Implicit Modeling of Object Topology by Guidance with Temporal View Attention

Peter Michael Goebel and Markus Vincze

Automation & Control Institute
Faculty of Electrical Engineering and Information Technology
Vienna University of Technology, A-1040 Vienna
{goe,vincze}@acin.tuwien.ac.at
http://www.acin.tuwien.ac.at

Abstract. Object recognition developed to the most common approach for detecting arbitrary objects based on their appearance. Statistical pattern analysis methods, are able to extract features from appearing images, enabling a classification of the image content; have reached certain maturity; and achieve excellent recognition on rather complex problems. However, these systems seem not directly scalable to human performance in cognitive sense and appearance does not attribute to understanding the structure of objects. Due to noise, occlusions, and illumination, objects are segmented often poorly with more or less drop outs in contour that yields poor recognition performance, and since object representation enables logical input of spatial arrangements to higher cognitive processes, scene interpretation in cognitive manner gets inhibited.
In another paper we proposed the architecture and a simulation of the five bottom layers of a cognitive vision model by implementing the striate visual cortex as the first level. Hence, in this work we focus on the concept of modeling object prototypes from geon recipes on biological formations, such as the circuit of Papez, and show how structure of such formation can be utilized for the modeling of objects. The proposed implementation is exemplified by an object similar to the Necker cube.

Key words: Cognitive Modeling, Cognitive Representation, Fuzzy Graphs, SubGraph Matching, Image Primitives

1 Introduction

The field of *Cognitive Systems* (CS) is concerned with high-level advanced *cognitive* capabilities that are enablers for the achievement of more intelligent goals such as scene understanding, and autonomous navigation in complex cluttered environments. Vision, as a key perceptional capability relates to rather difficult problems, such as visual *object recognition*, *representation* and *scene understanding* [Pin05]. A projection of observed objects from a 3D scene onto a 2D sensor array is commonly used. Here similarities to biology are found in the projection of a scene onto the retina of a mammal's eye. Human perception, accordingly to Gestalt theory [Kof35], tends to inherently assume the simplest and most regular

Proceedings of the ICVS International Cognitive Vision Workshop – ICVW 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

organization that is consistent with a given image. Geometric relationships, such as collinearity and parallelism, are constant properties of projections of collinear or parallel edges of the visible layout. This tendency underlies the organization of visible surfaces into objects [WS02].

Therefore, syntactical approaches utilizing this support were developed, e.g. so about twenty five years ago, Marr and Nishihara [Mar82] presented a model of recognition, restricted to the set of objects that can be described as generalized cones; Biederman [HB06] introduced 1987 *geon* theory, arguing that complex objects are made up of arrangements of basic component parts (i.e. geons that represent cubes, cylinders, spheres, etc); and Riesenhuber and Poggio [RP02] proposed a hierarchy of nested arrangements of local features such as lines and vertexes.Christou et al. [CTB99] studied whether contextual information regarding an observer's location within a familiar scene could influence the identification of objects. Results suggest that object recognition can be supported by knowledge of where we are in space and in which direction we are looking. Johnson-Laird's [JL83] mental model theory proposes reasoning as a semantic process of construction and manipulation of models in working memory of limited capacity. It provides a unified account of deductive, probabilistic, and modal reasoning.

However, in our approach, we use the ability of the system to change its viewpoint as an important issue for object representation learning. In particular, we are deriving recipes from "geon" like object definitions in order to support the concatenation of line and point primitives for building object prototypes, which are stored in brain memory.

The remaining sections of the paper are organized as follow: Section 2 recalls prerequisites needed for easy understanding of our approach. Section 3 explains the model with applying it to a given problem simulation. Section 4 concludes with an outlook on further work.

2 Prerequisites

In this section we give a short survey on graphs, fuzzy graphs, subgraph matching for object representation, cognitive working memory and attention models in order to provide background information as needed and relate it to our approach.

