an ill-posed problem as a "correct" segmentation can only (if ever) be achieved when it is known what shall be in the image (cp. approaches that use shape priors, such as the Level Set method of [FDP06]). Using segmentation, we do not expect to find something specific, but treat these very simple generic "blobs" as *possible* parts of an object concept. With other words, this is the most basic understanding of "proto-object" (as a set of grouped perceptually similar pixel values).

The second tool we are using is *vs2*, which is a tool for finding Gestalts in images by grouping edges bottom-up. For now, we only use its capability to find "closures", i.e., closed polylines. This tool does not claim to find *objects*, but only "proto-objects" which denote here likely polylines that have a high degree of non-accidentalness. The rationale underlying this technique is *perceptual grouping*, an established theory (reaching back to the works of Wertheimer and Köhler) in perceptual psychology dealing with how human vision groups a scene into probable objects due to principles such as proximity, similarity or symmetry. A good introduction can be found in [Pal99]. For details with respect to *vs2*, please refer to [Zil07].

Our simplest example is shown in Figure 5.1, where the original image 5.1(a) is either processed using the segmentation algorithm 5.1(b) or the edge-based closure finding technique 5.1(c).

The proto-objects found will be input to the ontological reasoning, but before they can be used, some basic postprocessing is done. When using the edge-based approach, we need to "convert" the contours to surfaces of the objects. In order to compute the relations of the possibly needed parts, we do the following:

1. Double entries, i.e., *very* similar polygons with nearly identical outline that might occur due to local clutter are deleted.
2. Polygons are sorted, uniquely coloured and painted in a new image.
3. Colour erosion is performed using a colour voting scheme in order to remove very small particles.



(a) A toy arch.

(b) Segmentation with [FH04], resulting in 20 blobs.

(c) Detecting closures [Zil07], resulting in 4 proto-objects.

Figure 5.1: Vision preprocessing – getting possible object concept parts.

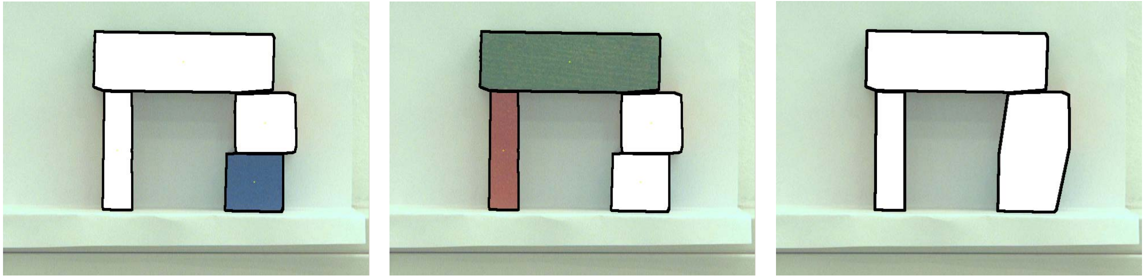
4. The hull points of all polygons are computed using a binary filtering approach.
5. An adjacency filter is applied which computes the relative position of the polygons to each other if they are adjacent[5].

The output of this procedure are the edges (more precisely: the minimal set of vertices) of the polygons as well as an adjacency map. For the rule-based approach, we additionally computed now the different kinds of "support" by applying a naïve physics approach. This is simply done by computing whether the centre of mass of the object further above is within the vertical limits of the lower object. If this is not the case, but it is adjacent, there is "instable support". Again, this is the most trivial approach; a more elaborate technique would involve three-dimensional information and probably even the integration of different viewpoints. Note that here again, a major advantage of separating explicit knowledge (e.g., of a relation named "support") from the vision processing becomes apparent: The underlying visual processing might be quite complex, yet the fact that there can be objects supporting others is a relation that just exists in any case. Once this information of objects and their relations is retrieved, it is input to the reasoner which we will describe next. As previously announced, the manner of representing the object concept arch in the ontology (restriction- or rule-based) determines how the interaction between the vision output (i.e., the part candidates) and the ontology looks like, therefore we split the explanation into these two approaches again.

5.3.4 Interaction Vision-Ontology with Restrictions

Using an ontology to describe the relations of object concept parts directly in terms of *restrictions* poses the following difficulty: With reference to the listing on page 112, we do not know what (if any) of the parts extracted from the image serves as *O1*, *O2* or *O3* of an arch. If we knew that, we would not need any reasoning as the function of their parts would determine their role (*O1* and *O3* are columns, *O2* a top-bar). As for this reason it is not possible to "label" the objects correctly and then ask the ontology whether they form an arch, with the restrictions-approach we can only

[5] Note that this, too, can only be done due to our assumption that we are dealing with robotics: In order to have a meaningful "is_left_of" relation, we need to assume an aligned camera.



(a) Direct query for Arch.

(b) Query for Stack.

(c) Using the stack as new object for the arch-query.

Figure 5.2: Restriction-approach with "closure detection" from Figure 5.1(c).

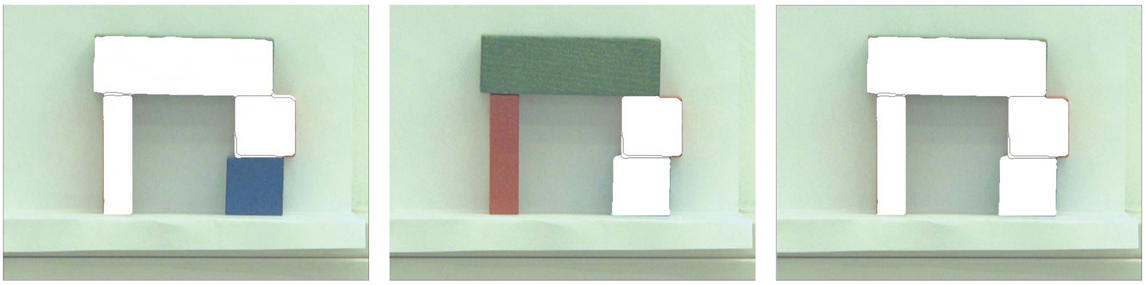
query the ontology what the definition of arch is and then do the reasoning *outside* the ontology. As can be seen, this includes an additional assumption namely that we already ask *for an arch*, i.e., we assume what is to be seen. With our argument from our theoretical chapter, this is not as bad as it might seem at first sight because due to our intentionality, we could predict to see an arch and consequently only look out for this concept. Still this reduces flexibility and error-handling.

With the restriction-approach, the reasoning would work as follows: Retrieve the definition of arch from the ontology and test *all possible combinations* of retrieved objects (from the image) whether a triple of them (only in case of the arch; even more for more complex concepts) complies with *all* constraints. If this is the case, instantiate a new arch with an appropriate name in the ontology. Results are shown in Figure 5.2. Needless to say, that if the number of parts extracted from the image is high, unmanageable combinatorial explosion is guaranteed. This holds especially for the case if we do not intelligently use additional constraints (e.g., only test "yOnTopAdjacent" constraints when checking column-top-bar relations. Here, the number of needed test scenarios is given as permutation without repetition. Even in the (perfectly vision preprocessed) scene of Figure 5.1(c) that only shows 4 detected object candidates and only 3 restrictions for defining the concept arch, we would hypothesise-and-test

$$(n)_k = \frac{n!}{(n-k)!} = \frac{4!}{(4-3)!} = 24 \tag{5.2}$$

times to find all possible combinations of the parts that might form an arch. Note the dependency on detected candidates. If segmentation is more fine-grained (and realistic), showing 20 blobs such as in Figure 5.1(b)[6], this number reaches 6840. The number gets even larger when subconcepts are found, e.g., when objects form together a new object stack, because then the number of objects to be related increases (this recurrence is needed for our desired abstraction). As said, additional information from vision preprocessing (via the adjacency map) prunes the search space. Nevertheless,

[6] Of course, we could "prune" the number of blobs and leave out very small ones, of which there are some in the Figure, reducing the number from 20 to 10 or less. However, this would imply that we make an assumption about how big the arch is supposed to be in the image, which is seldom a wise thing to do before any reasoning took place.



(a) One result of query for Arch. (b) One result of query for Stack. (c) Best result of query for Arch.

Figure 5.3: Rule-approach with segmentation from Figure 5.1(b).

doing reasoning "outside" the ontology gives away its big advantage: Namely, that its reasoner is highly specialised for exactly such a kind of relation-detection.

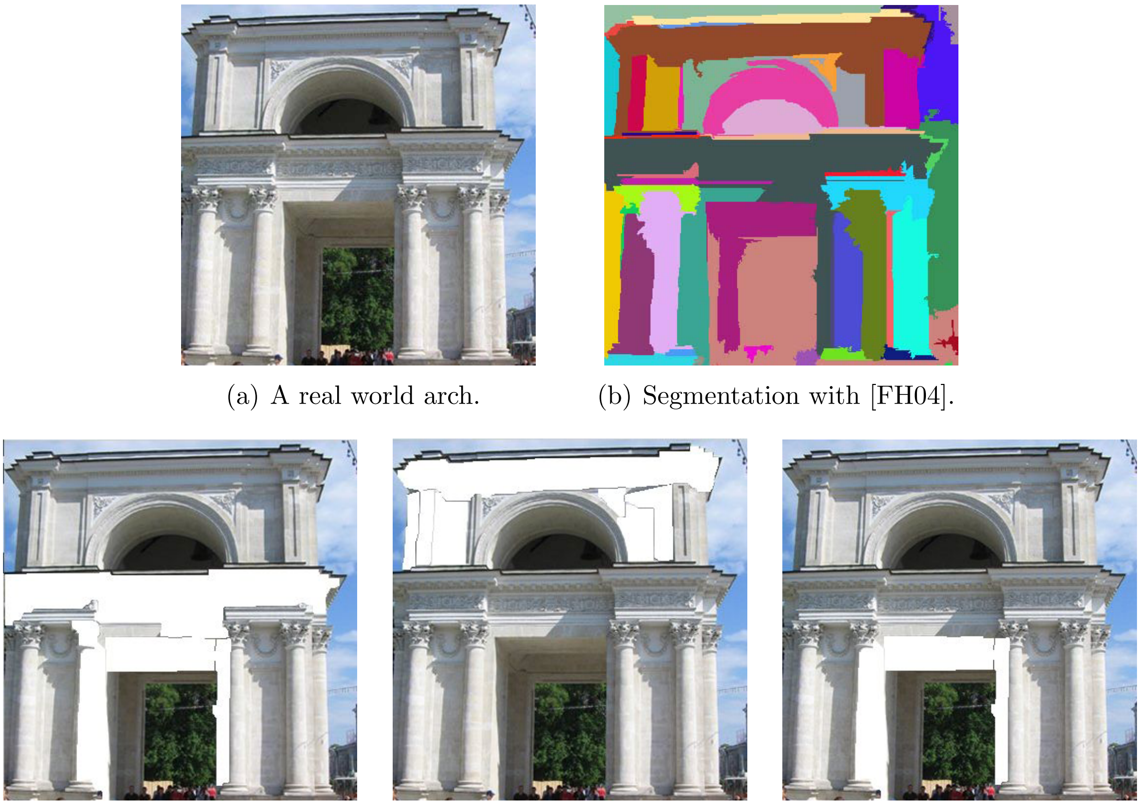
Note that due to the fact that we allow some objects to group together to stacks (using a naïve physics approach, Figure 5.2(b)), we are able to perform *abstraction*: The number of contributing parts is no longer important as long as the overall structure of an arch is maintained. However, this abstraction comes at a cost with this technique of doing reasoning "outside" the ontology: We need to know in advance that we should search for stacks *before* we search for arches. If we do not do this search for stack before, we only find "direct" arches, such as in Figure 5.2(a) and not the ones that we as humans would judge to represent the best result (Figure 5.2(c)).

