Abstraction, ontology and task-guidance for visual perception in robots

Matthias J. Schlemmer and Markus Vincze

Vienna University of Technology, Automation and Control Institute
1040, Vienna, Gusshausstrasse 27-29/E376, Austria
schlemmer@acin.tuwien.ac.at

Abstract. In this paper we summarise theoretical work on visual perception for cognitive robotics. We study the use of logic-based ontologies for situated perception, motivated by cognitive science and philosophy. General considerations on scope and aims of "cognitive vision" is given where a *functional* position is taken. The use of symbol-based reasoning following this approach and the central importance of the mapping from low-level sensor data to higher-level symbols is outlined. Special focus is given w.r.t. the localisation of different mechanisms (subsymbolic, symbolic) during perception and cognition and *cognitive functions* that seem important are named. A preliminary implementation approach for mapping simple Gestalts to object concepts and a general discussion on the use of logic for perception conclude the paper.

Keywords. cognitive robotics, visual perception, cognitive functions, ontology, symbol-based AI

1 Introduction

We study the use of ontologies in the field of visual perception for cognitive robotics. Core questions that correlate with the topic of this seminar are therefore: Is logic interesting for robotic vision? If yes, to what extent? The same questions hold for probability theory and of course for their possible combinations. We will show that this implies further general considerations which we will, however, only elaborate shortly for the sake of compactness. Furthermore, this paper is largely theoretical and only provides some first implementation results to show the feasibility of the approach.

We will see that questions relevant for robotic vision are highly correlated to well-known questions from the field of artificial intelligence; and due to the relation of artificial intelligence and cognitive science, issues arising in the discussion of human cognition also arise in computer science and robotics. The aim of this paper is to show that the use of "GOFAI-symbols"[1] still does make sense (despite their infamous history of limitations) – and that their storage in a logic-based ontology can be shown to provide a promising approach for cognitive

[1] Good Old Fashioned Artificial Intelligence

Dagstuhl Seminar Proceedings 08091
Logic and Probability for Scene Interpretation
http://drops.dagstuhl.de/opus/volltexte/2008/1608

robotics and especially for its link to visual perception. Here, the question that interests us concerns how to step from "uncertain" data to symbols which can further on be used to guide higher-level tasks, such as planning or predicting. To this end, we identify some *cognitive functions* that we judge to be important for so-called "cognitive vision". Furthermore, we will show that a logic-based ontology of object concepts might play the role of the "glue" between these functions on the one hand, and additionally the core role for transition from perceptual data to symbols on the other.

After some (very compactly presented) general considerations clarifying some notions in Section 2, we suggest our view on cognition from a functional point of view (Section 3). The section about ontology modelling (Section 4) will be kept small as this seminar has other contributions that get much more into detail with respect to the modelling issues and we will rather focus on some examples of the bridge between vision output and ontological knowledge representation. In the last Section (5) we critically discuss the use of logic for perception from a more general viewpoint.

2 General considerations

In order to clarify why we are attempting to use logic-based ontologies for "cognitive vision" we shortly explain the overall view and the goals that our work addresses. The term "cognitive" has lately been used as a buzzword in the engineering sciences for pointing to the fact that we need to tackle issues involving more higher-level functions of reasoning (or "thinking"); for perception in robots we interpret this as the demand to ultimately remove the outdated sense-compute-act cycle where perception is *only* concerned with the sense-part, detached from the other processes. There are highly elaborated techniques in computer vision dealing with specific problem statements. However, what vision for cognitive robotics truly needs is the look at – and better: the integration of – higher-level functions (see below).

Before looking closer at our view and the role of logic in it, we need to shortly state what we mean by "cognitive" or "intelligent"[2]. In our approach, intelligence can only be defined from the third person perspective, meaning that an agent (robot or human) behaves intelligently when we, as observers, judge it this way. This is, for sure, a strictly behaviourist view that is – for robotics – a means to evaluate the feasibility of different approaches (as we are lacking the possibility of a first-person view and therefore the possibility to judge *my* behaviour in a *conscious* manner).

Cognition, on the other hand, is quite a tricky term and there are totally different views on what it actually means. We take a rather abstract and general viewpoint: cognition is the superordinate concept of our perceptions and actions allowing for functional fit of the system (or in the radical constructivist term: "viability"). Without the aim to dive too deep into discussions from cognitive

[2] For a more detailed elaboration on these notions including "consciousness", please refer to [1].

science, we shortly state that the approach to the phenomenon "cognition" has correctly been shifted in the last years from "pure" symbol manipulation to more extensive views involving the dynamic, situated and embodied perception and action complex that aims at surviving 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 (also cf. [2]).

