Pursuing Artificial General Intelligence By Leveraging the Knowledge Capabilities Of ACT-R

Alessandro Oltramari, Christian Lebiere

Department of Psychology, Carnegie Mellon University, Pittsburgh (USA)

Abstract. Intelligence is a multifaceted phenomenon which makes trying to capture its very essence a slippery task. In this paper, we commit to a hybrid notion of intelligence, conceived as the combination of cognitive operations and knowledge resources that leads to purposeful behavior. Accordingly, this paper describes an artificial system that benefits from both **mechanism–centered** and **knowledge–centered** approaches. In particular, the system integrates the ACT-R cognitive architecture with SCONE, a knowledge-based system for ontological reasoning, to combine ACT-R's subsymbolic cognitive mechanisms with SCONE's knowledge representation and inference capabilities. We apply the hybrid system to computationally approximate human intelligent behavior in a task of visual recognition.

1 Introduction

'An architecture without content is like a computer without software - it is an empty shell"[1]

"What is intelligence?". From the dawn of Western Thought to the Contemporary (scientific) Age, scholars from different disciplines have struggled to answer this question. Despite the broad range of seemingly intelligent manifestations in the natural realm, the key to solve this problem relies on the very same *questioner*, i.e. on narrowing down the focus to the main features of *human* intelligence. In his 1950 seminal work [2], Alan Turing assessed the centrality of behavior to define intelligence: a suitable game needs to be designed where humans and machines have to answer to a human interrogator who is set in a room apart from the players; in this scenario, a machine will be considered intelligent if and only if it would be able to *imitate* human behavior to the extent of not being unmasked by the interrogator. As Turing pointed out, the type of the game is not important: what is central, instead, is that it allows to evaluate humans and machines' *behavior*, by their moves, strategies and, ultimately, answers. In this paper we are neither discussing the philosophical implications of the behaviorist perspective, nor providing a critical analysis of behaviorism with respect to internalism, where the 'faculty of mental representation' (as Kant would name it

[1] Quotation from [1], p. 18.

[3]) becomes a necessary condition for acknowledging intelligence [4]. Rather, we adopt a hybrid framework: trying to overcome the classic tension between task-specific *narrow* AI and task-independent *strong* AI, this article focuses on intelligence as *knowledge in action*, namely as "the combination of cognitive operations and knowledge resources that leads to purposeful behavior" [1]. In particular, we describe an artificial system that benefits from both **mechanism–centered** and **knowledge–centered** approaches to computationally approximate human intelligent behavior in a task of visual recognition[5].

2 Extending ACT-R with a Knowledge Component

Integration is the key to intelligent behavior: learning mechanisms determine which knowledge can be acquired and in which form and specific knowledge contents provide stringent requirements for mechanisms to be able to access and process them effectively. This mutual dependence between mechanism and knowledge is well reflected in the ACT-R cognitive architecture [6], a modular framework whose components include perceptual, motor and memory modules (see figure 1). After a brief introduction of ACT-R core features (section 2.1), we describe how the cognitive architecture can be leveraged by means of a dedicated knowledge compontent (section 2.2), fostering high-level deductive reasoning.

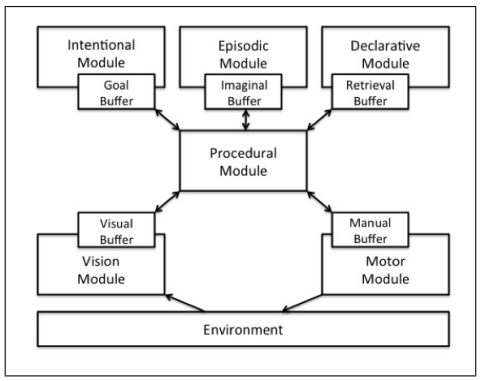

Fig. 1. ACT-R modular structure elaborates information from the environment at different levels.