5.3.5 Interaction Vision-Ontology with Rules

The rule-approach outlined further above does not describe object concepts via restrictions, but rather via *rules* that fire whenever their conditions are fulfilled. The results are shown in Figure 5.3. At first sight, the results look quite similar, which is due to this simple scene. However, a closer look at the segmentation that constitutes the input reveals that there are small patches "between" parts of the stack on the right as well as between the left column and the top-bar, which can only be handled due to the recursivity of the rules that are fired (The arch in Figure 5.3(c) actually consists of six parts.). This is also the reason why we can use the "original" object parts although we already found the stack (Figure 5.3(b)) and need not build up a new object that is the convex hull of both stack parts (as in Figure 5.2(c)). Furthermore, although a lot more possible object concept parts are input to the reasoner (20 in this case), reasoning is much faster as this is done *inside* the reasoning engine of the ontology. This means that we need not take care about what image patch takes the role of being, e.g., the top-bar, but rather let the reasoner find this by filling the appropriate variable in the rule. The gain in speed is mainly due to the fact that reasoners, such as *RACER* that we use, are highly specialised for these tasks[7].

Considering abstraction, of course, we still need a rule that describes how a stack looks like, but the composition of the arch is done by the reasoner alone, firing the rules

[7] We will not go into detail about how this is done. Information about *RACER*'s tableaux calculus and its technology in general can be found in [HTM01, HM03, HM04, HMW04].



(a) A real world arch.

(b) Segmentation with [FH04].

(c) One arch hypothesis.

(d) Another arch hypothesis.

(e) The “best” arch hypothesis.

Figure 5.4: Processing a real world arch: the arch can be found (contributing parts in white) without using a priori appearance or shape information.

as they are fulfilled. This way, abstraction appears much more natural and with less guidance by the user than in the first case.

Figure 5.4 shows a much more complex example of a natural image[8]. As can be seen from the segmentation, a lot more regions are input as possible object part candidates and from the (subset of) results, we can see that there is a large variety of hypotheses. Here, we see the burning need for additional ontological information: All of the results are true “in a sense” – this sense being dependent on the task, the intentionality of the system that investigates this image. “Situated perception” needs to have information at hand that accounts for that. For example, the second picture of the second row shows a correct arch – however, for a task like “drive through the arch”, this would be a false positive. For a task like “find all arch-like structures”, it would be a true positive. For a robot, probably, the “best” solution is the one to the right. It shows the function of arch from the perspective of the robot (again, if the task is to drive through). To put it in a nutshell: Without knowing what we want to do, we cannot judge which one is the right solution. *Vision alone cannot provide us with this information.* If we have the mentioned task, we can easily say that only those arches should be treated as correct

[8] This is the triumphal arch in Chisinau/Moldova; slightly clipped version of a photo used with kind permission of the photographer Mikhail Golubev.

outcome that have a certain height and the bottom lines of which are on the ground plane.

5.3.6 Summary

It is obvious that except for the advantages that we can explicitly separate abstract “description” knowledge from the code, we have not gained additional semantic information yet. Indeed, we could argue that this way, we can have different object concepts represented that can be triggered due to the task. Additionally – and this is what we are heading for with this approach – we have now concepts in our ontology that can be enriched with much more than spatial information. The next section is focusing on this issue. Especially with the latter technique (rule-based), we have the possibility to either search for all (spatial) object concepts that we know, i.e., to do what object re-cognition does on a more detailed level (e.g., using interest points) or to let rules fire that are selectively triggered according to what we currently search for. This latter possibility constitutes what we aim for and reflects situation-dependent perception. The following table summarises the differences of the two approaches:

	Using Restrictions	**Using Rules**
representation	explicit (classes with relations)	implicit (within rules)
description-complexity	higher	lower
readability	worse	better
reasoning	outside the ontology	inside the ontology
speed of reasoning	low	high
exploitation of advantages	very small	quite high

Note that the mentioned notion of *abstraction* is still vision-bound. We believe that such a technique is necessary because of two reasons: The first one – the “bottom-up reason” – is that vision is always prone to errors. We have argued that any computer vision tool that tries to interpret pixel data might “see” regions and segment it as such that we humans do not even see because we already subsumed it (top-down?) to another region, or we immediately noticed (unconsciously) that it is just a small shadow but not a new object part. The second – the “top-down reason” – is that this can (however, only to a very small extent) account for the fact that often “the big picture” is more important than details. For example, even in the simple example of the toy arch in the figures presented, the overall thing that is depicted is an arch – the fact, that the right column is made up of two parts is a detail that we also somehow “see”, but it is a detail the absence of which would not change the impression that there is an arch (e.g., if the right column would be similar to the left). All this is also related to the ill-posed problem of segmentation mentioned earlier. Both in the simple and in the complex arch examples, we could use different algorithms and parameters in order to get different segmentations. However, the problem stays the same: There will be parts that do not refer to parts in reality. We aimed to show that an overall concept is needed in order to hypothesise what could be in the image.

A quite different level of abstraction is in place if our background knowledge tells us more about the object concept than the spatial relations of their parts. In the next example, we will see what this might lead to. Additionally, we will see that such a binding of interesting object concepts can help the robot to choose the right vision tool to use in case a direct search fails. It is here, where the possibilities and the richness of ontologies show their advantage when coupled with computer vision.

5.4 A Situated Vision Example Integrating the Functions

Whereas the previous section only had the intention to show *one* possible representation of object concepts that can be directly used for vision-near *abstraction* in order to perform object concept recognition, we will now try to combine some more of the theoretical issues from the previous chapter with an illustrative example. This will also lift the secret how the different parts of this thesis fit together. We will do this by means of a very typical home robotics scenario in which the robot gets the task to bring the user his/her cup[9]. Before doing this, we need, however, to bring a disclaimer:

The rationale of this section is that we want to show that a State-of-the-Art ontology is suited to host all the different information needed for the variety of subtasks as well as the deployment of the "cognitive functions" extracted in Chapter 4 in an explicit and manageable manner. We by no means claim that this thought experiment solves the robotics or vision problems. We are, of course, aware that each of the steps are itself an own branch of scientific research. Hence, this is unfortunately not the solution to the holy grail of vision for cognitive robotics, but we are pleading for a change from "hacked" information to "neatly engineered" one, as such a comprehensive high-level framework is usually neglected. We are consequently putting information needed in the ontology and show how this helps to integrate the cognitive functions that we delineated in the previous chapter as being necessary for cognitive robotics and consequently for a robot companion. Of course, as with the rest of this thesis, we are focusing on perception-near functions and capabilities. A lot of steps in the following scenario is left out – especially those who deal with navigation, grasping and man-machine communication. Likewise, the cognitive functions mentioned in the following are understood in a perception-related manner. They will most likely take various additional forms in non-perceptual issues.

The ontology – or rather: the relevant part of the ontology – of the robot which we designed for this scenario is depicted in Figure 5.5. Describing the scenario, we will point to the crucial issues that have been central in this thesis: the cognitive functions on the one hand and the integration of different (ranging from more concrete to more abstract) vision techniques on the other hand.

[9] This is actually our group goal.

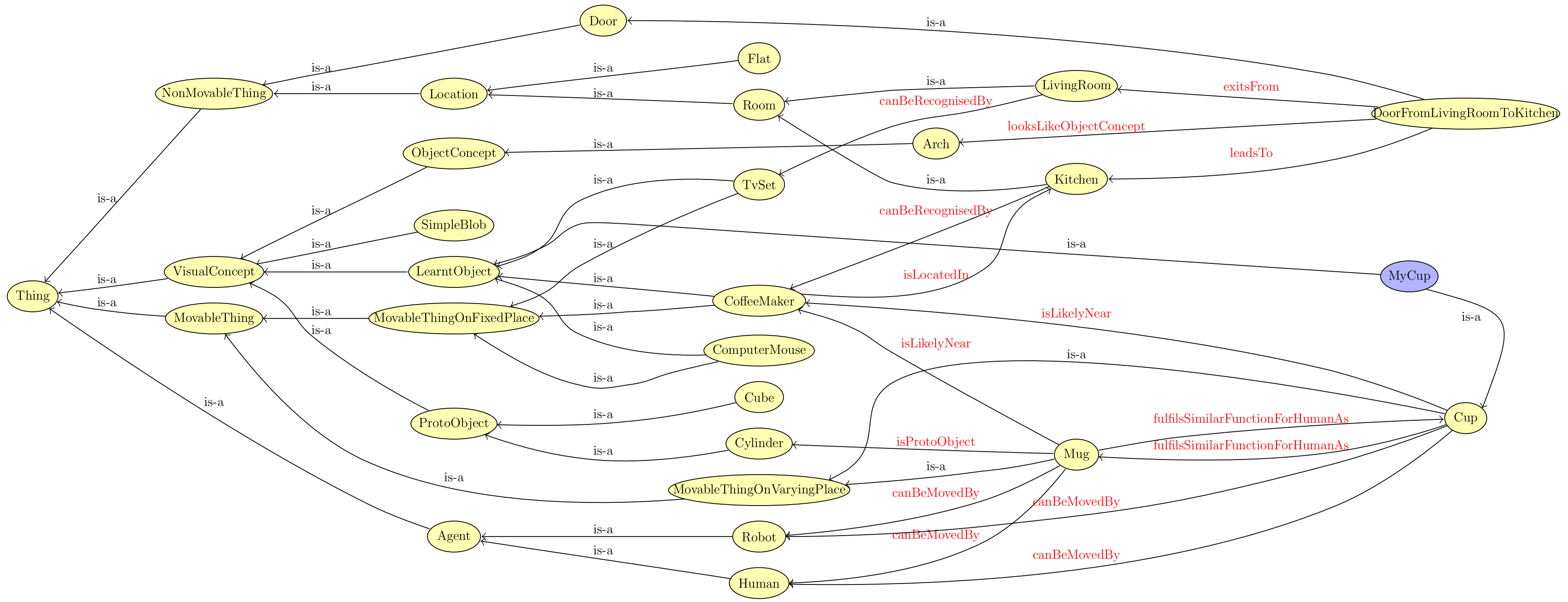


Figure 5.5: Preliminary Ontology Design for the robot companion scenario; generated using the Protégé plug-in *OWLPropViz* [Wac08] and coloured for visibility reasons.

Assume the user gives the robot the following command: *James, please bring me my cup!* – which is one of the typical example tasks that we mentioned in Chapter 1. From now on, the robot is guided by one of the cognitive functions that we identified in Chapter 4: `intentionality` – assuming that it has understood the command and answers to the name James, of course[10]. This principle will be liable for the following queries of the ontology as well as the actions and subgoals the robot sets itself. Of course, a lot of these steps can be done using well-known planning algorithms, but the content of information that is sought and extended in the ontology is guided by intentionality, i.e., the task that is performed.