2.1 Anthropomorphism?

As could be seen in the previous paragraphs, a mixture of anthropomorphic terms has been introduced to engineering. We believe that it is imperative to be cautious on where to draw the analogies from robots to humans (not only because of ethical reasons). Our statement about the notion of "cognition" in the previous Section already shows our main approach: Without wanting to go into details of the *implementation layer of cognition*, it is probably better to focus on the "functions" that the system needs to have in order to act in a way cognitive robotics usually aim at: to serve purposefully at the humans' side, who finally decides about the viability of the artificial system he created.

Artificial intelligence started off with the strictly so-called "cognitivistic" view that all human brain action is problem solving with symbols. This believe has been put into perspective by history and for sure there are capabilities of humans that cannot be reduced to the manipulation of symbols. However, it seems that humans do in fact have the capability of abstract logic-based reasoning. With [3] we therefore hold that it seems that evolution at some point found that using symbols rather than raw sensor data is more efficient for higher level planning of behaviour or understanding the environment.

These considerations allow us now to depict our view on how to tackle the issues that arise when trying to combine higher-level functions well-known to artificial intelligence (planning, problem solving) with the specialized but too narrowly aimed techniques of computer vision.

2.2 The bigger picture

A feasible approach for an inspiration for cognitive robotics that is not too anthropomorphic yet does still take it's approach from human perception and cognition might be the *functional level*, as already indicated further above. In the next section we name them and try to put them into a cycle that should show a possible explanation on how humans built up their specific view on the world – inspired by the tradition of radical constructivism, cf. [4] for an account w.r.t. robotics. Radical constructivism highlights the active building-up of knowledge (instead of the passive receiving of "information out there"), the tendency of a "living" system towards viability and the goal of cognition as personal organisation of the experiential world (as opposed to the discovery of an independently existing ontological reality).

As we will outline, we believe that for the functions that an artificial cognitive system should adopt, there is a need for an ontological knowledge repository. Though we believe that this ontology is actively built up by humans from the very beginning of his/her life on, we start with investigating how to use a "grounded" ontology that is pre-given by the system's designer. Underlying is the belief that without knowing how to *use* an ontology it is hard to design mechanisms how to built it up. Therefore, the more concrete question we are investigating is how we can combine such an ontological ("symbol-based" or "logical") knowledge repository with the uncertainties of perceiving – with other words: How to link the bottom-up delivered vision input (which is very often statistically and probabilistically given) with the symbol-world in which higher-level cognitive functions such as predicting or planning might take place.

The bigger picture that we believe is in place in humans, now finally goes like this (we are of the opinion that we therefore might implement a similar architecture in artificial agents on a *functional level* as well):

1. There are multiple layers (low-level to high-level) where "cognition" (of which perception is an integral part) takes place.
2. Depending on situation and task, there are different mechanisms taking place – either subsymbolic or symbolic.
3. Humans are able to switch from one mechanism to the other if the situation affords this.
4. Therefore, items having been perceived in a subsymbolic or probabilistic manner might get symbols for further symbolic computation if necessary.

A simple example should clarify this: Looking at Van Gogh's famous self-portrait from 1887, we might first only see a man with a red beard. If we don't know that it's 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 museum visit or plans for future cultural trips. 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.

Needless to say that the "amount of symbols generated" is further dependent on the individual's history: Being an expert in the history of art for sure influences what is usually seen as "way of perceiving" – which we could call here "dissemination of activation throughout the layers". It is far beyond the focus of this paper to investigate that this might involve a pragmatic explanation for the need of consciousness: Something gets consciously in our minds whenever a subsymbolic arrangement becomes a symbol. At this point, this symbol can be used for (often language-based) reasoning and more generally "thoughts about that thing".

All of this, of course, also again underlines that instead of a cycle like sense – compute – act, there is rather and at least a cycle like sense – predict and integrate cues – hypothesise – decide – act; and huge interaction and "mechanism-changing" between the steps as well. So where to locate logic, where probability? And even more interesting: How to combine them? We believe that logic can be nicely seen as those symbols allowing for higher-level thoughts that get fed by lower-level (e.g., probabilistic) input – and for perception which is our field of work, this means that on any layer there might be ontological representations as well as subsymbolic items handled by "cognition". Before showing our first attempt for a vision-based object concept ontology, let's have a closer look on a possible minimal set of functions for a cognitive robot.