2.1 ACT-R

ACT-R integrates declarative and procedural knowledge, the latter being conceived as a set of procedures (production rules) that coordinate information processing between its various modules: accordingly, an ACT-R model can accomplish specific goals on the basis of declarative representations elaborated

through procedural steps (in the form of *if-then productions*). At the symbolic level, ACT-R performs two major operations on *Declarative Memory* (DM): i) accumulating knowledge units (i.e., *chunks*) learned from internal operations or from interaction with the environment and ii) retrieving chunks that provide needed information. Both chunk learning and retrieval are performed through limited capacity buffers that constrain the size and capacity of the chunks in DM. ACT-R has accounted for a broad range of cognitive activities at a high level of fidelity, reproducing aspects of human data such as learning, errors, latencies, eye movements and patterns of brain activity (refer to [7] for more details). Although it is not our purpose in this paper to present the details of the architecture, two specific sub-symbolic mechanisms need to be mentioned here to sketch how the system works: i) *partial matching* - the probability that two different knowledge units (or *declarative chunks*) can be associated on the basis of an adequate measure of similarity (this is what happens when we consider, for instance, that a bag is more likely to resemble a basket than a baseball bat); ii) *spreading of activation* - when the same chunk is connected to multiple contexts, it contributes to distributionally activate all of them (e.g., a polysemous word like bag can be associated to different activities like travelling, shopping, eating, etc.)[2].

2.2 SCONE

Inasmuch as humans understand their surroundings by means of coupling perception with knowledge, the ACT-R cognitive architecture should be enabled to generalize over perceptual transductions by applying fine-grained models of the world to concrete scenarios. In order to fulfill this goal however, ACT-R needs to properly encapsulate those models – or *ontologies* – and exploit them for pattern recognition and high-level reasoning. Since ACT-R declarative module supports a relatively coarse-grained semantics based on slot-value pairs, and the procedural system is not optimal to effectively manage complex logical constructs (e.g., 2^{nd} order), a specific extension is needed to make ACT-R suitable to fulfill knowledge-intensive tasks. Accordingly, we engineered an extra module as a bridging component between the cognitive architecture and an external knowledge-base system, SCONE [8]. SCONE is an open–source knowledge-base system intended for use as a component in many different software applications: it provides a LISP-based framework to represent and reason over symbolic common–sense knowledge. Unlike most diffuse KB systems, SCONE is not based on Description Logics [9]: its inference engine adopts marker–passing algorithms [8] (originally designed for massive parallel computing) to perform fast queries at the price of losing logical completeness and decidability. In particular, SCONE represents knowledge as a *semantic network* whose nodes are locally weighted (*marked*) and associated to arcs (*wires*[3]) in order to optimize basic reasoning

[2] Section 3 will show in more details how these two mechanisms can be exploited by an artificial system to disambiguate visual signals

[3] In general, a *wire* can be conceived as a binary relation whose domain and range are referred to, respectively, as A-node and B-node.

tasks (e.g. class membership, transitivity, inheritance of properties, etc). The philosophy that inspired SCONE is straightforward: from vision to speech, humans exploit the brain's massive parallelism to fulfill all recognition tasks; if we want to build an AGI system that is able to deal with the large amount of knowledge required in common-sense reasoning, we need to rely on a mechanism that is fast and effective enough to simulate parallel search. Shortcomings are not an issue since humans are not perfect inference engines either. Accordingly, SCONE implementation of marker–passing algorithms aims at simulating a pseudo-parallel search by assigning specific marker bits to each knowledge unit. For example, if we want to query a KB to get all the parts of cars, SCONE would assign a marker M1 to the A-node CAR and search for all the statements in the knowledge base where M1 is the A-wire (domain) of the relation PART-OF, returning all the classes in the range of the relation (also called 'B-nodes'). SCONE would finally assign the marker bit M2 to all B-nodes, also retrieving all the inherited subclasses[4]. The modularization and implementation of an ontology with SCONE allows for an effective formal representation and inferencing of core ontological properties of world entities. In general we refer to ACT-R including the SCONE module as ACT-RK, meaning 'ACT-R with improved Knowledge capabilities' (the reader can easily notice the evolution from the original ACT-R architecture – figure 1 – to the knowledge-enabled one – figure 2). This integration allows for dynamic queries to be automatically submitted to an external ontology by ACT-RK whenever the perceptual information is incomplete, corrupted or when common-sense reasoning capabilities are needed to generalize over perceptual information filtered from the environment. In this way, ACT-RK is also able to overcome situations with missing input: mechanisms of partial matching and spreading activation [7] can fill the possible gap(s) in the input stream and retrieve the best–matching piece of background knowledge. In particular, in the second part of the paper we describe how an ACT-RK model can perform an action recognition task. Note that the integration of SCONE into ACT-R respects the general cognitive constraints of the architecture, especially in terms of limited-capacity buffers constraining communication between the module and the rest of the architecture. Also, the SCONE marker-passing algorithms are similar to ACT-R spreading activation, leaving open the possibility of a deeper integration of the two frameworks in future work. In principle, if it is true that ACT-R can *per se* deal with simple logical reasoning on the basis of its production mechanisms, when knowledge-intensive tasks come into play an external KBS like SCONE becomes a crucial plug-in for augmenting ACT-R scalability, computational efficiency, and semantic adequacy.

