M. Matthies, H. Malchow & J. Kriz (eds.)
Integrative Systems Approaches to Natural and Social Dynamics.
Springer-Verlag Berlin, 2001.

Vision as computation, or: Does a computer vision system really assign meaning to images?

Andreas Schierwagen, Leipzig

Prof. Dr. Andreas Schierwagen

Universität Leipzig, Institut für Informatik

Augustuspl. 10/11

D - 04109 Leipzig, Germany

e mail:
schierwa@informatik.uni-leipzig.de

Abstract

Computer vision (or *image understanding*) is generally defined as the construction of explicit, meaningful descriptions of the structure and the properties of the 3-dimensional world from 2-dimensional images. A conceptual framework for image understanding that is widely accepted is based on Marr's concept of visual perception as computational process (Marr 1982). Marr postulated a hierarchical architecture for vision systems with different intermediate representations and processing levels (low, middle and higher level vision). The methodology introduced by Marr - description of cognitive processes on the levels of computational theory, algorithm and implementation - serves as a guideline, even today, for the "classical ", symbolic AI and the cognitivist paradigm of Cognitive Science, respectively.

In this paradigm cognition is defined as manipulation of symbolic representations according to the rules of a formal syntax. Inherent to this approach is the so-called symbol grounding problem. This problem consists in explaining how a (natural or artificial) cognitive system is, or can become, causally linked with its environment, so that both its behavior and the underlying basis mechanisms, representations etc. are meaningful for the system itself, and get meaning not only from an external designer or observer.

Thus the understanding problem of computer vision presents itself as variant of the symbol grounding problem. In this paper we examine the type of semantics employed in knowledge-based image understanding. It turns out that in both conventional and symbol grounding systems the semantics is "borrowed" - an interpretation by users remains necessary.

It is argued that the depicted problems with image understanding and symbol grounding are matters of principle. Since machines do not have subjectivity, it is unreasonably to expect that they could ever have an understanding capacity. Approaches based on the computing paradigm will be unable to capture the historically determined, holistic nature of living beings and their embedding in an ecological niche, even if modern AI theories emphasize the agent - environment interaction. We conclude that computer vision (and AI in general) should take the tool perspective and use its possibilities in a direct and constructive manner.

Key words:

image understanding, physical symbol system, representation, semantics concepts, symbol grounding

1. Introduction

Computer vision (or *image understanding*) represents a subfield of Artificial Intelligence which aims at the analysis and interpretation of visual information. Image understanding is considered as a process starting from an image or from image sequences (i.e. 2-dimensional projections of a static or dynamic scene) and resulting in a computer-internal description of the scene. The problems of image understanding are at the core of current efforts to enable a machine to make "intelligent" interactions with their environment. Sensors are used to obtain information from its 3D environment which can occur in the form of natural speech, images, noises etc.. This information is then processed in order to arrive at different forms of internal representation, again enabling the machine to interact with the environment, may it be in linguistic form or by actions of a robot.
The internal representations form the "knowledge" or the "models" of knowledge-based computer vision. According to the conventional methodology, the complexity of the processing steps is mastered by formulating and studying each cognitive problem on three mutually independent levels - the levels of computational theory, algorithm and implementation (Marr 1982).
Knowledge-based computer vision defines itself as part of "traditional" AI. The central concepts are symbol processing and representation; on the one hand Newell and Simon´s (1976) hypothesis of the physical symbol system serves as a theoretical framework, on the other hand Marr's concept of visual perception as a computational process.
This paradigm defines cognition as manipulation of symbolic representations according to the rules of a formal syntax. Inherent to this approach is the so-called symbol grounding problem. It consists in explaining how a (natural or artificial) cognitive system is or can become causally linked with its environment, so that both its behavior and the underlying basis mechanisms, representations etc. are meaningful for the system itself, and get meaning not only by an external designer or observer.
Thus the understanding problem of computer vision presents itself as a variant of the problem of whether semantic machines are possible or not. The present contribution characterizes the semantics concept of classical AI as one of internalistic semantics by means of which access to the world cannot be achieved. Further, the symbol grounding approach, which has been developed to enable a symbol system to get access to its environment, is reviewed. With this approach, the grounding of the internal representations in sensory (visual etc.) "experience" is aimed at. It turns out that a solution of the understanding problem in the actual sense, however, is not achieved. It is

argued that the speech of (image- or language-) understanding systems in AI should be only metaphorical.