3 Functions

The view presented in the last Section raises the questions on how to address these different approaches and how to represent the "symbols". We are not supposing that all thought is symbol manipulation, however, obviously a lot of (at least seemingly) meaningful ("intelligent") human behaviour seem to be *explainable* by symbol manipulation. It seems we are able to solve problems by thinking about general terms, to link abstract concepts to given concrete appearances and to generalise instances to object classes. It is in this sense what we mean by "symbols" (and which we consequently want to link with low-level vision techniques). They are visual object concepts, meaning that they are classes of things that hold for a plurality of varying instances, which – philosophically spoken – contains the things-in-theirselves, the constituting substrates of object classes. For sure, there are properties of object classes that can not only be extracted by visual input (e.g., what constitutes a pepper mill that looks like a wine bottle is its *function* of being able to crush pepper corn, not its appearance as bottle).

The point is that whether human cognition "really" does *imply* the use of a "symbolic layer" or not is irrelevant. From an engineer's point of view being on the search for intelligently behaving robots ("cognitive robots") that are able to fulfill their tasks even if they are not pre-programmed and so to achieve viable solutions, it is the *functional level* which is important – and those functions can be best explained by using symbol manipulation. So the main question that arises is at which layer we shall switch from "subsymbolic" or "probabilistic" methods to symbols (object concepts in our approach). Does it make sense to draw a strict line or shall we rather look for a multi-layer, multi-connected framework (which, of course not only involves visual perception but rather the whole situated embodiment).

After all these theoretic underpinning, let's have a closer look at the functions in order to choose a feasible starting layer of ontological representations. Fig. 1 shows our ideas of a (minimal?) set of functions that a cognitive robot should implement.

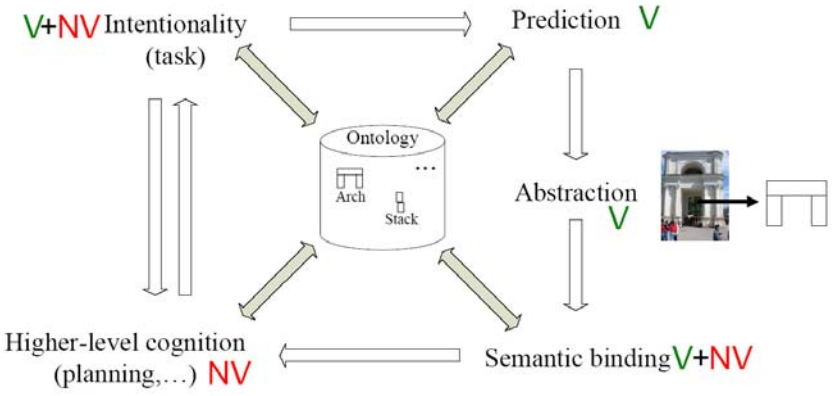


Fig. 1. Using an ontology as "glue" between cognitive functions of a robot. V means "visual function" and NV "nonvisual function" – referring to our approach.

In the diagram, the emphasis lies on the transition between visual functions ("V") and higher-level non-visual ("NV") cognition functions. By this, we refer to the fact that vision should be able to extract information that can be used for further "symbol-based" tasks, such as planning. For us, the use of an ontology for scene interpretation could be the means to achieve this.

Starting from *intentionality* of the system, which could simply be the task of the robot (e.g., get out of the room) there is already an interplay between the visual and the non-visual layer, showing their interconnectedness. Whereas it might be a higher-level goal to "detect a door" to get out, the door itself is a *visual concept* that needs to be *detected*. This is why we are suggesting to use an ontology of known object concepts on which further knowledge can be connected (see later). The importance of intentionality is motivated from the phenomenological philosophy of Edmund Husserl and specifically from A. Meinong's conception [5][3].

The next function is *prediction* which we consider as a highly relevant cognitive function because it primes the system towards what it will perceive. As this is highly related to perception and less to internal thoughts, we consider it a perceptual (and for our approach therefore: visual) function. Some believe that prediction is one – if not "the" – core function of the human brain, cf. [7].