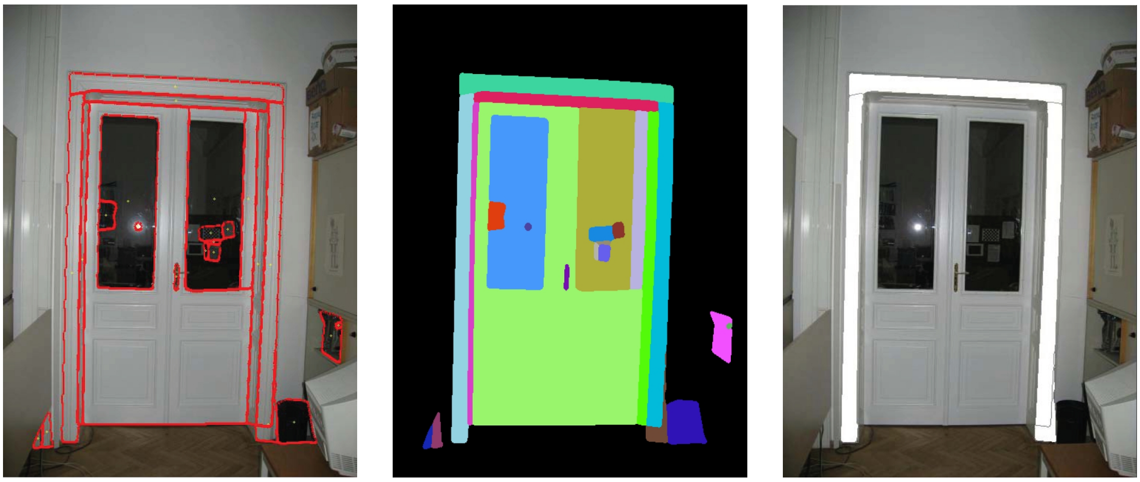
The robot might now want to do some constraint checking. Here is one point where `symbol binding` steps in. The robot needs to not only map "my cup" (and, of course, "bring" and "me" as well) to a concept or individual in its database (in our case: *MyCup*), but also check whether the task as such makes sense. In our case, one possibility could be that the robot assesses whether the task of bringing and the object of cup fit together. As you can see from Figure 5.5, *MyCup* is defined to be of the type *Cup*, which itself is connected to the agent *Robot* by a *canBeMovedBy* relation.

At this point, we need to make a note: This ontology design (Figure 5.5) is only one of an infinite number of possibilities how to design an ontology for the robot. We deliberately chose a simple and straight-forward one. For example, a more complex extension here would be to connect *Cup* with a simple taxonomical *is-a* relation to some concept like "kitchenware" which is further on connected to "HouseholdStuff" which itself could be a subclass of "BringableObject". In this case, the robot would need to go the hierarchy further up until it either arrives at "BringableObject" (as opposed to, e.g., "NonBringableObject"). Additionally, symbol binding can go much further: *Cup* could be connected via a "mayHold"-relation to "Liquids" and thus reveal even more information what a cup is used for. So our ontology design really only shows how to exploit the architectural advantages of an *OWL*-ontology in a straightforward manner.

Back to our example, the robot now knows that it can solve the task – provided it finds the cup, which is the perceptual challenge we are interested in. Here, we can exploit the experience of the robot (or in case of preprogramming it: the practical experience orientation of its designer), which may have led to a connection in the ontology that tells the robot that the instances of *Cup* are *likelyNear* the *CoffeeMaker* which itself is not only a *MovableThingOnAFixedPlace*, but also *isLocatedIn* Kitchen. This naturally leads to the question how to get there, which itself could – as suggested in the Figure – be queried in the ontology with the *leadsTo* question. The delivered answer that *DoorFromLivingRoomToKitchen leadsTo Kitchen* can be exploited to search for that door.

Up until here, we have already encountered totally different concepts which can be grouped as we did into *Agent* concepts (in our case, just holding *Human* and *Robot*), *MovableThing* concepts, of which we already encountered the subclass *MovableThingOnFixedPlace* and *NonMovableThing*, which holds the just encountered concept of *Door*

[10] In the following, we use a `typewriter font` to underline the cognitive functions that we discussed in Chapter 4 and *italics* for the concepts and relations defined in the ontology diagram of Figure 5.5.



(a) A door with closures found by [Zil07].

(b) Blobs after postprocessing (see Section 5.3.3).

(c) Best hypothesis of arch found.

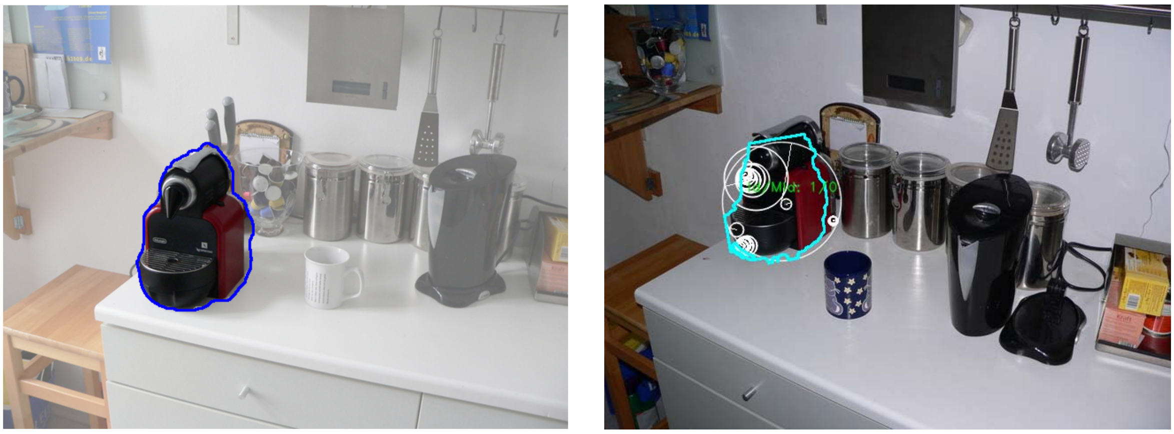
Figure 5.6: Detection of a door by detecting its surrounding "arch" (door frame).

with the more specific subclass of *DoorFromLivingRoomToKitchen*[11]. Some of the concepts are rather typical for robotics (e.g., *Agent*) whereas others are rather typical for perceptual questions. This is one of the advantages of using such an ontology and this is why it is well suited for our goal of combining the different aspects: You can define concepts and relations that are suited for typical perceptual purposes of the robot. One example will be the definition of special perception-associated concepts and relations that we will use now.

DoorFromLivingRoomToKitchen is connected via a *looksLikeObjectConcept* relation to the concept of *Arch*, which itself is a subclass of *ObjectConcept*. The idea is that the robot knows some of these object concepts that could be, for example, detected via the rule approach which we presented further above. Note that we could have, of course, also connected *Door* with *Arch*, but we chose to connect the more specific subclass *DoorFromLivingRoomToKitchen* as another door in the flat (the robot's spatial ecological niche) might rather be a double wing door and thus look differently and can be detected as another object concept. In our case, the door to the kitchen is supposed to have a clearly visible door frame and is thus detectable by the arch concept. Figure 5.6 shows our concept detection mechanism from Section 5.3 applied to a door. Note that when the robot decides (due to its knowledge in the ontology) to use this mechanism, it does indeed perform `abstraction` as shown in the previous section. Furthermore, it `predicts` what is likely to be seen as it stands in the living room and knows there must be a door somewhere, i.e., an arch in this case, and triggers the appropriate rule.

Assuming that the robot found its way to the kitchen using all its complex planning and object avoidance skills, it is again in need of deliberate thinking. For example, it could check whether it is really in the kitchen or it wants to orient itself towards the coffee maker. This is necessary, because though a robot companion might maintain a

[11] This could, in fact, also be modeled as "individual" rather than as concept, cp. the subclass-vs.-instance-problem mentioned in Section 5.2 – more on design issues can be found in [HJM+07].



(a) One of the learning images for the coffee maker, the manually segmented region of interest is shown in blue.

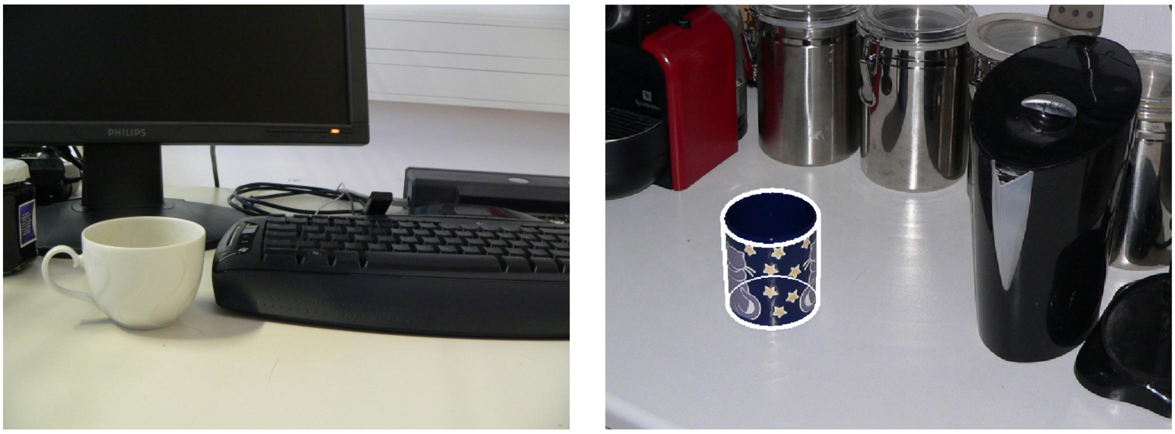
(b) Detecting the coffee maker with interest points.

Figure 5.7: The coffee maker is learnt appearance-based via interest points and detected as “proof” that the robot is in the kitchen.

map of the flat at some point, it might not have one at the beginning, it might be used in a different flat for some reason or localisation just fails. Here again, concepts are much more general than a specific map of a specific flat.

We suppose that the robot has learnt the appearance of the coffee maker in this household before (for example, by having been shown by the user). We assume here that it is reasonable that some of the objects handled are not stored as concepts of proto-objects but really as instances with a specific appearance, e.g., by storing an aggregate of SIFT-keys as we did for our tracking approach in Chapter 3. This holds for objects for “typical” and “often used” home robotics tasks in particular. We would argue that here again, however, `prediction` is especially of importance. The reason is simple: It is well-known to computer vision that the check for all objects known to the robot is computationally very exhaustive and furthermore, the more objects stored the more complex the classification gets. Using `prediction`, however, and the “non-visual” information that we are now in the kitchen, reduces immensely the search space. The robot predicts to see kitchenware; therefore it only needs to search for this class of items in its camera frame.

In the ontology design diagram (Figure 5.5), we have an arc from *CoffeeMaker* to *Kitchen* which is called *canBeRecognisedBy* and points to exactly the fact just mentioned. Hence, using now *is-a LearntObject* rather than *looksLikeObjectConcept* invokes its object recognition technique, which is an implementation similar to [LLS07]: A previously generated codebook, i.e., a collection of clusters of detected interest points, is used to redetect an object. Hereby, every interest point votes for the centre of the respective object, those centres with most votes “win”. Figure 5.7 shows one of the learning images, where the part in which the interest points are searched has been masked manually, along with the frame in which the coffee maker is detected later on, using the information acquired before. Interest points are marked in white and the rough estimated outline of the coffee maker is projected in blue.



(a) *MyCup* on an office table, instead of in the kitchen. The robot needs to find another solution.

(b) Detecting the proto-object "cylinder" (superimposed in white) with an extension of [Zil07] as alternative solution to *MyCup*.

Figure 5.8: A cup and a mug – functionally related via the ontology and therefore serving as alternative solutions.

Back to our storyline. The robot now has verified that it is in the kitchen, and its intentionality still drives it to search for *MyCup*. The robot knows that it *isLikelyNear* the *CoffeeMaker*, thus, `intentionality` and `prediction` along with the `symbol binding` capability (which tells the robot in this case, where to look) helps to guide attention to a relevant part of the kitchen. We can assume that an often used part, such as *MyCup* would actually be represented as appearance-based learnt object. Let us assume it looks like the one in Figure 5.8(a). However, if the robot is not able to find it (say, there is no second entry in the ontology that tells the robot that *MyCup* might also be *likelyNear* the computer keyboard), and there is no learnt proto-object or object concept that enables the robot to search for a more generic *Cup*, it is imperative that the robot "knows" what a cup is: Here, the main reason for `symbol binding` as we propagandise by using an ontology becomes apparent. Again, we are talking about robot companions, and – as discussed in Chapter 4 in the context of the so-called frame-of-reference problem – this is why symbol binding does not (only) refer to the "understanding" what the cup is in the robot's meaning, but the robot needs to know what the cup means for the human.