2. Historical Outline of Computer Vision

The science of computer vision has passed through a number of paradigm shifts over the last four decades (viz. e.g. Neumann 1993, Crowley and Christensen 1995). It has its infancy in the 1950's when first attempts were undertaken to use the new computing machines to process images.
During the period 1965-1975 vision was mainly considered as pattern recognition. In this approach an object is described by a feature record. The similarity of objects is defined by the quantifiable degree of the agreement of the feature records which describe the objects. The book of Duda and Hart (1973) gives an informative overview of work from this time.
The pattern recognition approach soon encountered several fundamental difficulties. In particular, the problem of segmenting an image into significant chunks which could be classified proved to be generally unsolvable. It became obvious that segmenting requires more than only measurements in the image. Only by regarding the intended use can suitable segmenting be defined. Eventually it became generally accepted that machine vision requires an understanding of the world which was represented in the image. This led to a modification of the viewpoint to the position that vision was an application field for AI techniques.
Thus, the approach to investigate vision as image understanding was established. This reorientation took place in the 1970's, when new techniques were developed in AI for programming expert systems, in particular techniques of knowledge representation and inference. The expectation was that it would be possible with these techniques to provide the world knowledge needed for the analysis and understanding of images. From this period the anthology of Hanson and Riseman (1978) gives a representative overview of work.
The image understanding approach also soon encountered barriers which limited its success. Above all, the task to enter and formalize the necessary world knowledge proved to be feasible only for restricted domains. The segmenting problem cannot be solved with the image understanding approach. An important reason is that most AI techniques are rather sensitive to flaws of the image segmenting. Initial segmenting represents still today an important problem because of which many promising algorithms fail.

Another approach argued that understanding an image requires going back from the 2D pattern of grey or color values to the 3D form of the objects which generated the pattern. This recovery approach was developed by Marr (1982) and his colleagues at MIT into an influential concept, still strong today, for machine vision. Various techniques were specified with the goal to reconstruct the form of imaged objects on the basis of image features such as shading, texture, contour, movement etc. .

These so-called Shape-from-X techniques turned out to be ill-posed in the mathematical sense. A problem is well-posed when its solution exists, is unique and depends continuously on the given data. Ill-posed problems fail to satisfy one or more of these criteria. This means, for the case of a single static image, an unambiguous reconstuction is not possible in general. Uniqueness with the recovery can often be achieved if controlled camera movements are used, i.e. if images of the scene are taken from different views. Thus, the research area of active vision was introduced by Bajcsy (see Bajcsy 1988) and promoted by Aloimonos et al. (1988). Active vision techniques use algorithms of constant or linear complexity.

The contribution of active vision first was still embedded in the context of the recovery approach. Since the 1990's, modeling a vision system as an active agent has represented a lively research area. Thus, attention has been payed to criticism at the conception of AI machines as knowledge-based systems. Computer vision is no longer to be considered as a passive recovery process, but has to include the process of selective data acquisition in space and time. Further, a good theory of vision should provide the interface between perception and other cognitive abilities, such as reasoning, planning, learning and acting. In the framework of this approach, the aspects of attention, orientation to targets and purpose become important (Sommer 1995, Schierwagen and Werner 1998).

At the same time there are projects which resume the knowledge-based approach. The starting point is the assumption that object recognition includes the comparison of the objects with internal representations of objects and scenes in the image understanding system (IUS). From a computational perspective (on the level of algorithm and representation) different possibilities of implementation result. While Marr (1982) tried to put the data-driven recovery of the visual objects into practice, an "image-based" approach has been suggested (see Tarr and Buelthoff (1998) for review). This approach does not need recovery in the sense of computing 3D representations. Image-based models represent objects by their image from a specific viewpoint. In order to determine the perceptual similarity between an input image and known objects, robust matching

algorithms are required. Tarr and Buelthoff (1998) plead in summary for a concept of object recognition which incorporates aspects of both recovery and image-based models.