Next comes *abstraction* that allows us to extract qualitative knowledge from quantitative data. Referring to our discussion of the different layers and the different subsymbolic and symbolic mechanisms in Section 2.2, it is here where we locate our chosen layer to switch from subsymbolic to symbolic. Still we have elbowroom here, as we are either applying quite general techniques to extract proto-objects (such as simple closed contours as will be shown in Section 4)

[3] It is far beyond the scope of this paper to further elaborate on the philosophical details of this and the following topics. For a slightly more comprehensive account, please refer to [6].

or more fine-grained techniques that are specialised for detecting, e.g., doors as such. It is here, where specialized generic object recognition tools from computer vision can be used, e.g. [8]. Either way, the output of this step is a qualitative representation of what is in the image (ideally primed by and corresponding to the prediction step).

When now trying to bind semantics to the abstract qualitative information extracted from the image, the ontology gets important again. As we will see later, one of the big advantages of using State-of-the-Art ontologies is the possibility to store very different information in it, enriching the previously defined "visual object concepts" needed for the abstraction step with data on what the object can be used for, what its possible connecting objects are and lots more. To once more apply this metaphor on thoughts about humans: It is here, where all the previous experiences are triggered that allow us to link seen information with similar situations and problem solutions. It is also here, where we are able to smoothly leave the visual domain and dive into what we termed "higher-level cognition" in the diagram. Using knowledge about what we saw, we are able to plan what to do next, ideally ending in a new task and therefore new predictions what to see next – closing the loop.

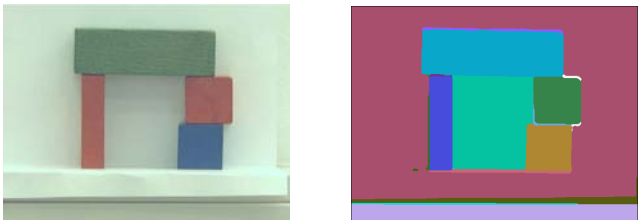
There is still one big point missing in the diagram, namely *generalisation*. There is a huge pile of literature on human concept formation, the reference work is [9]. However, it is still largely unknown *how* humans actually perform this and likewise machine learning techniques are yet not capable of learning abstract concepts from different sources and exemplar distributions in the large-scale manner than would actually be needed. This cognitive function is also not tackled here – our focus lies on the use of pre-given ontological representations within a cognitive robot. We are, however, aware that by circumventing the learning of the concepts with all their rich semantics, we are undermining the whole constructivist idea of personal representations that are bound to one's embodiment.

4 Implementation

We will now shortly outline a first and very preliminary approach to modelling the ontology as well as the interaction with the vision part. This work is twofold, as on the one hand we are interested in the general layout of the object concept ontology. On the other hand, we would like to present work on how the transition from detected low-level image patches to symbols (in terms of object concepts) can be performed.

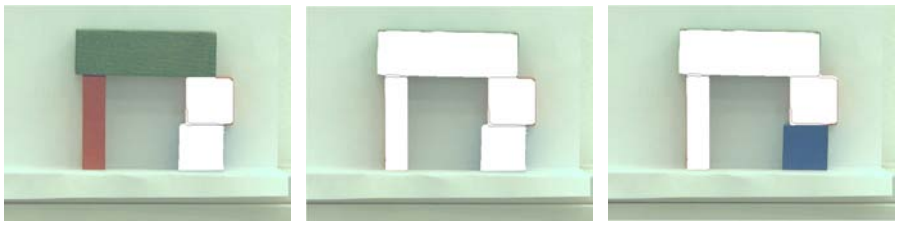
As we are at the moment only dealing with vision, we decided to define the "constituting substrate" of object concepts via the spatial relations of simpler parts. As stated, the big advantage of ontologies lies in the fact that any further information can enrich this basal definition later on. As representation and reasoning system we are using RACER [10].

A stack is therefore defined as two objects, one on top of the other; an arch is the combination of two columns that are not adjacent and a top-bar that



(a) A real-world example arch. (b) Segmentation with [11].

Fig. 2. Vision pre-processing for the transition from low-level perceptual data to symbols in terms of visual object concepts.



(a) Found stack. (b) Best arch hypothesis. (c) Another hypothesis.