Why this is important for a thesis concerned with perception is obvious: Most easily, there is an entry in the ontology which directly links the cup to another concept which is also bound to a perception technique. This is indicated with the *fulfilsSimilarFunctionForHumanAs*-link from *Cup* to *Mug*. This is, of course, very simple and rather a shortcut as a richer ontology could likewise connect *Cup* and *Mug* via some *isADrinkingContainer* or *mightHoldLiquids*-link. This, then, is a question of reasoning to retrieve the best suited "function" for the human. What we emphasise is that, here, another kind of `abstraction` comes into play which is as necessary to cognitive robotics as is the visual abstraction having tackled earlier. For the robot might now search for an object with similar function, end up with the concept *Mug* and thus not only see that this, too, *isLikelyNear* the *CoffeeMaker* but also get the information that it *isProtoObject Cylinder*.

Again, the robot can now invoke yet another vision technique that is integrated in its ontology: It searches for any mug, therefore the actual appearance is irrelevant and it is truly only interested in the "proto-mug" as this shall deliver any mug in the image. This is exactly what it can do, because the cup, its actual task, might not be deliverable, however, it can still (maybe) satisfy the user by delivering a "similar" object[12]. Hence, the robot can now – still using its attention to zoom in to places next to the coffee maker (due to the knowledge of *isLikelyNear*) – search for the proto-object of a cylinder, detect the mug as shown in Figure 5.8(b), bring it to the user and thus fulfil its task. Likewise, the detection of this proto-object could be performed using the Superquadric approach of Chapter 3. There are various possibilities that might all be governed by the ontological knowledge about objects and scenes.

The last exercise – bringing the mug to the user – involves its grasping, where techniques of visual servoing can be applied. Here, we finally also returned to the roots of this thesis: One possibility to manage the tracking part of the visual servoing is to use the approach presented in Chapter 3. Having localised the object that shall be handled as presented above, the system can now apply the tracking technique which uses interest points that are localised now, in the actual lighting situation with the actual instance under inspection. Note that when the task started in the living room, the system had no clue that it will use interest points on the blue mug which it might even never have seen before, in order to retrieve it and fulfil its assignment.

5.5 Résumé

In this chapter, we started with showing how to overcome some of the limitations that automatically arise when information that vision algorithms use are not made explicit. After having outlined the main advantages of using a State-of-the-Art ontology language and tool, such as *OWL* or *RACER*, we started with a very simple example of detecting an arch with the help of such an explicit information representation. Main rationale was that when dealing only with spatial relations we can – on a quite basic level – represent the "constituting substrate" of an arch without being bound to accidentals such as appearance information or size parameters. Here, we compared a restriction-based with a rule-based approach and showed that the latter is better suited for our objective than the former. Nevertheless, "explicit knowledge" as propagandised throughout this thesis, is of course better to be represented directly in the class of the object (as in the restriction-based approach) and not in rules. Consequently, further research projects might tackle the generation of these rules *out of* such a detailed class description. Especially interesting would be the generation of the restrictions and rules out of the experience of the robot, and this way performing `generalisation`, i.e., the one cognitive function that has also been extracted from our theoretical considerations yet has not been tackled. In particular, difficulty lies in finding the necessary negative

[12] It is clear that it would likewise be possible just to not deliver anything and ask the user what he/she wants to get instead. However, the peculiarities of human-machine-interaction are not our major concern.

statements (e.g., the column on the left does *not* touch the column on the right in case of the arch).

The last section of this chapter then tried to show how this basic visual abstraction technique can play one role in a more comprehensive ontological framework that conceptually tackles tasks in a home robotic scenario. We showed how different visual modalities can be incorporated – by always keeping it *situated*, i.e., bound to a task in a specific situation.

Concluding, we will now have a look on typical objections concerning the approach of using logic and vision as we did in the basic arch example. One dispute concerns the use of symbols, which we already defended in Sections 4.5.6 and 4.7.2. Once this stance is accepted, we should ask the related question whether the use of *logics* makes sense at all. Vision is often rather thought of as something "fuzzy", something deploying "probabilities" and "uncertainties". It is, of course, true that we cannot possibly describe the whole world in logic statements. We would, however, object to the often heard consequence that this implies that it is senseless to even try to model a part of it in such a way. The point is that there might nevertheless be "stable" concepts, especially concerning object concepts, that are in fact connected to higher-level thoughts in similar ways as we tried to show in our situated vision example (Section 5.4). For instance, we humans will probably agree that a mug can almost always be used if no cup can be found. We will always find exceptions, and when we do, we might adjust our world-view. That is fine – it is not said that an ontology needs to be fixed all through the lifetime of the robot – but at each point having a current view, a lot of additional information can be retrieved using a network of relations and concepts, such as the one shown. We do agree, however, that it is a drawback, that we cannot (yet?) explicitly store probabilities of relations. Still then, we would need to ask where the probabilities come from, how they change due to experience and how they can help us at all when doing reasoning.

To put it in a nutshell, we would say that modelling the world *only* symbolically and *only* logically might be ill-posed, but modelling some object concepts that are relatively stable and connect them to other information (e.g., "associations" to function or typical show-ups) might still be useful. This is also due to the fact that we are working in a limited domain (home robotics), which implies a clear niche and robots that always have intentionality.

Another objection, namely that perception is not always perfect and a search via strict rules is therefore ill-posed, has to be taken serious as well. As we saw, of course, a minimal quality of the image processing is still important, no matter whether detecting object concepts, proto-objects or stored instances. We would plead to use a representation that is *rather* of coarse than of fine granularity – this is just a design recommendation that we found useful. Of course, the respective interplay of task, niche and perceptual capability will influence this decision. But a very detailed description of how a door looks like, including its handle, possible windows and high-resolution structure, is not what we mean by "object concept".

We concluded with a scenario that showed that "the use of ontologies for situated perception in robots"...

1. ...assists designer and robot to manage explicit information,
2. ...can lead to a guidance of what perception technique should be used when and
3. ...helps in retrieving additional non-visual information.

Additionally, we could show that the cognitive functions that we demanded in Chapter 4 can be deployed by the robot this way. The function of `generalisation` would be, of course, to “learn” the ontology that we pre-gave the robot and thus build up its own representation of the world – following a truly constructivist approach.

Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.

WINSTON CHURCHILL
[Chu42]

Chapter 6

Summary and Future Work

So here we are, finally conducting the last "movement" of our "symphony" that brought us through various issues from computer vision and cognitive robotics. In the following, we will reprise the major themes that we tackled and hope to end in an *Allegro con brio.*

We started with two typical computer vision problems and their solutions in Part I. In Chapter 2, we showed an approach to exploit appearance information for assisting a model-based object tracking tool. We computed adjacent matrices of texture measures and combined it with a sliding window technique in order to find the biggest deviation between the matrices where we assumed to have found the edge. We could show that our approach enhances tracking robustness in situations where colour edge detection alone fails (e.g., in case of cluttered surfaces delivering too many spurious colour edges).

Chapter 3 then presented an integrated system for handling basic object shapes that aimed to reduce necessary interaction by the human operator. In order to achieve this, the system is able to learn the object model in Superquadric parameters, detect it in the scene (both done with a laser scanner) and finally track it throughout an image sequence (with a monocular camera). This system integrates two different vision modalities, namely a three-dimensional point cloud that only captures shape and two-dimensional appearance information in terms of interest points. Our contribution here was the development of an appearance-based interest point monocular tracking approach for objects in 3D Superquadric representation.

In summary, Part I dealt with techniques that use a quantitative representation of the object handled. Any information (except for the shape parameters used in the "wire-frame"- and "Superquadric"- models, respectively) is "inside" the code. There is no notion of "object", only of aggregates of pixel data that the systems use. Indeed, they are stand-alone versions of concrete solutions to very specific job definitions.

Based on the insights from this low-level vision part, we started to investigate what these techniques lack in order to be suited for *situated perception in robots.* Part II started in Chapter 4 with the call to tackle the right problems instead of focusing on (usually very narrow) solutions. In order to talk about the right problems we conducted a thorough analysis of theoretical issues involved in the question how perception could interact with higher-level reasoning by using a common ontology. Thus, we first exposed our general lines of thought, including questions of ontology and epistemology and tackled issues such as anthropomorphism and rationalism. We came up with the view that we might need a clear look on *cognitive functions* that define general, versatile and "intelligent" behaviour – the latter notion we roughly defined as the third-person perspective of the principle of cognition. As functions we enumerated and justified *intentionality*, *prediction*, *abstraction*, *symbol binding* and *generalisation* as especially relevant for situated perception of a robot companion. We found the idea of using a common ontology well suited for connecting those and other, partly non-visual, functions and defended the view that it should need "concepts" in the psychological sense, whereby we focused – according to the topic – on *perceptual object concepts.* The scientific contribution of this chapter was a structured theoretical account of many issues that are consulted every now and then in the respective engineering disciplines without thoroughly looking on their implications. Especially in their discussion and fertilisation for the context of computer vision and cognitive robotics lay our main interest.

The rest of the thesis finally showed an implementation of these ideas by using an ontology in the narrow engineering sense. It introduced the use of *OWL* and *RACER* to store the object concepts in order to first show how visual abstraction can be achieved with only qualitative information stored in the knowledge base and second how a comprehensive example of a typical home robotic task (the delivery of a drinking vessel to the user) might be solved by integrating various different vision techniques and applying the cognitive functions shown earlier. Here, the use of *OWL* and the bridge from quantitative vision data to qualitative information in the ontology in terms of object concepts constituted the novelty of our approach.

6.1 Central concern

Main themes throughout the thesis were the question how computer vision techniques can be integrated as robotic vision techniques and used to account for the mandatory situatedness, the notion of object and why this can only be reached by integrating non-visual information, the focus on cognitive functions and, of course, the logical approach to structure common knowledge in an ontology. A special concern, too, was the focus on the importance of *explicit, qualitative information* as usually descriptions used in computer vision stay on a quantitative level and mostly "inside the code" which hampers the step from shallow *data* to grounded, situated and hence flexible *information.* Related to that is the distinction between *proto-object* and *object.*

We chose to present both typical computer vision tasks (Part I) and theoretical considerations (Part II) and could in fact call the first part "robotics as technology" and the second "robotics as science" (a distinction made by D. Parisi). We found that a technical PhD-thesis should comprise both aspects and tried to combine these works – although quite distinct at first sight – via the consideration that an integrated theory needs to know both.