3. Knowledge-based machine vision today

Since not all researchers consistently followed the repeated conceptual changes in image understanding research, different viewpoints continue to exist next to each other. Neumann (1993, p. 567) suggested a definition comprising various approaches: "Image understanding is the recovery and interpretation of a scene on the basis of images allowing at least one of the following operations:

- output of a linguistic scene description
- answering linguistic inquiries concerning the scene
- collision-free navigation of a robot in the scene
- regular grasping and manipulating of objects in the scene. "

This definition includes the interpretation of images and thereby emphasizes understanding. The suggested operational term of understanding is to ensure that it is not the programmer of the IUS that accomplishes the understanding, but actually the system. Inputs to the system are camera images from which during a multi-level process a representation of the environmental scene which caused the images is obtained. A scene is thereby a spatial-temporal window of the environment. Static scenes are, in general, 3-dimensional, and dynamic scenes are 4-dimensional. An image is a 2D projection of a static scene; dynamic scenes lead to image sequences.
The computer-internal description of the scene serving as output consists of two parts: (i) information about the spatial-temporal relations of the scene objects and (ii) interpretation of scene content, particularly object recognition. The internal representation of the scene description is realised by knowledge representation methods and inference techniques (in particular spatial reasoning).
The conceptual framework within which vision is examined in cognitivist AI is represented in Fig. 1. Image understanding is described as a process of four cooperating, task specific subprocesses which in each case require specific intermediate representations. The primary image analysis proceeds from the digital raster image in which the radiometric characteristics (intensity