Fig. 3. Using the ontological definition for finding a real world arch. (a) shows a stack found which could serve as column for the arch. (b) is the best hypothesis (judged by the human observer), and (c) is correctly found yet would be judged not to be perfect.

connects them on top. Abstraction gets important when thinking about arches whose columns are made of more than one part. Therefore, we allow any number of contributing parts as long as simple physical laws are obeyed (a stack is only a stack if the upper object's barycenter is within the width of the lower object).

Referring to the discussion on where to start off with symbols (see Section 2.2), we would like to give two scenarios. The first one assumes vision to be a) perfect and b) perfectly tuned to the current task. E.g., if we are provided with an "arch-detector", we can immediately feed our higher-level cognition (e.g. a planner aiming to drive through the arch) with precise location information in image 2(a).

The second scenario shows our approach to manage the sensor-to-symbol transition a few layers down, namely by using an ontological representation of "arch" which only affords the spatial relations of its parts as explained above. E.g., we might only detect patches of similar colour, for example by a segmentation algorithm based on a graph cut method, as in Fig. 2(b) with [11]. To be able to find the arch in this case, we are performing the following steps:

1. Extract the patches and treat them as possible object parts
2. Provide the reasoner with these parts
3. Apply rules that specify object concepts as particular arrangements of simple objects

(a) A door in our office.

(b) Closure detection using [12].

Fig. 4. Vision pre-processing for the transition from low-level perceptual data to symbols in terms of visual object concepts.

4. Retrieve the hypotheses found by the ontological reasoner

The result of the reasoning is shown in Fig. 3. Following these first steps, we are planning to implement further techniques to weigh the hypotheses or to use them as input for higher-level cognition tasks.

Fig. 4 shows a different example, where the pre-processing is deliberately very differently, namely by using a Gestalt-based grouping tool [12], the output of which is depicted in Fig. 4(b). Some of the possible arch hypotheses found are shown in Fig. 5. Especially interesting is Fig. 5(c) as this hypothesis points to a very important fact: It shows us that we applied a too weak description of a "good arch". Here we see that we are need of investigations about what constitutes an object concept; what the defining parts are and why humans judge some instances as "good exemplars" and some as "bad exemplars". For us roboticists, we are in the situation of having to neatly define what the arch needs to look like *for* the robot *in reference to* its tasks.

5 Discussion

Hopefully we could motivate our use of ontologies for situated perception of a cognitive robot so far. In this Section, we would like to meet some of the objections of the use of logic for scene interpretation on a quite general level. Afterwards, we give a short conclusion.

5.1 Does using logic for scene interpretation make sense at all?

First of all, the use of a logic-based knowledge representation always raises the philosophical question of whether there is any ontological reality "out there"

(a) Best hypothesis for arch found.

(b) Another very good hypothesis.

(c) Bad hypothesis.

Fig. 5. Using the ontological definition of arch for finding a door in an office (from Fig. 4). Subfig. 5(c) shows that a more restrictive definition of arch might be useful in order to sort such "meaningless" solutions out.

at all. We take here a position that tries to combine the view of Nicolai Hartmann [13] and the view of radical constructivism. For the former, the epistemological and the ontological questions are highly connected. The gnoseological problem involves the ontological one and *cannot* be treated without it – but it's not the same. The latter, the seemingly contradictory view of radical constructivism, excludes questions about an independently existing reality in order to focus on concepts like autopoiesis and viability. We take a middle way as we do believe that *for robotics*, we are forced to assume an independent reality of objects that we can provide the robot with (at least a minimal bootstrapping set), but we admit that in order to achieve the overall *goal* of the system, namely autonomous and viable behaviour, we need to have this "reality" *tuned* in an agent-specific manner.

To be more specific: The symbols with which the system should operated, are created by creating the "world of the robot", i.e., the agent-specific representation of objects. I think this point is often not made clear, that the use of constructivist approaches does not necessarily exclude the use of symbolic manipulation. What is in fact to be rejected from a constructivist view, is however that those symbols are a 1:1 account of an independently existing outside world.

One very important and justified objection can be posed once the general assumption of using symbols is accepted: The World is not (that strictly) logical or to put it more explicitly: not everything can be presented with logic statements → Therefore it is senseless to represent the world this way. To clarify our viewpoint: It is true that the world cannot totally be represented as symbols (this is what the history of AI showed us), however, we think that there *are* perceptual object concepts (visual ones in our case) and that they *are* in fact connected to higher-level information that is used by us. For example, thinking about a car

does automatically raise ideas, thoughts and associations about what it means to own one, how it's like to drive one, etc. *without* demanding a specific appearance in front of our eyes. So the representation of a "subset" of the world in symbols is possible, and exactly here we see the transition for vision techniques to higher-level cognition. Additionally, we still have a quite down-to-earth goal in our minds: the limited domain of robotics, ideally for home robotics some day.