2.1 About Graphs

An undirected graph (see Fig. 1) $G = (V, E)$ consists of a set V of vertices and a set E of edges whose elements are unordered pairs of vertices. The edge $e = (u, v) \in E$ is said to be incident with vertices u and v, where u and v are the end points of e. Then these two vertices are called adjacent. The set of vertices adjacent to v is written as $A(v)$, and the degree of v is the number of vertices adjacent to v and is denoted as $|A(v)|$. The graph in Fig.1 is defined with $V = \{a, b, c\}, n = |V| = 3$ vertices; $E = \{e_1, e_2, e_3\}, m = |E| = 3$ edges; and $A(v) = \{a, b\}, |A(v)| = 2$. When the edges E are assigned with weights

Proceedings of the ICVS International Cognitive Vision Workshop – ICVW 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

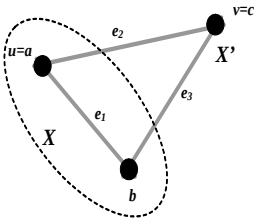

Fig. 1. An example graph, with $m = 3, n = 3$; e_2 incident to u and v; and the partition (X, X')

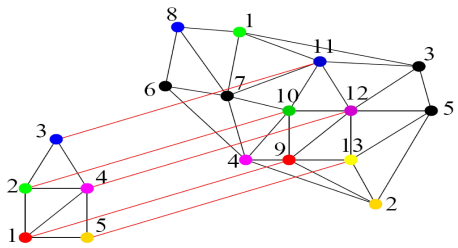

Fig. 2. A subgraph matching; the object graph is matched to the scene graph [SS05].

$w(e_i)$, the graph gets a weighted graph $G = (V, E, w)$. A partition (X, X') is defined as the proper disjoint subsets of V. The complement of $X \subseteq V$ is denoted $X' = V - X$. The open neighborhood of X is defined as $\Gamma(X) = \{v \in X' | (u, v) \in E \; for \; some \; u \in X\}$; an induced subgraph $\langle X \rangle$ is the graph $H = X, F$ where $F = \{(u, v) \in E | u, v \in X\}$. An alternating sequence of distinct adjacent vertices and their incident edges is called a path; when a $u \ldots v$ path exists, the graph G is connected; otherwise G splits in a number of subgraphs; $G/\{u\}$ means the vertex u deleted from G.

Fuzzy Graphs Conjectured from [BBP02], taxonomy of fuzzy graphs can be classified in five primary types: (i) fuzzy set of crisp graphs; (ii) crisp vertex set and fuzzy edge set; (iii) crisp vertices and edges with fuzzy connectivity; (iv) fuzzy vertex set and crisp edge set; and (v) crisp graph with fuzzy weights.

Thus to change the Graph in Fig.1 to a fuzzy graph, we add membership functions as weights $\alpha, \beta, \ldots; \mu$, and define $V = \{a, \alpha, b, \beta, c, \gamma\}, n = |V| = 3$ for fuzzy vertices; and $E = \{e_1, \mu_1, e_2, \mu_2, e_3, \mu_3\}, m = |E| = 3$ for fuzzy edges. The membership functions are chosen in accordance to the specific task at hand.

Relational Graph Matching Relational matching algorithms [CFSV04] offer the advantage of drawing on structural constraints in the matching process without the need for calibration. Processing of natural images with intrinsic variabilities of patterns, noise, and occlusions often yields incomplete graph representations to which the matching has to be tolerant. Matching on graphs generally leads to NP-hard problems, however, the compactness of information finally provides advantages over other representations.

In this paper we follow the combinatorial subgraph matching by the semidefinite program (SDP) convex relaxation approach of Schellewald and Schnörr [SS05]. In their approach, see Fig. 2, *model graphs* G_K (shown at the left) representing object views are matched to *scene graphs* G_L (shown at the right) by *bipartite matching*. They assume $K \leq L$ with $K = |V_K|, L = |V_L|$ vertices and a distance function $w_{i,j}$, which measures the similarity of the vertex pairs $i \in V_K$ and $j \in V_L$. Thus an optimal matching in the bipartite graph

Proceedings of the ICVS International Cognitive Vision Workshop – ICVW 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

$G_B = (V_K \cup V_L, E)_{(i,j)\in E}$ for all pairs (i, j) with corresponding weights $w(i, j)$ can be found. We extend in our approach their work to graphs representing scenes in 3D space for pattern completion as described in section 2.4.

2.2 Attention and Search in Vision

In general, searching is the task retrieving subsets from available data collections. It is also finding the smallest distance between a given template of something to be searched, and a formation where it is to search for, using an appropriate metric. The aim of visual search is to find the odd item – discriminated by saliency, color, size, orientation, depth or movement.

Attention is the cognitive process of selectively concentrating on one item, while ignoring others. It is tied very closely to perception. Feature search appears easy, since no attention is needed; conjunction search appears hard since attention is needed. With using saliency maps [IKN98], simulating the elements of a visual scene that are likely to attract the attention of human observers gets possible. Bak et al. [BTW87] suggest that depth information may guide our perceptual system into a self-organized state to assist us in resolving ambiguous information in object perception.