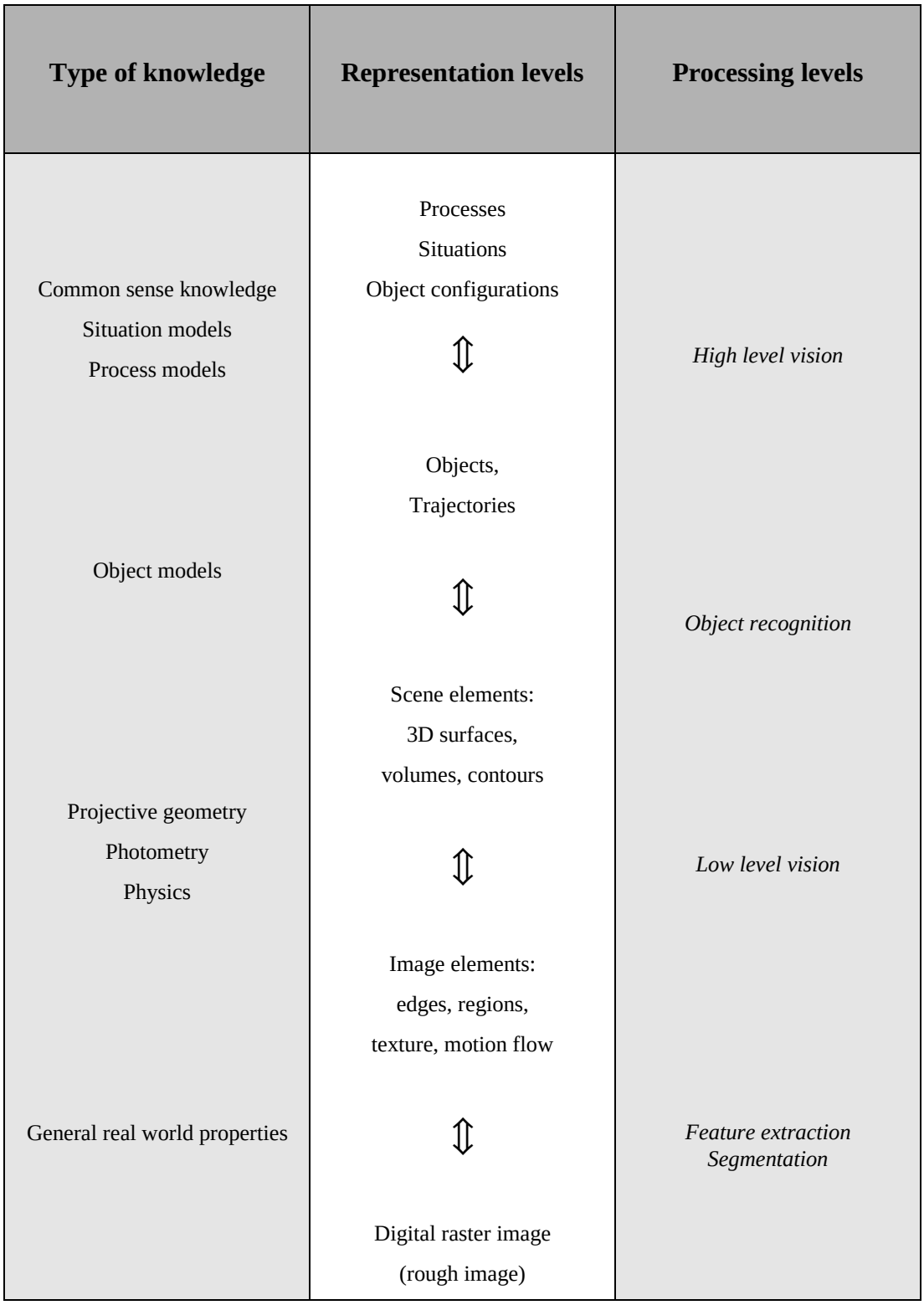

Fig. 1: Image understanding as a hierarchical, knowledge-based process. Represented are the different types of knowledge which are used to infer the scene description from the image (left), and the intermediate representations on the different levels (centre) produced by the corresponding processing steps (right). Adapted from (Neumann 1993).

and colour) of each pixel are recorded, to the determination of image elements (edges, homogeneous areas, texture, etc.).
Low level image interpretation aims at interpreting image elements as scene elements i.e. as results of the mapping of parts of a 3D scene. Processes of this level are to solve a central task of image understanding: the extraction of real world characteristics from image properties. It includes in particular the recovery of 3D object shapes by means of the shape-from-x techniques.
In the following processing step - object recognition - objects are identified in the image data extracted so far, and on the basis of the scene elements. A crucial role here is played by the a-priori knowledge of which displays are produced by the camera if objects are seen from different views. This a-priori knowledge is represented by the object models of the knowledge base.
The higher level image interpretation summarizes further processing steps which aim at detecting "object and time-transcending connections, e.g. interesting object configurations, special situations, coherent motion sequences, etc.. Analogous to object recognition, a-priori knowledge of what one wants to detect plays an important role here" (Neumann 1993, p. 570). The content of the resulting description depends not only on the scene or the corresponding image, but also on the question or the context in which the output is to be used.
Although current knowledge-based IUS's do not show a strict partitioning into hierarchically organized subprocesses, they are still oriented at the sketched conceptual framework. They have an interactive-hierarchical architecture, in which partial results of earlier processing steps trigger processes on higher levels whose results feed back to the processing steps of lower levels. Examples are knowledge-based systems for the integration of machine vision and natural language processing (see e.g. Hildebrandt et al. 1995, Herzog et al. 1996, Pauli et al. 1995 and the references therein).

4. "Understanding" with AI machines

We turn now to the question which concept of understanding has been used in knowledge-based computer vision, and whereby or at which step during the processing of images this "understanding" takes place.
The definition of image understanding (paragraph 3) includes on the one hand that a name is assigned to an object. This can be achieved by various matching algorithms, i.e. results of the low and middle processing level are compared on the high representation level (image or scene description) for matching stored object models. In the case of a robot the problem presents itself

differently. It is not "explicit" understanding by designation which is important, but the "implicit" proof for understanding by showing adapted behavior in its interaction with the environment. Thus, in both cases a Turing test serves to judge (by us as observers) whether the IUS understands the scene: the (linguistic or sensomotor) behavior is used as criterion for understanding. In AI and in the philosophy of mind, the validity of Turing tests with respect to the understanding capacity of symbol systems is controversial. Criticisms are directed against the physical symbol system hypothesis (PSSH) of Newell and Simon (1976) which states that a physically implemented sign manipulator - a physical symbol system - possesses sufficient and necessary means for "intelligence". For these authors the symbol concept is completely defined within the structure of the symbol system, even if a connection to the designated object is required. The form of the symbols is arbitrary, and their interpretation takes place according to social agreement between observers/users of the symbol system. According to this hypothesis intelligent behavior consists of the following steps: generation of symbols by the sensory apparatus, then manipulation of these symbols (for instance with inference techniques or algorithmic search) in order to create a symbol or a symbol structure as output.