For the sake of completeness, we name the advantages although they are probably well-known. (Technical) ontologies (like OWL [14] or the RACER System [10] that we currently use, see Section 4) allow for:

- Automatic reasoning (consistency checking and deduction of implicit knowledge)
- Storing semantic knowledge (in the sense of, e.g., affordance-, task related or functional information) along with the concepts
- Abstraction and interconnection of concepts – e.g., it is possible to include very different information, e.g., the current bodily state so to emulate findings from psychology according to which the body signals highly influence the way of perceiving (e.g., [15, 16])
- Human language definitions that help during development and testing
- Finally, with the help of ontologies we are able to separate "common sense knowledge" from vision processing

5.2 Conclusion

In this paper we tried to outline our understanding of using logic for situated perception in cognitive robotics. Therefore, we motivated the use of symbols in terms of "object concepts" stored in a logic-based ontology. Scene interpretation does, in our view, not stop with the detection of features but contrarily, is in deep need of involving higher-level *cognitive functions*, of which a necessary subset has been identified. General consideration on using the "human metaphor" for perception in robots as well as our view on the layers on which a necessary transition from sensor data to symbols is performed, has been given. Finally, a preliminary approach for performing this transition on quite a low level by using the ontological definition of object concepts has shown that logic has its right in the search of situated perception of cognitive robots.

Acknowledgements

This work is has been funded by the European Commission's Sixth Framework Programme under contract no. 029427 as part of the Specific Targeted Research Project XPERO (Robotic Learning by Experimentation) as well as by the Austrian Science Foundation under the grant #S9101 (Cognitive Vision).

References

1. 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. Springer, Vienna/Austria (To Appear 2009) 302–319
2. Clark, A.: Mindware – An Introduction to the Philosophy of Cognitive Science. Oxford University Press, New York (2001)
3. Sloman, A.: Cosy-pr-0505: A (possibly) new theory of vision. online at: `http://www.cs.bham.ac.uk/research/projects/cosy/papers/` (2008)
4. Ziemke, T.: The Construction of 'Reality' in the Robot. Foundations of science **6** (2001) 163–233
5. Pupils of A. Meinong, ed.: A. Meinong's Gesammelte Abhandlungen (Collected Works). Volume 2: Abhandlungen zur Erkenntnistheorie und Gegenstandstheorie (Works on epistemology and object-theory). J.A. Barth, Leipzig (1913) Reprinted in: Alexius Meinong Gesamtausgabe (A. Meinong complete edition), Volume 2, Graz (1971).
6. Schlemmer, M.J., Vincze, M., Favre-Bulle, B.: Modelling the thing-in-itself – a philosophically motivated approach to cognitive robotics. In: Proceedings Epigenetic Robotics 2007. (2007)
7. Hawkins, J.: On Intelligence. Times Books, New York (2004)
8. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: Proceedings of the Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic (2004)
9. Bruner, J., Goodnow, J., Austin, G.: A study of thinking. Wiley, New York (1956)
10. Haarslev, V., Möller, R.: Racer: A core inference engine for the semantic web. (2003) 27–36
11. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. International Journal of Computer Vision **59** (2004) 167–181
12. Zillich, M.: Making Sense of Images: Parameter-Free Perceptual Grouping. PhD thesis, Vienna University of Technology (2007)
13. Hartmann, N.: Grundzüge einer Metaphysik der Erkenntnis (Outlines of a metaphysics of cognition). 5. edn. Walter de Gruyter & Co., Berlin (1965)
14. (W3C), W.W.W.C.: OWL Web Ontology Language Overview. online at *http://www.w3.org/TR/owl-features/* (2008)
15. Proffitt, D.R., Stefanucci, J., Banton, T., Epstein, W.: The role of effort in perceiving distance. Psychological Science **14** (2003) 106–112
16. Proffitt, D.R., Bhalla, M., R., G., Midgett, J.: Perceiving geographical slant. Psychonomic bulletin & review **2** (1995) 409–428