With this findings in mind, we decided in our approach to use 3D data rather than 2D appearance models to gain advantage for object recognition.

2.3 Revisiting Brain Memory

Since brain memory is not directly observable, a consensus between cognitive psychology, neuropsychology, and neuroanatomy declares that functional memory regions are defined by patterns of neuronal activity, caused by certain events within our environment. The brain is always confronted with too much information, thus it attempts to use attention to extract only salient features *"from the given story"*.

Recent developments in fMRI[1], PET[2] neuroimaging enable to observe such activities in vivo [Dav06]. Brain memory formation is not permanent, since it develops with retrieval, association, and forgetting. In most cases information is not forgotten; rather, new information is acquired that interferes with the old. Thus, new information is linked to old information by association and that allow grouping events and items in categories. Associations are formed between stimuli, which are presented in close approximation of space and time to existing memory areas, where highly emotional events are remembered best [HFO06].

Working Memory The concept of working memory was popularized in 1974 by Baddeley and Hitch [Bad00], constituted by a system of two slave memories, the *short-term* (STM) and the *long-term* (LTM) memory. Since a general discussion of related models is beyond the scope of this contribution, we refer to Miyake and Shah [MS99] for a comparison of ten different models of working memory.

[1] functional magnetic resonance imaging
[2] positron emission tomography

Proceedings of the ICVS International Cognitive Vision Workshop – ICVW 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Mammalian learning focuses on solving specific tasks, i.e. by task learning, where a rapid learning of new things requires another formation than that is utilized with classical backpropagation in cortical memory. Evidence from psychology research, for example, is given by considering the AB-AC paired associates list learning task. Here A represents one set of words that is associated with two different sets of other words, B and C. After studying the $A \cup B$ list, the subjects are tested by asking them to give the appropriate B associate for each of the A words. Then, subjects study the $A \cup C$ list often and are subsequently tested on both lists for recall of the associates after each iteration of learning the $A \cup C$ list. Although subjects do exhibit some level of interference on the initially learned $A \cup B$ list as a result of learning the $A \cup C$ list, they still remember a reasonable percentage of the $A \cup B$ list. Attempts to train a backpropagation neuronal network to perform the AB-AC task yields suffering the net from catastrophic interference [MN89].

This findings are quite similar to the well known XOR problem with the neural perceptron model and give evidence that the brain appears to have developed two specialized systems – the posterior cortex, which uses slow interleaved learning – and the hippocampus that uses sparse pattern separated representations, which enables rapid sequential learning.

2.4 O'Reilly's Hippocampal Episodic Memory Model

In our work, we utilize the model of O'Reilly et al. [HFO06], which neither formulates a structural distinction between declarative and procedural knowledge nor postulates a central executive; the processing appears rather distributed within the entire cortex.

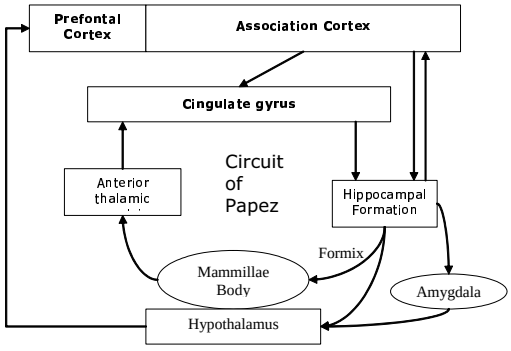

Fig. 3. The Papez circuit refers to a neural circuitry linking limbic centers. Cells in the mammillary body of the hyperthalamus project to the anterior thalamus. Cells here project to the cingulate gyrus, from which cells in turn project to the hippocampus. Fibers arising from the hippocampus course through fornix to the hypothalamus, closing the circuit [HFO06].

Proceedings of the ICVS International Cognitive Vision Workshop – ICVW 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

The limbic system is a group of subcortical brain areas and consists principally of the *hippocampus* (rapid learning), *cingulate gyrus* (motor control and action selection), *hypothalamus* (vital functions), *anterior thalamus*, *amygdala* (emotions), and the *mammillary bodies*. The three recent areas are part of the *Circuit of Papez*, Fig. 3, where information from STM circulates until it becomes "permanent" in LTM. The *thalamus* holds many specialized subdivisions (nuclei) that provide sensory input to the cortex[3]. The *hippocampus* forms the central axis of the limbic system. It is critical to spatial learning and awareness, navigation, episodic/event memory, and associational recollection.