As an example we may consider a suitable programmed system, passing the Turing test for image understanding. According to the claims of "strong AI", such a system understands the scene and represents at the same time the explanation for how humans understand this scene. Searle (1980) formulated one of the most well-known arguments against the PSSH. In the context of image understanding (Fig. 1) it reads as follows: The early, near-signal processing steps are followed by symbol processing steps (object recognition, higher image interpretation) to which Searle's argumentation concerning the "Chinese room" can be applied (Searle 1980). Although an observer of the IUS will have the impression that it understands the scene, this is not true: the underlying algorithms, formulated by a programmer in a certain programming language, do not own meaning from and for themselves. Searle has developed the arguments (Searle 1990) which are supported by analyses of the semantics concept of computer science (viz. for example, Hesse 1992, p. 285; Lenz and Meretz 1995, p. 70).

Following Searle, Harnad (1990) criticized the claim of the symbol processing approach that meaningful programs could arise from rule-following symbol manipulations. He proposed a conceptual model to ground symbols in the environment of a system. The symbol grounding problem consists in answering the question: "How can the semantic interpretation of a formal system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?" (Harnad 1990, p. 335). As a candidate solution, Harnad suggested to connect symbols causally by

non-symbolic (iconic and categorial) intermediate representations with the objects to which they refer. Neural networks were considered as the appropriate tool to produce the intermediate representations. A hybrid connectionist / symbolic system was envisaged to typify the conjunction of sensory experience and symbol.

As Harnad emphasized, with symbols grounded in this way the compositionality of the system would easily be achieved. An IUS could thus understand a complex scene, as elementary objects were grounded in sensory experience, and the inherent meaning of complex objects, object constellations etc. would result, in accordance with Frege's principle of compositionality.

Later Harnad conceded that by symbol grounding, probably the only thing that can be achieved, is to limit the interpretation possibilities for the symbols. It cannot be guaranteed that the semantics of the symbols is intrinsic, i.e. independent of interpretation (Harnad 1993, 1994). With the symbol grounding approach a correlation semantics can be implemented, which can be advantageous for a technically oriented AI. In Harnad's words: "... the fact that our own symbols do have intrinsic meaning whereas the computer's do not, and the fact that we can do things that the computer so far cannot, may be indications that even in AI there are performance gains to be made (especially in robotics and machine vision) from endeavouring to ground symbol systems" (Harnad 1990, p. 340).

5. Conclusion

In this contribution the conceptual framework was presented, within which image understanding is described as a hierarchical, knowledge-based process. We considered the understanding problem of computer vision, i.e. the question of whether the performance of an IUS is limited to the manipulation of signs (signals or symbols), or whether it is possible that the signs can get intrinsic meaning.

To answer this question, we considered the type of semantics which is employed in image understanding. Conventional IUS's possess an internalistic semantics, i.e. the meaning of a symbol is seen in the conceptual role which it plays with respect to the other symbols. In this way the IUS cannot acquire access to the world; its semantics is "borrowed", an interpretation by users is necessary. The semantics of (hybrid) symbol grounding systems is of correlative nature, i.e. a symbol gets meaning through the correlation of sign and designatum. This means that, also in these systems, interpretation remains necessary. The symbols, however, are not completely arbitrary in their interpretation. In both cases it is not possible that symbols can possess an intrinsic meaning.

There is evidence that the sketched difficulties with image understanding are matters of principle. The PSSH excludes semantic aspects while describing intelligent behavior, as independence is postulated between the syntactic and the semantic level, i.e. the rule-based manipulations of signs and the respective semantic interpretation. Haugeland (1981, p. 23) formulated it so: "... if you take care of the syntax, the semantics will take care of itself."