3 Simulating visual intelligence with an ACT-RK model

'Visual intelligence' is the human capability to understand a scene by means of recognizing the core interactions holding between the most salient entities

[4] We refer the reader to [8] for details concerning marker–passing algorithms.

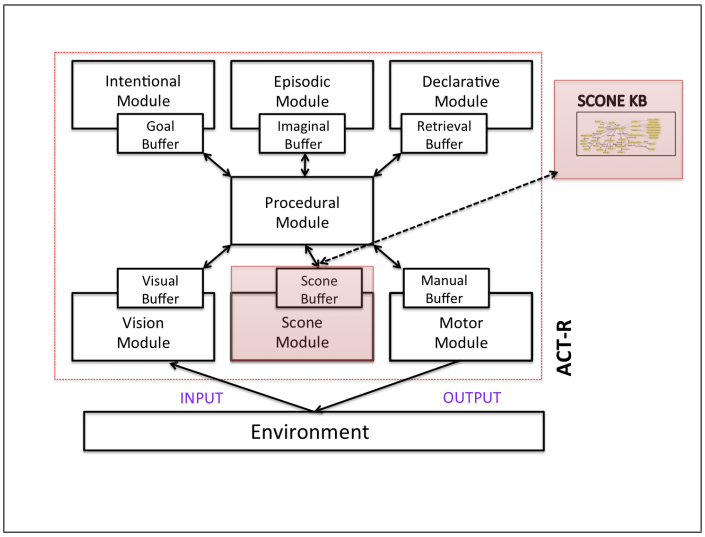

Fig. 2. The ACTR-RK framework

detected from the environment. In this sense, perceptual data, conceptual representations and reasoning are combined together by humans to *make sense* of a scene: for instance, when we *see* a dog chasing a flying stick thrown by a person, first we identify the type of entities into play (dog, person, stick) and then we break the complex event into smaller components (e.g., the person extending the arm from the back, the dog jumping and running, the stick falling on the ground, etc.), inferring its teleological features (make the dog play and bring back the stick) and causal nexus (when the person's hand releases the stick, it starts moving on air with a curved trajectory whose range depends on the exerted force). It is clear that we are not just *seeing* with the eyes but our mental representations and cognitive processing are also involved. Reproducing this capability at the machine level requires a comprehensive infrastructure where low-level visual detectors and algorithms couple with high-level knowledge representations and processing: this is the goal of the DARPA Mind's Eye program[5], where an artificial visual systems is considered to be (*behaviorally*) intelligent if it is able to process a video dataset of various human actions[6] and output the probability distribution (per video) of a pre-defined list of verbs, including 'walk', 'run', 'carry', 'pick-up', 'haul', 'follow', 'chase', 'exchange', 'open', 'close', etc.[7]. Performance is measured in terms of consistency with human responses to stimuli (*Ground Truth*): subjects have to acknowledge the presence/absence of every

[5] `http://www.darpa.mil/Our_Work/I2O/Programs/Minds_Eye.aspx`
[6] `http://www.visint.org/datasets.html`.
[7] This list has been provided by DARPA.

verb in each video. In order to meet these requirements, we devised an ACT-RK model to work in a human-like fashion, trying to disambiguate the scene in terms of the most reliable perceptual and conceptual structures. Because of space limitations, we can't provide the details of a large-scale evaluation: nevertheless, in what follows we discuss an example to describe the functionalities of the system.

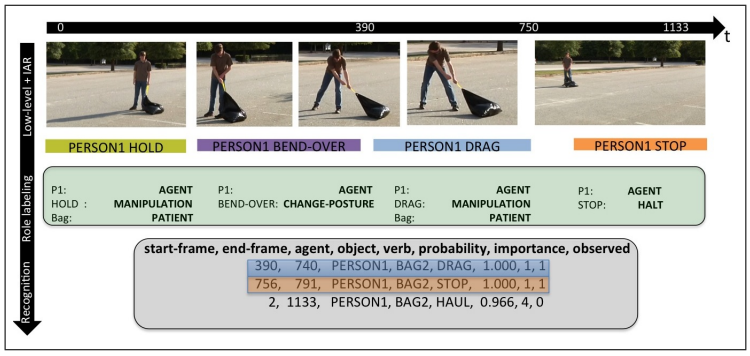

Fig. 3. The horizontal arrow represents the video time frames while the vertical one represents the interconnected levels of processing. The box in the middle displays the results of semantic disambiguation of the scene elements, while the box in the bottm contains the schema of the output, where importance reflects the number of components in a pattern (1-4) and *observed* is a boolean parameter whose value is 1 when a verb matches a visual detection and 0 when the verb is a result of cognitive processing.