The hippocampus is mostly feedforward directed and sits at the top of the cortical hierarchy; receives a wide range of information from various cortical areas; and encodes the current state of the environment in such a way that some fraction of the original patterns can be used to retrieve the original. Hence, in the hippocampus two mechanisms are competing – *pattern separation*, operating during encoding of new patterns – and – *pattern completion*, enabling partial cues to trigger activation of previous stored information.

Pattern separation can be achieved by considering the concept of a unit's activation threshold to get excitation as it requires to overcome inhibitory competition from other units. The central idea is that sensitivity to the conjunction of activity in the input produced by high thresholds yields pattern separation. To work this optimally, it is important that different receiving units are maximally activated by different input patterns, which can be achieved by having a high level of variance in the weights of partial connectivity with the inputs.

Pattern completion is necessary to retrieve already persistent patterns instead of storing every new activation in an separate area.

3 Explaining the Object Modeling

Utilizing findings of psychological vision perception to cognitive modeling, our recent work presented a new practical concept for the modeling of visual object representations, shown in Fig. 4 that claims to close the gap between appearance based image processing and cognitive models. We conjectured Peschl's "Modes of knowing and modes of coming to know" [Pes06] providing the cognitive functionality, where the vision model layers I-1 to I-5 form Peschl's behavioral level-1.

3.1 The Cognitive Constructivists Framework

The cognitive constructivists framework [GV07] is organized in five cognitive levels (I) ... (V) of abstraction:

[3] e.g. lateral geniculate nucleus, LGN, in the early visual pathway

Proceedings of the ICVS International Cognitive Vision Workshop – ICVW 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

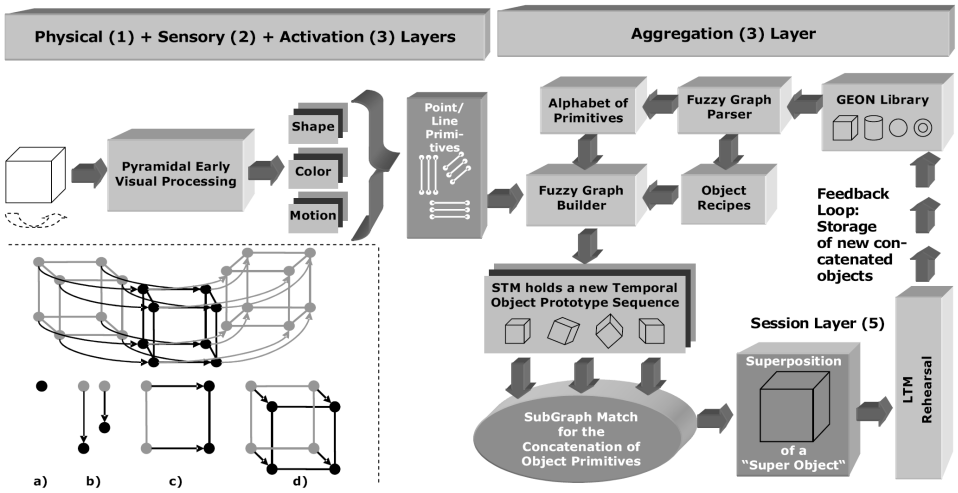

Fig. 4. The conceptual model with focus on the aggregation layer as the contribution of this work. The cube gets observed from a 3D stereo camera; the pyramidal early vision processing layers (1..3) generate line and point primitives; with the aggregation layer (4), the primitives are concatenated by object recipes, derived from geon objects in a library; the object prototypes are matched together and the superposition representation is generated for LTM update within session layer (5). From the LTM representation, the Geon library can be updated, which yields learning of objects by their structural description.

(I) level of executive behavior: Realized as a list of observations, it describes a phenomenon on its behavioral level. On this level, the proposed cognitive vision model concept, contributed by the authors, is embedded with the

- **(i) physical layer -** implements the interface to the hardware; the output of this layer is a multichannel raw data time series.
- **(ii) sensory layer -** provides the temporal preprocessing stage similar to area V1 in the striate cortex (SC); preprocessing is realized by the application of the (pyramidal) contourlet transform [DV03], thus an image is decomposed into contour (i.e. edge), directional, and motion information.
- **(iii) activation layer -** generate a list of primitives similar to area V2 of the SC, representing shape, color, movement, and time.
- **(iv) aggregation layer -** primitives of layer (iii) are concatenated by fuzzy graphs, the outcomes are object prototypes, gathered from different viewpoints; the prototypes are then matched and stored in STM.
- **(v) session layer -** the object prototypes are used for learning and exercising (rehearsal) of the long term memory; the outcome is a knowledge base with knowledge about learned objects in arbitrary pose.