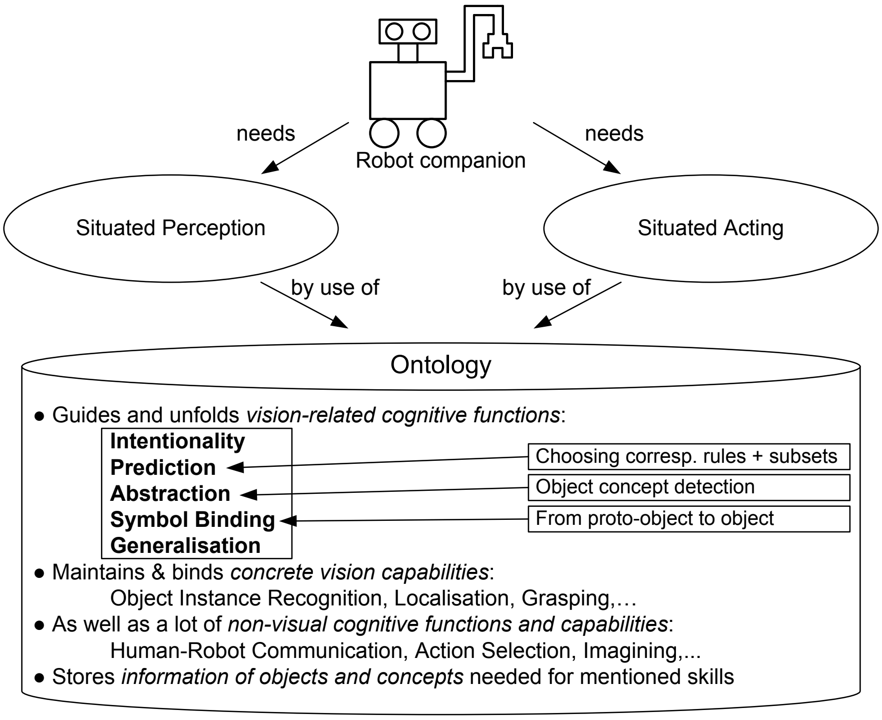

needs
needs
Robot companion
Situated Perception
Situated Acting
by use of
by use of
Ontology
• Guides and unfolds *vision-related cognitive functions*:
Intentionality
Prediction
Abstraction
Symbol Binding
Generalisation
Choosing corresp. rules + subsets
Object concept detection
From proto-object to object
• Maintains & binds *concrete vision capabilities*:
Object Instance Recognition, Localisation, Grasping,…
• As well as a lot of *non-visual cognitive functions and capabilities*:
Human-Robot Communication, Action Selection, Imagining,...
• Stores *information of objects and concepts* needed for mentioned skills

Figure 6.1: Conclusion: Robot companions need a shared ontology that integrates and unfolds cognitive functions and concrete capabilities. The focus of this thesis is marked.

One debatable aspect of this thesis is that the choice of a specific focus – vision in our case – is already again a "breaking-up" of the system into pieces, which might contradict a holistic view. However, we claim that the use of the functional layer actually provides the necessary connection to the whole system.

This thesis did not aim at giving a thorough evaluation of this or that algorithm or representation scheme. Neither was our major concern to advance well-known stand-alone vision applications, such as object recognition or object class recognition, in terms of speed or recognition rate. The goal was to provide the theoretical fundament of future robot companion research and to show what is needed in order to achieve the versatile and autonomous functionalities that we aim at. Similarly as the concept of *embodiment* has revolutionised the field of research with respect to "complete agents", we think that *envisionment* should impinge the field of robot companion research. Figure 6.1 shows once again the main ideas and needs for future research having this objective.

6.2 Open Research Questions

Of course, this really is not the end, but maybe the end of the beginning: One topic left out totally is what we subsumed under the notion of the cognitive function of *generalisation*: Learning the object concepts is everything else than trivial – especially if we stick to a logical description as used in the ontology of Chapter 5. However, using rules to do machine learning might in fact be more plausible in science theory than accumulating probabilities. For this, we propose that the stances of W.v.O. Quine, for whom total science is like "[...] a field of force whose boundary conditions are experience [...]" [Qui53], as well as Popper's falsificationist approach [Pop35] provide suitable starting points. This means that the whole body of theories of the agent is permanently prone to revision and possible partly rejection. One first example could be that the "arch" that we had as very simple example could be learnt by showing various examples, doing the same vision preprocessing and trying to assemble bits of information that are stable throughout the images in order to come up with a rule similar to the hard-coded one presented[1].

Another interesting question that can be derived from our research is to what extent this approach can be combined and mapped onto Jean Piaget's work whose groundbreaking theory of *schemas* is the foundation for understanding how humans acquire and use knowledge in situations in order to act in them – in a consistent manner (e.g., [Pia54]). Following [MFH88], a cognitive schema is thus an entity of arrangement with which new experiences can be integrated in an existing knowledge system. This is especially relevant for a project like ours as the very strong statement of [Rum80] – "They [schemata] are the fundamental elements upon which all information processing depends [...]" – is in fact a direct account of situatedness and points to the interesting interplay of empirical data and pre-given (possibly innate) knowledge. According to the authors, all cognitive activities (be it perceptions or further interpretations) are working with those schemata in the sense of establishing, comparing, selecting and using them.

One of the driving forces of this thesis was our group goal of an autonomous and intelligently behaving robot companion and the entailed questions on cognitive functionality needed. Therefore, the main open question concerns the integration of different techniques, frameworks and information bits needed to achieve this goal. The challenge is still open, we tried to provide the theoretical framework to this end. We could only partially provide a sketch on the ontology needed and there is still the question how much scene knowledge needs to be integrated beforehand. Furthermore, our example implementation aimed at showing what is possible. In the future, this can be enriched, including the maintenance of multiple hypotheses-strands, imagined outcomes of actions, spatio-temporal information, and, of course, lots more specific vision techniques including using 3D-sensors and the like.

The materialization of the personal robot butler might still be long way off, but we are heading for it and it keeps being worth dreamt of.

[1] Thanks to David Hogg from the University of Leeds for pointing to this.

Bibliography

[AC06] P. Avero and M.G. Calvo. Affective Priming with Pictures of Emotional Scenes: The Role of Perceptual Similarity and Category Relatedness. *Spanish Journal of Psychology*, 9(1):10–18, 2006.

[AH97a] P. Agre and I. Horswill. Lifeworld analysis. *Journal of Artificial Intelligence Research*, 6:111–145, 1997.

[AH97b] M. Ahissar and S. Hochstein. Task Difficulty and the Specifity of Perceptual Learning. *Nature*, 387:401–406, 1997.

[Ari99] Aristotle. *The Metaphysics.* Penguin Classics, New York, NY, USA, 1999.

[Ari06] Aristotle. *The Categories.* Dodo Press, 2006.

[Ayr03] M. Ayromlou. *Cue Integration Techniques for Robust Feature Tracking.* PhD thesis, Vienna University of Technology, Vienna, Austria, 2003.

[Bar81] A.H. Barr. Superquadrics and Angle-Preserving Transformations. *IEEE Computer Graphics and Applications*, 1(1):11–23, 1981.

[BBC93] L. Birnbaum, M. Brand, and P. Cooper. Looking for Trouble: Using Causal Semantics to Direct Focus of Attention. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 49–56. IEEE Computer Press, 1993.

[Ben99] M. Bennett. Introduction. In M. Bennett, editor, *Developmental Psychology – Achievements and Prospects*, chapter 1, pages 1–12. Psychology Press, Philadelphia, PA, USA, 1999.

[BGG03] V. Bruce, P.R. Green, and M.A. Georgeson. *Visual Perception – Physiology, Psychology and Ecology.* Psychology Press, Hove, Great Britain, 4^{th} edition, 2003.

[Bie06] G. Biegelbauer. *Efficient Part Feature and Object Detection by Fitting Geometric Models to Range Image Data.* PhD thesis, Vienna University of Technology, Vienna, Austria, 2006.

[Blo97] N. Block. On a Confusion about a Function of Consciousness. In N. Block, O. Flanagan, and G. Güzeldere, editors, *The Nature of Consciousness*, chapter 20, pages 375–416. MIT Press, Cambridge, MA, USA, 1997.

[Bor23] E. Boring. Intelligence as the Tests Test It. *New Republic*, 36:35–37, 1923.

[Bor99] M. Bornstein. Human Infancy: Past, Present, Future. In M. Bennett, editor, *Developmental Psychology – Achievements and Prospects*, chapter 2, pages 13–35. Psychology Press, Philadelphia, PA, USA, 1999.

[Bro91] R. Brooks. Intelligence Without Representation. *Artificial Intelligence*, 41:139–159, 1991.

[BSV06] G. Biegelbauer, M. Schlemmer, and M. Vincze. Learning the Object Model for Automatic Detection and Tracking for Robot Grasping. In *Proceedings of the* 8^{th} *International IFAC Symposium on Robot Control (SYROCO)*, Bologna, Italy, 2006.

[BT78] H. Barrow and J. Tenenbaum. Recovering Intrinsic Scene Characteristics From Images. In A. Hanson and E. Riseman, editors, *Computer Vision Systems*, pages 3–26. Academic Press, New York, NY, USA, 1978.

[BTG06] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *Proceedings of the* 9^{th} *European Conference on Computer Vision (ECCV)*, volume 3951 of *Lecture Notes in Computer Science (LNCS)*, pages 404–417. Heidelberg, Germany, 2006.

[Cas99] R. Case. Conceptual Development. In M. Bennett, editor, *Developmental Psychology – Achievements and Prospects*, chapter 3, pages 36–54. Psychology Press, Philadelphia, PA, USA, 1999.

[CBGG97] A. Cohn, B. Bennett, J. Gooday, and N.M. Gotts. Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *Geoinformatica*, 1:1–44, 1997.

[CCPR02] A. Chella, M. Cossentino, R. Pirrone, and A. Ruisi. Modeling Ontologies for Robotic Environments. In *Proceedings of the* 14^{th} *International Conference on Software Engineering and Knowledge Engineering (SEKE)*, pages 77–80, 2002.

[CH01] A.G. Cohn and S.M. Hazarika. Qualitative Spatial Representation and Reasoning: An Overview. *Fundamenta Informaticae*, 46(1-2):1–29, 2001.

[Cha95] D. Chalmers. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.

[Cha96] D. Chalmers. *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press, Oxford, United Kingdom, 1996.

[Chr99] T. Christaller. Cognitive Robotics: A New Approach to Artificial Intelligence. *Artificial Life and Robotics*, 3(4):221–224, 1999.

[Chu42] W. Churchill. Cited from The Churchill Centre, online at: `http://www.winstonchurchill.org` (12/2008), 1942. In London, United Kingdom.

[CL97] D. Crevier and R. Lepage. Knowledge-Based Image Understanding Systems: A Survey. *Computer Vision and Image Understanding*, 67:161–185, 1997.

[Cla01] A. Clark. *Mindware – An Introduction to the Philosophy of Cognitive Science*. Oxford University Press, New York, NY, USA, 2001.

[Com08] Committee of CogRob08. Website of the Bi-Annual Workshop on Cognitive Robotics. Online at: `http://www.cse.yorku.ca/cogrob08/index.html` (12/2008), 2008.

[Con90] J.H. Connell. *Minimalist Mobile Robotics: A Colony-Style Architecture for an Artificial Creature*. Academic Press, San Diego, CA, USA, 1990.

[Del57] R. Delaunay. La Lumière (1912). In Pierre Francastel, editor, *Du Cubisme à l'art abstrait*, page 147. S.E.V.P.E.N., Paris, France, 1957. Cited from: Gordon Hughes, *Coming into Sight: Seeing Robert Delaunay's Structure of Vision*, in: October 102, pp 87-100, MIT Press, 2002.

[Elm90] J. Elman. Representation and Structure in Connectionist Models. In G. Altman, editor, *Cognitive Models of Speech Processing*, chapter 17, pages 345–382. MIT Press, Cambridge, MA, USA, 1990.

[Epi87] Epictetus. *Discourses*. Hayes Barton Press, 1887. Translated by G. Long.

[Eva03] J. Evans. In Two Minds: Dual-Process Accounts of Reasoning. *Trends in Cognitive Sciences*, 7(10):454–459, 2003.

[Eva08] J. Evans. Dual Processing Accounts of Reasoning, Judgement, and Social Cognition. *Annual Review of Psychology*, 2008. In press.