Figure 3 schematizes the ACT-RK model core functions, namely to semantically parse temporally-ordered atomic events previously extracted from low-level computer vision systems [10], e.g. 'hold' (micro-state) and 'bend-over', 'drag', 'stop' (micro-actions), associating frames and roles to visual input from the videos. This specific information is retrieved from the HOMinE ('Hybrid Ontology for the Mind's Eye project') ontology, in particular from a fragment of the ontology which has been built on top of the FrameNet lexical resource [11]: frames and semantic roles are assembled in suitable chunk types and encoded in the declarative memory of ACT-RK[8]. As with human annotators performing semantic role labeling [12], the model associates verbs denoting atomic events to corresponding frames. When related mechanisms are activated, the model retrieves the roles played by the entities in the scene, for each atomic event[9]: e.g., 'hold' evokes the *manipulation* frame, whose core role *agent* can be associated to 'person1' (as showed in light-green box of the figure). In order to prompt a choice within the patterns of action encoded in the ontology (see table 1), sub-symbolic

[8] HOMinE has been implemented into SCONE KBS and represents an extension of the SCONE core ontology for action types, as the reader can see in figure 5.

[9] Entities and atomic events are visually recognized using suitable features detectors, object tracking algorithms and SVM classifiers.

computations for *spreading activation* are executed [7]. Spreading of activation from the contents of frames and roles triggers the evocation of related ontology patterns.

Action	Role1	Role2	Role3	Role4	C1	C2	C3	C4
Arrive	self-mover	theme			walk	stop		
Give	agent	carrier	agent		holding	transport	drop	
Take	carrier	agent	agent		transport	drop	holding	
Exchange	agent	agent	agent		give	take	swap	
Carry	agent	carrier	agent		holding	transport	pull	
Pick-up	protagonist	agent	protagonist	agent	bend-over	lower-arm	stand-up	holding
Put-down	agent	protagonist	agent	figure1	holding	bend-over	lower-arm	on
Haul	protagonist	agent	agent	agent	bend-over	extend-arm	holding	drag

Table 1. An excerpt of the roles and atomic components (C1-C4) constituing the patterns of actions for the model

The core sub-symbolic computations performed by the ACT-RK model can be expressed by the equation in figure 4.

$$A_i = \ln \sum_j t_j^{-d} + \sum_k W_k S_{ki} + \sum_l MP_l Sim_{li} + N(0,\sigma)$$

Fig. 4. Equation for Bayesian Activation Pattern Matching

- **1st term**: the more recently and frequently a chunk i has been retrieved, the higher its activation and the chances of being retrieved. In our context i can be conceived as a pattern of action (e.g., the pattern of HAUL), where t_j is the time elapsed since the j^{th} reference to chunk i and d represents the memory decay rate.
- **2nd term**: the contextual activation of a chunk i is set by the attentional weight W_k given the element k and the strength of association S_{ki} between an element k and the chunk i. In our context, k can be interpreted as the value BEND-OVER of the pattern HAUL in figure 3.
- **3rd term**: under partial matching, ACT-RK can retrieve the chunk that matches the retrieval constraints to the greatest degree, combining the similarity Sim_{li} between l and i (a negative score that is assigned to discriminate the 'distance' between two terms) with the scaling mismatch penalty MP. In our context, for example, the value PULL could have been retrieved, instead of DRAG. This mechanism is particularly useful when verbs are continuously changing - as in the case of a complex visual input stream.

– **4th term**: randomness in the retrieval process by adding Gaussian noise.