Proceedings of the ICVS International Cognitive Vision Workshop – ICVW 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

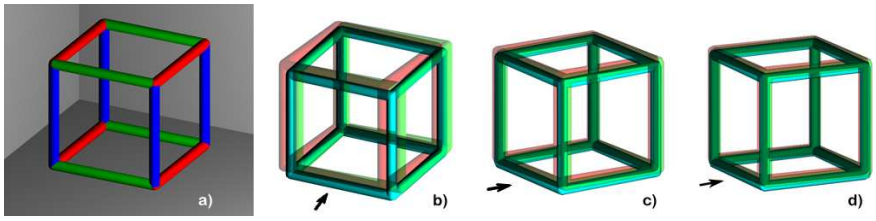

Fig. 5. The trinocular stereo simulation of the method similar to the Necker cube as example: (a) shows the 3D cube and its surrounding environment; (b) is the stereoscopic view of (a) with robot's x-coordinate position 0.5, (c) 1.5, and (d) 2 meters.

(II) level of hidden patterns of behavior: The patterns are the result of more or less complex inductive and constructive processes. The class of scientific explanations are situated on this level.

(III) level of causes: This level concerns the exploration and the construction of causes; the resulting knowledge is the source for a deeper understanding of a phenomenon.

(IV) level of potentiality: This level changes the perspective from the mode of (constructive) perception to the mode of externalization; new physical realities are created or existing (physical) realities are changed.

(V) level of reflection This step has the potential of fundamentally questioning the knowledge that has been constructed so far by reflecting on the knowledge, its premises, and on the construction and learning processes that have led to that knowledge; completely unexpected results and new perspectives can be brought up that have never been considered before.

3.2 The Early Processing

Fig. 5 shows a cube 3D simulation setup of the proposed approach. In the model, we are considering a mobile robot is moving around the cube (a). Its trinocular stereoscopic views (b)..(d) are processed by the early processing stages of the cognitive vision model [GV07]. Herein, a 4D-temporal feature map of a trinocular stereo image setup (Fig.5b) is triggered by an activation function, generating line primitives. Fig. 6 shows eight directional sets of such line primitives.

The principle of language syntax is used by the "geon" library, shown in Fig. 4, thus a fuzzy graph parser derives recipes according on the defined object alphabet (i.e. of cube, cylinder, cone, sphere, etc.).

The recipes use a fuzzy graph construction, shown Fig. 4 lower left corner box, utilizing A* heuristic search strategy that apply together *lowest-cost-first* and *best-first* searches, optimizing path cost as well as heuristic information in its selection of current best path [PMG98]. Thus, in (a) the graph starts with A* heuristics seeking the first correspondences within a 3D fuzzy plane, and with (b) 1D recursive seek for corresponding vertexes, defining edges within the

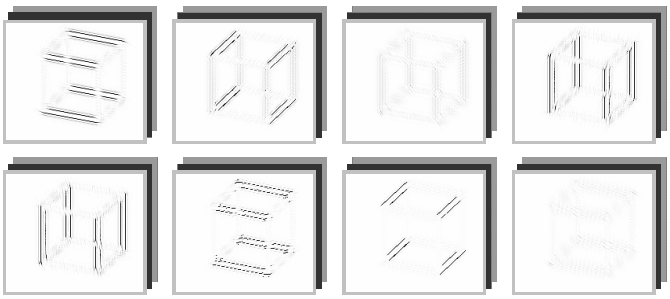

Fig. 6. The responses of the highest level of the pyramidal contourlet transform [DV03], with eight directions are shown $[0 + 7 \times (\Theta = 45^\circ)]$. The line segments appear broken due to the self occlusions by the front parts of the object.

plane; (c) 2D recursive seek for corresponding edges follows; and (d) 3D recursive seek for corresponding planes concludes the search. A* uses an estimate from the Hough [DH72] space of extracted line segments for the next path in search and the object with best best fitness is selected for storage in STM. Finally, the circuit of Papez is implemented herein for LTM rehearsal of objects.