[FB81] M.A. Fischler and R.C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[FDP06] M. Fussenegger, R. Deriche, and A. Pinz. A Multiphase Level Set based Segmentation Framework with Pose Invariant Shape Priors. In *The 7th Asian Conference on Computer Vision (ACCV)*, volume 3851/2006 of *Lecture Notes in Computer Science (LNCS)*, pages 674–683. Springer, Heidelberg, Germany, Jan 2006.

[Fel03] J. Feldman. The Simplicity Principle in Human Concept Learning. *Current Directions in Psychological Science*, 12(6):227–232, 2003.

[Fel06] J. Feldman. An Algebra of Human Concept Learning. *Journal of Mathematical Psychology*, (50):339–368, 2006.

[FH04] P. Felzenszwalb and D. Huttenlocher. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[Fod75] J. Fodor. *The Language of Thought.* Cromwell, New York, NY, USA, 1975.

[Fod98] J. Fodor. *Concepts: Where Cognitive Science Went Wrong.* Oxford Cognitive Science Series. Oxford University Press, New York, NY, USA, 1998.

[FP03] D. Forsyth and J. Ponce. *Computer Vision – A Modern Approach.* Prentice Hall, Upper Saddle River, NJ, USA, 1st edition, 2003.

[GD06] L. Gupta and S. Das. Texture Edge Detection using Multi-resolution Features and SOM. In *Proceedings of the* 18th *International Conference on Pattern Recognition (ICPR)*, volume 2, pages 199–202, 2006.

[Gru93] T. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[HA02] S. Hochstein and M. Ahissar. View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron*, 36:791–804, 2002.

[Ham90] D.W. Hamlyn. *In and Out of the Black Box: On The Philosophy of Cognition.* Blackwell, Oxford, United Kingdom, 1990.

[Har65] N. Hartmann. *Grundzüge einer Metaphysik der Erkenntnis (Outlines of a metaphysics of cognition).* Walter de Gruyter & Co., Berlin, Germany, 5th edition, 1965.

[Har90] S. Harnad. The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42:335–346, 1990.

[HB04] J. Hawkins and S. Blakeslee. *On Intelligence.* Times Books, New York, NY, USA, 2004.

[HJM^{+}07] M. Horridge, S. Jupp, G. Moulton, A. Rector, R. Stevens, and C. Wroe. A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools. Technical Report edition 1.1, University of Manchester, 2007. Online at: `http://www.co-ode.org` (12/2008).

[HKRS08] P. Hitzler, M. Krötzsch, S. Rudolph, and Y. Sure. *Semantic Web: Grundlagen.* Springer, Heidelberg, Germany, 2008.

[HM01] V. Haarslev and R. Möller. Racer System Description. In R. Goré, A. Leitsch, and T. Nipkow, editors, *Proceedings of the International Conference on Automated Reasoning (IJCAR)*, pages 701–705. Springer, 2001.

[HM03] V. Haarslev and R. Möller. Racer: A Core Inference Engine for the Semantic Web. pages 27–36, Sanibel Island, FL, USA, 2003.

[HM04] V. Haarslev and R. Möller. Optimization Techniques for Retrieving Resources Described in OWL/RDF Documents: First Results. In *Proceedings of the* 9th *International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, pages 163–173, Whistler, BC, Canada, 2004.

[HMW04] V. Haarslev, R. Möller, and M. Wessel. Querying the Semantic Web with Racer + nRQL. In *Proceedings of the KI-2004 International Workshop on Applications of Description Logics (ADL)*, Ulm, Germany, 2004.

[Hof04] T. Hofweber. Logic and Ontology. Entry of the *Stanford Encyclopedia of Philosophy.* Online at: `http://plato.stanford.edu/entries/logic-ontology/` (12/2008), Oct. 2004.

[Hop82] J.J. Hopfield. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. In *Proceedings of the National Academy of Sciences of the USA*, volume 79, pages 2554–2558, 1982.

[HS88] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proceedings of the* 4th *ALVEY Vision Conference*, pages 147–151, Manchester, United Kingdom, 1988.

[HSD73] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, 1973.

[HT98] G.D. Hager and K. Toyama. X Vision: A Portable Substrate for Real-Time Vision Applications. *Computer Vision and Image Understanding*, 69(1):23–37, 1998.

[HT03] C. Hudelot and M. Thonnat. A Cognitive Vision Platform for Automatic Recognition of Natural Complex Objects. In *Proceedings of the* 15th *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 398–405, 2003.

[HTM01] V. Haarslev, M. Timmann, and R. Möller. Combining Tableaux and Algebraic Methods for Reasoning with Qualified Number Restrictions. In *Proceedings of the International Workshop on Description Logics (DL)*, pages 152–161, Stanford, CA, USA, 2001.

[Hus54] E. Husserl. *Husserliana VI: Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie (The Crisis of European Sciences and Transcendental Phenomenology).* Martinus Nijhoff, Den Haag, The Netherlands, 1954.

[Hus64] E. Husserl. *Erfahrung und Urteil – Untersuchungen zur Genealogie der Logik (Experience and Judgement).* Claasen Verlag, Hamburg, Germany, 3rd edition, 1964.

[HYMLJ05] J. Han-Young, H. Moradi, S. Lee, and H. JungHyun. A Visibility-Based Accessibility Analysis of the Grasp Points for Real-Time Manipulation. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 3111–3116, Edmonton, Canada, 2005.

[Jac81] B. Jacobs. *Angst in der Prüfung – Beiträge zu einer kognitiven Theorie der Angstentstehung in Prüfungssituationen (Fear in Examination – Contributions to a cognitive theory of fear development in exam situations).* Fischer, Frankfurt/Main, Germany, 1981.

[Jac85] P.C. Jackson. *Introduction to Artificial Intelligence.* Courier Dover Publications, Mineola, NY, USA, 2^{nd} edition, 1985.

[Joh99] M. Johnson. Developmental Cognitive Neuroscience. In M. Bennett, editor, *Developmental Psychology – Achievements and Prospects*, chapter 9, pages 147–164. Psychology Press, Philadelphia, PA, USA, 1999.

[Jor86] M.I. Jordan. Attractor Dynamics and Parallelism in a Connectionist Sequential Machine. In *Proceedings of the 8^{th} Conference on Cognitive Science*, pages 531–546, Amherst, MA, USA, 1986.

[Kan78] T. Kanade. Region Segmentation: Signal vs. Semantics. In *Proceedings of the 4^{th} International Joint Conference on Pattern Recognition*, pages 95–105, 1978.

[Kan99] I. Kant. *Critique of Pure Reason (1787).* Cambridge University Press, Cambridge, United Kingdom, 1999.

[KC02] D. Kragic and H. Christensen. Model Based Techniques for Robotic Servoing and Grasping. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 299–304, Lausanne, Switzerland, 2002.

[Kei99] F.C. Keil. Cognition, Content, and Development. In M. Bennett, editor, *Developmental Psychology – Achievements and Prospects*, chapter 10, pages 165–184. Psychology Press, Philadelphia, PA, USA, 1999.

[KG05] Y. Katz and B.C. Grau. Representing Qualitative Spatial Information in OWL DL. In *Proceedings of the 1^{st} International Workshop: OWL Experiences and Directions*, Galway, Ireland, Nov. 2005.

[KGB07] K. Kveraga, A.S. Ghuman, and M. Bar. Top-down Predictions in the Cognitive Brain. *Brain and Cognition*, 65(2):145–168, 2007.

[KK05] V. Kyrki and D. Kragic. Integration of Model-based and Model-free Cues for Visual Object Tracking in 3D. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1554–1560, Barcelona, Spain, 2005.

[KK06] V. Kyrki and D. Kragic. Tracking Unobservable Rotations by Cue Integration. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2744–2750, Orlando, FL, USA, 2006.

[KKK03] S. Kim, I. Kweon, and I. Kim. Robust Model-Based 3D Object Recognition by Combining Feature Matching with Tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, volume 2, pages 2123–2128, 2003.

[KS03] J. Krivic and F. Solina. Contour Based Superquadric Tracking. In *Knowledge-Based Intelligent Information and Engineering Systems*, volume 2774 of *Lecture Notes in Computer Science (LNCS)*, pages 1180–1186. Springer, Heidelberg, Germany, 2003.

[Kub99] W. Kubinger. *Ein neues Verfahren zur Bewertung regionenbasierter Farbsegmentierungsverfahren (in German)*. PhD thesis, Vienna University of Technology, Vienna, Austria, 1999.

[Lef06] G. Lefrançois. *Psychologie des Lernens (Title of the Original: Theories of Human Learning)*. Springer, Heidelberg, Germany, 4^{th} edition, 2006.

[LF05] V. Lepetit and P. Fua. Monocular Model-Based 3D Tracking of Rigid Objects: A Survey. *Foundations and Trends in Computer Graphics and Vision*, 1(1):1–89, 2005.

[LHM00] C.P. Lu, G.D. Hager, and E. Mjolsness. Fast and Globally Convergent Pose Estimation from Video Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(6):610–622, 2000.

[Lli01] R. Llinas. *i of the vortex: From Neurons to Self.* MIT Press, Cambridge, MA, USA, 2001.

[LLS07] B. Leibe, A. Leonardis, and B. Schiele. Robust Object Detection with Interleaved Categorization and Segmentation. *International Journal of Computer Vision*, 77(1-3):259–289, 2007.

[Low04] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[Mar82] D. Marr. *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Freeman & Co., San Francisco, CA, USA, 1982.

[Mat87] H.R. Maturana. Kognition. In S.J. Schmidt, editor, *Der Diskurs des Radikalen Konstruktivismus*, pages 89–118. Suhrkamp, Frankfurt/Main, Germany, 1987.

[Mat02] M.J. Matarić. Situated robotics. In *Encyclopedia of Cognitive Science.* Nature Publishing Group, Macmillan Reference Ltd, 2002.

[MB03] B.A. Maxwell and S.J. Brubaker. Texture Edge Detection Using the Compass Operator. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, 2003.

[McC00] J. McCarthy. What is artificial intelligence? Online at: `http://www-formal.stanford.edu/jmc` (01/2009), Stanford University, April 2000.

[MCUP02] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *Proceedings of British Machine Vision Conference (BMVC)*, page 384393, Cardiff, United Kingdom, 2002.

[Mey08] Meyers Lexikon. Definition: Kognition – Meyers Lexikon online. Online at: `http://lexikon.meyers.de/wissen/Kognition` (12/2008), 2008.

[MFH88] H. Mandl, H. Friedrich, and A. Hron. Theoretische Ansätze zum Wissenserwerb (Theoretical Approaches to Knowledge Acquisition). In H. Mandl and H. Spada, editors, *Wissenspsychologie (psychology of knowledge)*, pages 123–160. Beltz Psychologie Verlags Union, Weinheim, Germany, 1988.

[Min06] M. Minsky. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind.* Simon & Schuster, New York, NY, USA, 2006.

[MJ91] J. Morton and M.H. Johnson. CONSPEC and CONLERN: A Two-Process Theory of Infant Face Recognition. *Psychological Review*, 98(2):164–181, 1991.

[ML06] E. Margolis and S. Laurence. Concepts. Entry of the *Stanford Encyclopedia of Philosophy*. Online at:
`http://plato.stanford.edu/entries/concepts` (12/2008), Feb. 2006.

[MNW99] R. Möller, B. Neumann, and M. Wessel. Towards Computer Vision with Description Logics: Some Recent Progress. In *Proceedings of the ICCV Workshop on Integration of Speech and Image Understanding*, pages 101–115, 1999.

[MS05] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(10):1615–1630, 2005.

[MTB03] N. Maillot, M. Thonnat, and A. Boucher. Towards Ontology Based Cognitive Vision. In *Computer Vision Systems*, volume 2626 of *Lecture Notes in Computer Science (LNCS)*, pages 44–53. Springer, Heidelberg, Germany, 2003.