As we saw, IUS's can at best evoke the illusion that they would understand; in fact it is we (as users, programmers etc.) who lend meaning to these systems. Thus, it's our existence by which the physical symbol structures of an IUS etc. can be instantiated semantically. In other words, talking about image understanding (or about language comprehension) with machines is a category error: Machines do not have subjectivity, and therefore it is unreasonable to expect that they could ever have an understanding capacity in the true sense.

This assessment is not restricted to AI systems based on symbol processing, but also has validity for alternative (connectionist, enactive) approaches (cf. for example, Lenz and Meretz 1995, D'Avis 1998, Ziemke 1999). The reason is that computationalist approaches in general are unable to capture the historically rooted, holistic nature of living beings and their embedding in the ecological niche. This is also true for the very recent attempts of situated and embodied AI, despite its emphasis of agent - environment interaction. After all, cognitions are not computations, and for computer vision (and AI in general) this can only mean that we should take the tool perspective and use the possibilities of these tools in a direct and constructive manner.

References

Aloimonos Y, Weiss I, Bandopadhay A (1988) Active vision. Int J Comp Vision 7: 333-356

Bajcsy R (1988) Active perception. Proc IEEE 76: 996-1005

Crowley JL and Christensen HI. (1995) Vision as Process. Springer, Berlin

D'Avis W (1998) Theoretische Lücken der Cognitive Science. J Gener Philos Sci 29: 37-57

Duda R, Hart P (1973) Pattern Classification and Scene Analysis. Wiley, New York

Hanson A, Riseman E (1978) Computer Vision Systems. Academic Press, New York

Harnad S (1990) The symbol grounding problem. Physica D 42: 335-346

Harnad S (1993) Symbol grounding is an empirical problem. In: Proc 15th Annual Conference of the Cognitive Science Society. Boulder, CO, pp 169-174

Harnad S (1994) Computation is just interpretable symbol manipulation; cognition isn't. Mind and Machines 4: 379-390

Haugeland J (1981) Semantic engines: An introduction to mind design. In: Haugeland J (ed) Mind Design. MIT Press, Cambridge, MA/ London, England, pp 1-34

Herzog G, Blocher A, Gapp K-P, Stopp E and Wahlster W (1996) VITRA: Verbalisierung visueller Information. Informatik Forschung und Entwicklung 11: 12-19

Hesse W (1992) Können Maschinen denken - eine kritische Auseinandersetzung mit der harten These der KI. In: Kreowski H-J (Hg.) Informatik zwischen Wissenschaft und Gesellschaft. Springer, Berlin [etc.], pp 280 - 289

Hildebrandt B, Moratz R, Rickheit S, Sagerer G (1995) Integration von Bild- und Sprachverstehen in einer kognitiven Architektur. Kognitionswissenschaft 4: 118-128

Lenz A, Meretz S (1995) Neuronale Netze und Subjektivität. Vieweg, Braunschweig/Wiesbaden

Marr D (1982) Vision. WH Freeman, San Francisco

Neumann B (1993) Bildverstehen - ein Überblick. In: Görz G (ed) Einführung in die künstliche Intelligenz. Addison-Wesley, Bonn [etc.], pp 559-588

Newell A, Simon HA (1976) Computer science as empirical enquiry: Symbols and search. Commun ACM 19: 113-126

Pauli J, Blömer A, Liedtke C-E, Radig B (1995) Zielorientierte Integration und Adaptation von Bildanalyseprozessen. KI 3: 30-34

Schierwagen A, Werner H (1998) Fast orienting movements to visual targets: Neural field model of dynamic gaze control. In: 6th European Symposium on Artificial Neural Networks - ESANN ‘98, D-facto publications, Brussels, pp 91-98

Searle JR 1980) Minds, brains and programs. Behav Brain Sci 3: 417-457

Searle JR (1990) Is the brain a digital computer? Proc Adr Amer Philos Assoc 3: 21-37

Sommer G (1995) Verhaltensbasierter Entwurf technischer visueller Systeme. KI 3: 42-45

Tarr MJ and Bülthoff HH (1998) Image-based object recognition in man, monkey and machine. Cognition 67: 1-20

Ziemke T (1999) Rethinking grounding. In: Riegler A, vom Stein A, Peschl M (eds) Does Representation Need Reality? Plenum Press, New York, pp 87-100