Hence, the object candidates, concatenated by the fuzzy builder in Fig. 3, using the geon-recipe alphabet and as stored as a sequence in STM, are matched to the scene graph auto-defined by the geon-recipe alphabet. This matching yields the superposition of the outcomes from all object candidate views into the representation of the derived object.Thus possible perturbations within single object views, such as missing vertices or edges are superseded with the information provided by other object candidate views.

The rehearsal strategy of presenting object prototypes from different viewpoints as time series for learning is used to select object prototypes for storage in LTM, updating the geon library. Now, new objects can be represented and used in a playground by higher cognitive processes.

4 Conclusion

In this work, building of object prototypes, supported by the proposed model was given with focus on learning with a hippocampal formation. Primitives, stemming from image sequences are detected and aggregated into internal object representations by applying fuzzy graph constraints and guidance from automatically generated "geon" recipes. Due to lack of space, a detailed explanation of other parts of the model will be presented elsewhere. As example, the Necker cube was given and simulated in a 3D simulation setup. The modeling approach will be utilized by future work for the representation and recognition of other object types with occlusions in cluttered, and noisy environments. The model can be extended to include auditory perception, which may provide advantage for cognitive learning.

Proceedings of the ICVS International Cognitive Vision Workshop – ICVW 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

References

[Bad00] A. D. Baddeley. The episodic buffer: a new component of working memory?. *Trends in Cognitive Science.*, 4:417–423, 2000.

[BBP02] M. Blue, B. Bush, and J. Puckett. Unified approach to fuzzy graph problems. *Fuzzy Sets and Systems.*, 125(3):355–368, 2002.

[BTW87] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality:an explanation of 1/f noise. *Physical Review Letters.*, 59:381–384, 1987.

[CFSV04] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. A (sub)graph isomorphism algorithm for matching large graphs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10):1367–1372, 2004.

[CTB99] C. G. Christou, B. S. Tjan, and H. H. Bülthoff. Viewpoint information provided by a familiar environment facilitates object identification. *TR Max-Planck institute f. biol. cybernetics.*, 68, 1999.

[Dav06] L. Davachi. Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology.*, 16:693–700, 2006.

[DH72] R. Duda and P. Hart. Use of the hough transformation to detect lines and curves in pictures. *Comm. ACM*, 15, 1972.

[DV03] M. N. Do and M. Vetterli. The contourlet transform. *IEEE Transactions on image processing.*, 14(12):2091–2106, 2003.

[GV07] P. M. Goebel and M. Vincze. Vision for cognitive systems: A new compound concept connecting natural scenes with cognitive models. In *LNCS, Springer, to appear.*, 2007.

[HB06] K. J. Hayworth and I. Biederman. Neural evidence for intermediate representations in object recognition. In *Vision Research.*, in Press 2006.

[HFO06] T. E. Hazy, M. J. Frank, and R. C. O'Reilly. Banishing the homunculus: Making working memory work. *Neuroscience*, 139, 2006.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 20(11):1254–1259, 1998.

[JL83] P. N. Johnson-Laird. *Mental Models, Towards a Cognitive Science of Language, Inference and Consciousness.* Harvard Press, 1983.

[Kof35] K. Koffka. *Principles of Gestalt Psychology.* Hartcourt, NY, 1935.

[Mar82] D. Marr. *Vision: A Computational Approach.* Freeman & Co., 1982.

[MN89] M. McCloskey and N.Cohen. *Catastrophic interference in connectionist networks: The sequential learning problem.* In G. H. Bower (ed.) The Psychology of Learning and Motivation, Academic Press., 1989.

[MS99] A. Miyake and P. Shah. *Models of Working Memory.* Cambridge, 1999.

[Pes06] M.F. Peschl. Modes of knowing and modes of coming to know. *Constructivist Foundations*, 1(3):111—123, 2006.

[Pin05] A. Pinz. Object categorization. *Foundations and Trends in Computer Graphics and Vision.*, 1(4):257–353, 2005.

[PMG98] D. Poole, A. Mackworth, and R. Goebel. *Comp. Intell.* Oxford Press., 1998.

[RP02] M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Current Opinion in Neurobiology.*, 12:162–168, 2002.

[SS05] C. Schellewald and C. Schnörr. Probabilistic subgraph matching approach based on convex relaxation. In *Proceedings of EMMCVPR, LNCS, Springer.*, volume 3757, pages 171–186, 2005.

[WS02] R. F. Wang and E. S. Spelke. Human spatial representation: insights from animals. *Trends in Cognitive Sciences.*, 6(9), 2002.