[MTB04] N. Maillot, M. Thonnat, and A. Boucher. Towards Ontology Based Cognitive Vision. *Machine Vision and Application (MVA)*, 16(1):33–40, 2004.

[MTH04] N. Maillot, M. Thonnat, and C. Hudelot. Ontology Based Object Learning and Recognition: Application to Image Retrieval. In *Proceedings of the* 16^{th} *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 620–625, 2004.

[MTS$^+$06] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2006.

[Nag74] T. Nagel. What Is It Like To Be A Bat? *The Philosophical Review LXXXIII*, 4:435–450, 1974.

[Nel99] K. Nelson. The Developmental Psychology of Language and Thought. In M. Bennett, editor, *Developmental Psychology – Achievements and Prospects*, chapter 11, pages 185–204. Psychology Press, Philadelphia, PA, USA, 1999.

[Nil98] N. Nilsson. *Artificial Intelligence: A New Synthesis.* Morgan Kaufman, San Francisco, CA, USA, 1^{st} edition, 1998.

[Nør94] T. Nørretranders. *Spüre die Welt – Die Wissenschaft des Bewußtseins (Feel the World – The Science of Consciousness).* Rowohlt, Reinbek bei Hamburg, Germany, 1^{st} edition, 1994.

[NS63] A. Newell and H. Simon. GPS, a Program that Simulates Human Thought. In E. Feigenbaum and J. Feldman, editors, *Computers and Thought.* McGraw-Hill, New York, NY, USA, 1963.

[NS76] A. Newell and H.A. Simon. Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3):113–126, 1976.

[NS96] B. Neumann and C. Schröder. How Useful Is Formal Knowledge Representation for Image Interpretation. In *Proceedings of the ECCV Workshop on Conceptual Descriptions from Images*, pages 58–69, 1996.

[NW03] B. Neumann and T. Weiss. Navigating through Logic-Based Scene Models for High-Level Scene Interpretations. In J.L. Crowley et al., editors, *Proceedings of the International Conference on Computer Vision Systems (ICVS)*, volume 2626 of *Lecture Notes in Computer Science (LNCS)*, pages 212–222. Springer, Heidelberg, Germany, 2003.

[Ohl89] P. Ohler. Kognitive Theorie der Filmwahrnehmung: der Informationsverarbeitungsansatz (Cognitive Theory of Movie Perception: the Information Processing Approach). In K. Hickethier and H. Winkler, editors, *Filmwahrnehmung*, pages 43–58. Edition Sigma, Berlin, Germany, 1989.

[OL90] D.N. Osherson and H. Lasnik, editors. *An Invitation to Cognitive Science.* MIT Press, Cambridge, MA, USA, 1990.

[Pal99] S.E. Palmer. *Vision Science – Photons to Phenomenology.* The MIT Press (Bradford Books), Cambridge, MA, USA, 1999.

[Pan07] J. Panksepp. Simulating the Primal Emotions of the Mammalian Brain: The Affective Feelings of Mental Life and Implications for AI-Robotics. DVD Video Footage of the ENF – The 1^{st} International Engineering & Neuro-Psychoanalysis Forum, July 2007. Institute of Computer Technology, Vienna University of Technology.

[PBGM95] D.R. Proffitt, M. Bhalla, R. Gossweiler, and J. Midgett. Perceiving Geographical Slant. *Psychonomic Bulletin & Review*, 2(4):409–428, 1995.

[Pia54] J. Piaget. *The Construction of Reality in the Child.* Basic Books, 1954.

[Pin05] A. Pinz. Object Categorization. *Foundations and Trends in Computer Graphics and Vision*, 1(4):255–353, 2005.

[Pla04] Plato. Parmenides. In *Sämtliche Werke – Band 3.* rowohlts enzyklopädie, Reinbek bei Hamburg, Germany, 35^{st} edition, 2004.

[PLF05] J. Pilet, V. Lepetit, and P. Fua. Real-Time Non-Rigid Surface Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 822–828, Washington, DC, USA, 2005.

[Pop35] K.R. Popper. *Logik der Forschung (The Logic of Scientific Discovery).* Springer, Vienna, Austria, 1935.

[Pop96] K.R. Popper. *The Myth of the Framework. In Defence of Science and Rationality.* Taylor & Francis Ltd., New York, NY, USA, 1996.

[PS01] R. Pfeifer and C. Scheier. *Understanding Intelligence.* MIT Press, Boston, MA, USA, 2001.

[PSBE03] D.R. Proffitt, J. Stefanucci, T. Banton, and W. Epstein. The Role of Effort in Perceiving Distance. *Psychological Science*, 14(2):106–112, 2003.

[PU03] W. Pidcock and M. Uschold. What are the Differences Between a Vocabulary, a Taxonomy, a Thesaurus, an Ontology, and a Meta-Model? Online at: `http://www.metamodel.com` (12/2008), Jan. 2003.

[Pup13] Pupils of A. Meinong, editor. *A. Meinong's Gesammelte Abhandlungen (Collected Works)*, volume 2: Abhandlungen zur Erkenntnistheorie und Gegenstandstheorie (Works on epistemology and object-theory). J.A. Barth, Leipzig, Germany, 1913. Reprinted in: Alexius Meinong Gesamtausgabe (A. Meinong complete edition), Volume 2, Graz (1971).

[Put60] H. Putnam. Minds and Machines. In S. Hook, editor, *Dimensions of Mind*, pages 148–180. NYU Press, New York, NY, USA, 1960.

[Pyl01] Z. Pylyshyn. Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80:127–158, 2001.

[Qui53] W.v.O. Quine. *From a Logical Point of View.* Harvard University Press, Cambridge, MA, USA, 1953.

[Qui69] W.v.O. Quine. *Epistemology Naturalized. In: Ontological Relativity and Other Essays.* Columbia University Press, New York, NY, USA, 1969.

[RCC92] D.A. Randell, Z. Cui, and A.G. Cohn. A Spatial Logic Based on Regions and Connection. In B. Nebel, C. Rich, and W. Swartout, editors, *Proceedings of the* 3rd *International Conference on Knowledge Representation and Reasoning*, pages 165–176, Los Altos, CA, USA, 1992. Morgan Kaufmann.

[RMG$^+$04] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic Objects in Natural Categories. In D.A. Balota and E.J. Marsh, editors, *Cognitive Psychology*, Key Readings in Cognition, chapter 29, pages 448–471. Psychology Press, New York, NY, USA, 2004.

[RN03] S. Russell and P. Norvig. *Artificial Intelligence – A Modern Approach.* Prentice Hall, Pearson Education, Upper Saddle River, NJ, USA, 2nd edition, 2003.

[Ros73] E. Rosch. Natural Categories. *Cognitive Psychology*, 4:328–350, 1973.

[Ros77] E. Rosch. Human Categorization. In N. Warren, editor, *Advances in cross-cultural psychology*, volume 1. Academic Press, London, United Kingdom, 1977.

[RT99] M. Ruzon and C. Tomasi. Color Edge Detection with the Compass Operator. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 160–166, 1999.

[Rum80] D. Rumelhart. Schemata: The Building Blocks of Cognition. In R. Spiro, B. Bruce, and W. Brewer, editors, *Theoretical issues in reading comprehension: Perspectives from Cognitive Psychology, Linguistics, Artificial Intelligence and Education*, pages 33–58. Lawrence Erlbaum Associates, Philadelphia, PA, USA, 1980.

[Rus99] J. Russell. Playing a Passing Game: Rationalism, Empiricism, and Cognitive Development. In M. Bennett, editor, *Developmental Psychology – Achievements and Prospects*, chapter 14, pages 253–271. Psychology Press, Philadelphia, PA, USA, 1999.

[SB90] F. Solina and R. Bajcsy. Recovery of Parametric Models from Range Images: The Case for Superquadrics with Global Deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 12(2):131–147, 1990.

[SBV06] M.J. Schlemmer, G. Biegelbauer, and M. Vincze. An Integration Concept for Vision-Based Object Handling: Shape-Capture, Detection and Tracking. In N. Zheng, X. Jiang, and X. Lan, editors, *Advances in Machine Vision, Image Processing, and Pattern Analysis – International Workshop on Intelligent Computing in Pattern Analysis/Synthesis (IWICPAS), Xian, China*, volume 4153/2006 of *Lecture Notes in Computer Science (LNCS)*, pages 215–224. Springer, Heidelberg, Germany, 2006.

[SBV07] M. Schlemmer, G. Biegelbauer, and M. Vincze. Rethinking Robot Vision – Combining Shape and Appearance. *International Journal of Advanced Robotic Systems (ARS)*, 4(3):259–270, 2007.

[SC05] A. Sloman and J. Chappell. The Altricial-Precocial Spectrum for Robots. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 1187–1192, Edinburgh, United Kingdom, 2005.

[Sch87] S.J. Schmidt. Der Radikale Konstruktivismus: Ein neues Paradigma im interdisziplinären Diskurs (Radical Constructivism: A New Paradigm in Interdisciplinary Discourse). In S.J. Schmidt, editor, *Der Diskurs des Radikalen Konstruktivismus*, pages 11–88. Suhrkamp, Frankfurt/Main, Germany, 1987.

[Sch98] N. Schneider. *Erkenntnistheorie im 20. Jahrhundert – Klassische Positionen (Epistemology in the 20th century)*, volume 9702 of *Reclams Universal-Bibliothek*. Reclam, Stuttgart, Germany, 1998.

[Sch08] M.J. Schlemmer. There's more to vision than meets the eye – Erkenntnistheoretische Konzepte des 20. Jahrhunderts im Lichte der visuellen Wahrnehmung (in German). Master's thesis, University of Vienna, Vienna, Austria, 2008.

[SDF04] A. Shahrokni, T. Drummond, and P. Fua. Texture Boundary Detection for Real-Time Tracking. In T. Padla and J. Matas, editors, *Proceedings of the 8th European Conference of Computer Vision (ECCV)*, volume 3022, pages 566–577, 2004.

[SDM02] E. Di Sciascio, F.M. Donini, and M. Mongiello. Structured Knowledge Representation for Image Retrieval. *Journal of Artificial Intelligence Research*, 16:209–257, 2002.

[Sea97] J. Searle. Reductionism and the Irreducibility of Consciousness. In N. Block, O. Flanagan, and G. Güzeldere, editors, *The Nature of Consciousness*, chapter 27, pages 451–460. MIT Press, Cambridge, MA, USA, 1997.

[Slo01] A. Sloman. Evolvable Biologically Plausible Visual Architectures. In T. Cootes and C. Taylor, editors, *Proceedings of British Machine Vision Conference (BMVC)*, pages 313–322, Manchester, United Kingdom, 2001.

[Slo02] A. Sloman. Getting Meaning Off the Ground: Symbol-grounding *vs* Symbol-tethering. Online at: `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/` (12/2008), March 2002.

[Slo06] A. Sloman. How to Put the Pieces of AI Together Again. In *Proceedings of the* 21st *AAAI Conference on Artificial Intelligence*, Boston, MA, USA, July 2006. AAAI Press.

[Slo07] A. Sloman. Consciousness in a Multi-layered Multi-functional Labyrinthine Mind. Poster-Presentation at PAC '07 (Conference on Perception, Action and Consciousness), July 2007. Online at: `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/` (12/2008).

[Slo08a] A. Sloman. Could a Child Robot Grow Up To be A Mathematician And Philosopher? Invited Talk at the University of Liverpool. Online at: `http://www.cs.bham.ac.uk/research/projects/cogaff/talks/` (12/2008), Jan. 2008.

[Slo08b] A. Sloman. Kantian Philosophy of Mathematics and Young Robots. In M. Autexier et al., editor, *Intelligent Computer Mathematics*, volume 5144 of *Lecture Notes in Computer Science (LNCS)*, pages 558–573. Springer, Heidelberg, Germany, 2008.

[Sol08] M. Solms. What is the "Mind"? A Neuro-Psychoanalytical Approach. In D. Dietrich, G. Fodor, G. Zucker, and D. Bruckner, editors, *Simulating the Mind – A Technical Neuropsychoanalytical Approach*, pages 115–122. Springer, Vienna, Austria, 2008.

[SP06] G. Schweighofer and A. Pinz. Robust Pose Estimation from a Planar Target. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(12):2024–2030, 2006.

[Spe98] E.S. Spelke. Nativism, Empiricism, and the Origins of Knowledge. *Infant Behavior and Development*, 21(2):181–200, 1998.

[ST04] M. Solms and O. Turnbull. *Das Gehirn und die innere Welt – Neurowissenschaft und Psychoanalyse (Original Title: The brain and the inner world).* Patmos, Düsseldorf, Germany, 2004.

[Ste99] R. Sternberg. Looking Back and Looking Forward on Intelligence: Toward a Theory of Successful Intelligence. In M. Bennett, editor, *Developmental Psychology – Achievements and Prospects*, chapter 16, pages 289–308. Psychology Press, Philadelphia, PA, USA, 1999.

[SV05] M. Schlemmer and M. Vincze. Texture Edge Detection Using Statistical Feature Matrices. In D. Chetverikov, L. Czuni, and M. Vincze, editors, *Proceedings of the Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition (HACIPPR/OAGM)*, pages 335–342, Veszprém, Hungary, 2005. Österreichische Computer Gesellschaft.

[SV07] M.J. Schlemmer and M. Vincze. On an Ontology of Spatial Relations to Discover Object Concepts in Images. In 6th *EUROSIM Congress on Modelling and Simulation*, Ljubljana, Slovenia, 2007.

[SV08a] M.J. Schlemmer and M. Vincze. A Functional View on "Cognitive" Perceptual Systems Based on Functions and Principles of the Human Mind. In D. Dietrich, G. Fodor, G. Zucker, and D. Bruckner, editors, *Simulating the Mind – A Technical Neuropsychoanalytical Approach*, pages 302–319. Springer, Vienna, Austria, 2008.

[SV08b] M.J. Schlemmer and M. Vincze. Abstraction, Ontology and Task-Guidance for Visual Perception in Robots. In A.G. Cohn, D.C. Hogg, R. Möller, and B. Neumann, editors, *Logic and Probability for Scene Interpretation*, number 08091 in Dagstuhl Seminar Proceedings. Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, Germany, 2008. Online at: `http://drops.dagstuhl.de/opus/volltexte/2008/1608` (12/2008).

[SVFB07] M.J. Schlemmer, M. Vincze, and B. Favre-Bulle. Modelling the Thing-In-Itself – A Philosophically Motivated Approach to Cognitive Robotics. In L. Berthouze, C.G. Prince, M. Littman, H. Kozima, and C. Balkenius, editors, *Proceedings of the* 7th *International Conference on Epigenetic Robotics (EpiRob) – Modeling Cognitive Development in Robotic Systems*, number 134 in Lund University Cognitive Studies (LUCS), pages 149–156, Piscataway, NJ, USA, 2007.

[SWH$^+$06] A. Sloman, J. Wyatt, N. Hawes, J. Chappell, and G. Kruijff. Long Term Requirements for Cognitive Robotics. In *Proceedings of the* 5th *International Cognitive Robotics Workshop (CogRob)*, July 2006.

[SWM04] M.K. Smith, C. Welty, and D.L. McGuinness. OWL Web Ontology Language Guide – W3C Recommendation. Online at: `http://www.w3.org/TR/owl-guide/` (12/2008), Feb 2004.

[TG80] A. Treisman and G. Gelade. A Feature Integration Theory of Attention. *Cognitive Psychology*, 12:97–136, 1980.

[Tho02] M. Thonnat. Knowledge-Based Techniques for Image Processing and for Image Uunderstanding. *Journal de Physique 4*, 12(1):189–236, 2002.

[TJ05] M. Tuceryan and A.K. Jain. Texture Analysis. In C. Chen and P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision*, chapter 2.1, pages 235–276. World Scientific Publishing Company, 3rd edition, Jan. 2005.

[TK04] G. Taylor and L. Kleeman. Integration of Robust Visual Perception and Control for a Domestic Humanoid Robot. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 1010–1015, Sendai, Japan, 2004.

[Ver04] D. Vernon. Cognitive Vision. Technical report, ECVision – The European Network for Cognitive Computer Vision Systems, 2004. Online at: `http://www.eucognition.org/ecvision/about_ecvision/Cognitive_Vision.pdf` (12/2008).

[vG92] E. von Glasersfeld. Cybernetics – Declaration of the American Society for Cybernetics. In C.V. Negoita, editor, *Cybernetics and Applied Systems*, chapter 1, pages 1–5. Marcel Dekker, New York, NY, USA, 1992.

[Vie06] M. Viezzer. *Autonomous Concept Formation: An Architecture-Based Analysis*. PhD thesis, University of Birmingham, School of Computer Science, Birmingham, United Kingdom, 2006.

[Vol81] G. Vollmer. *Evolutionäre Erkenntnistheorie (Evolutionary Epistemology)*. S. Hirzel, Stuttgart, Germany, 3^{rd} edition, 1981.

[VSGA05] M. Vincze, M. Schlemmer, P. Gemeiner, and M. Ayromlou. Vision for Robotics – A Tool for Model-Based Object Tracking. *IEEE Robotics & Automation Magazine – Special Issue: "Software Packages for Vision-Based Control Of Motion"*, 12(4):53–64, 2005.

[VTR93] F. Varela, E. Thompson, and E. Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, Cambridge, MA, USA, 1993.

[Wac08] L. Wachsmann. The OWLPropViz plug-in. Online at: `http://www.wachsmann.tk/owlpropviz/` (12/2008), 2008.

[WC92] C.-M. Wu and Y.-C. Chen. Statistical Feature Matrix for Texture Analysis. *Graphical Models and Image Processing (CVGIP)*, 54(5):407–419, 1992.

[WCF89] J. Wolfe, K.R. Cave, and S.L. Franzel. Guided Search: An Alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419–433, 1989.

[Win70] P.H. Winston. *Learning Structural Descriptions from Examples*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1970.

[Wit98] L. Wittgenstein. *Logisch-philosophische Abhandlung – Tractatus logico-philosophicus*. Suhrkamp, Frankfurt/Main, Germany, 1^{st} edition, 1998. Translation taken from: `http://schulers.com` (12/2008).

[WNR+06] H. Wang, N. Noy, A. Rector, M. Musen, T. Redmond, D. Rubin, S. Tu, T. Tudorache, N. Drummond, M. Horridge, and J. Seidenberg. Frames and OWL Side by Side. In *Proceedings of the 9^{th} International Protégé Conference*, Stanford, CA, USA, 2006.

[Wol94] J. Wolfe. Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.

[Wor04] World Wide Web Consortium (W3C). OWL Web Ontology Language Overview. Online at: `http://www.w3.org/TR/owl-features/` (12/2008), Feb. 2004.

[Wuk90] F.M. Wuketits. *Evolutionary Epistemology and Its Implications for Humankind.* SUNY Press, Albany, NY, USA, 1990.

[YKBK05] Y. Youngrock, A. Kosaka, P. Jae Byung, and A.C. Kak. A New Approach to the Use of Edge Extremities for Model-based Object Tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1871–1877, Barcelona, Spain, 2005.

[Yov08] Y. Yovell. Return of the Zombie – Neuropsychoanalysis, Consciousness, and the Engineering of Psychic Functions. In D. Dietrich, G. Fodor, G. Zucker, and D. Bruckner, editors, *Simulating the Mind – A Technical Neuropsychoanalytical Approach*, pages 251–258. Springer, Vienna, Austria, 2008.

[Zie01] T. Ziemke. The Construction of 'Reality' in the Robot. *Foundations of science*, 6(1-3):163–233, 2001.

[Zil07] M. Zillich. *Making Sense of Images: Parameter-Free Perceptual Grouping.* PhD thesis, Vienna University of Technology, Vienna, Austria, 2007.

[ZTB04] N. Zlatoff, B. Tellez, and A. Baskurt. Image Understanding and Scene Models: A Generic Framework Integrating Domain Knowledge and Gestalt Theory. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 2355–2358, 2004.

Curriculum vitae

Personal Information

Matthias J. Schlemmer

Academic degrees: Mag.phil. Dipl.-Ing.

Nationality: Austrian

DOB: 23/07/1979 in Vienna

Address: Schönburgstraße 9/3A
1040 Vienna, Austria

E-Mail: matthias@schlema.com

Work experience

07/2004–	**Automation & Control Institute**, Vienna Univ. of Technology Research Assistant
07/2002–12/2003	**ZT Prentner IT GmbH**, Vienna, Austria Project Assistance
09/2000–01/2001	**Dankl+Partner (consultants of a large pharmaceut. comp.)** Vienna, Austria Maintenance
08/2000	**Institute for Economic- and Business Management** Montanuniversität Leoben, Austria Summer Internship
10/1998–12/2006	**Austrian Red Cross**, Vienna Paramedic and ambulance driver, head of educational division Volunteer work

Education

07/2004–	**Doctorate (PhD) studies of Electrical Engineering** Vienna University of Technology, Austria Specialization: Cognitive aspects of computer vision and robotics
10/2002–06/2008	**Master of Philosophy** (with distinction) University of Vienna, Austria Specialization: Epistemology, philosophy of mind
10/1998–06/2004	**Master of Science in Electrical Engineering** (with distinction) Vienna University of Technology, Austria Specialization: Computer Engineering

10/1997 **Qualification to work as a paramedic**
Austrian Red Cross, Vienna

09/1989–05/1997 **BG XVIII (comprehensive school)**
Vienna, Austria
Matura (school leaving examination) with distinction

Selected Publications

Schlemmer M.J., Vincze M.: *A functional view on "cognitive" perceptual systems based on functions and principles of the human mind.* In: Dietrich, D.; Fodor, G.; Zucker, G.; Bruckner, D. (Eds.): "Simulating the Mind – A Technical Neuropsychoanalytical Approach", p. 302-319; Springer, October 2008.

Schlemmer M.J., Vincze M., Favre-Bulle B.: *Modelling the Thing-In-Itself - a philosophically motivated approach to cognitive robotics.* International Conference on Epigenetic Robotics (EpiRob); Piscataway, NJ, USA, November 2007

Schlemmer M.J., Vincze M.: *On an ontology of spatial relations to discover object concepts in images.* EUROSIM Congress on Modelling and Simulation; Ljubljana, Slovenia, September 2007

Schlemmer M.J., Biegelbauer G., Vincze M.: *Rethinking Robot Vision - Combining Shape and Appearance.* In: Journal for Advanced Robotic Systems (ARS); September 2007

Schlemmer M.J., Biegelbauer G., Vincze M.: *An Integration Concept for Vision-Based Object Handling: Shape-Capture, Detection and Tracking.* International Workshop on Intelligent Computing in Pattern Analysis/Synthesis; Xi'an, China, August 2006

Schlemmer M.J., Vincze M.: *Texture Edge Detection Using Statistical Feature Matrices.* Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition; Veszprem, Hungary, May 2005

Vincze M.; Schlemmer M.J.; Gemeiner P.; Ayromlou M.: *Vision for Robotics: a tool for model-based object tracking.* In: IEEE Robotics & Automation Magazine, IEEE Volume 12, Issue 4; December 2005

Vienna, March 3, 2009