As mentioned in 2.1, *partial matching* based on similarity measures and *spreading of activation* based on compositionality are the main mechanisms used by the model: in particular, we constrained semantic similarity within verbs to the 'gloss–vector' measure computed over WordNet synsets [13]. Base–level activations of verbs actions have been derived by frequency analysis of the American National Corpus[10]: in particular, this choice reflects the fact that the more frequent a verb, the more likely it is to be activated by our system. Additionally, strengths of associations are set (or learned) by the architecture to reflect the number of patterns to which each atomic event is associated, the so-called 'fan effect' controlling information retrieval in many real-world domains [14]. Last but not least, the ACT-RK model can output the results of extra-reasoning functions by means of suitable queries submitted to HOMINE via the scone module. In the example in figure 3, object classifiers and tracking algorithms could not detect that 'person1' is dragging 'bag2' by pulling a rope: this failure in the visual algorithms is motivated by the fact that the rope is a very thin and morphologically unstable artifact, hence difficult to be spotted by state-of-the-art machine vision. Nevertheless, HOMINE contains an axiom stating that: "For every x,y,e,z such that P(x) is a person, GB(y) is a Bag and DRAG(e,x,y,T) is an event e of type DRAG (whose participants are x and y) occurring in the closed interval of time T, there is at least a z which is a proper part of y and that participates to e"[11]. Moreover, suppose that in a continuation of the video, the same person drops the bag, gets in a car and leaves the scene (see figure 5). The visual algorithms would have serious difficulties in tracking the person while driving the car, since the person would become partially occluded, assume an irregular shape and would not properly lit. Again, ACT-RK could overcome these problems in the visual system by using SCONE to call HOMINE and automatically perform the following schematized inference: 1) cars move; 2) every car needs exactly one driver to move[12]; 2) drivers are persons; 3) driver is located inside a car; 4) a car moves then the person driving it also moves in the same direction. Thanks to the inferential mechanisms embedded in its knowedge infrastructure, the ACT-RK model is not bound to visual input as an exclusive source of information: in human-like fashion, it has the capability of coupling visual signals with background knowledge, performing high-level reasoning and disambiguating the original input perceived from the environment. In particular, the chunks created through the vision module on the basis of computer vision algorithms (schematized by the boxes on top of the video snippets in figure 5) are represented according to suitable chunk types in the declarative memory

[10] http://www.americannationalcorpus.org/

[11] Note that here we are paraphrasing an axiom that exploits Davidsonian event semantics [15] and basic principles of formal mereology (see [16] and [17]). Also, this axiom is valid if every bag has a rope: this is generally true when considering garbage bags like the one depicted in figure3, but exceptions would need to be addressed in a more comprehensive scenario.

[12] With some exceptions, especially in California, around Mountain View!

and used as input to the scone module. That module then becomes an (internal) information source in its own right, treated by the cognitive architecture in a similar way to the (external) visual information stream.

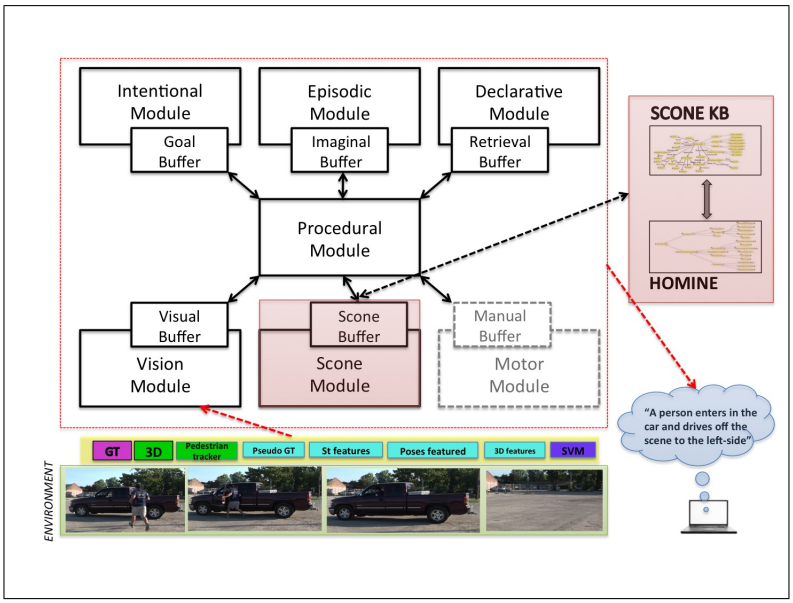

Fig. 5. A Diagram of the ACT-RK model querying HOMinE ontology through SCONE.

4 Conclusion

In this paper we outlined the core infrastructure of a high-level artificial visual intelligent system, focusing on the underlying grounding principles and presenting some functional examples. This system can be conceived as an ACT-RK model, namely an instance of the cognitive mechanisms of ACT-R and of the reasoning operations of SCONE KBS: in this respect, it can be seen as an attempt at accomplishing a complex task on the basis of a general approach to Artificial Intelligence, where cognitive mechanisms are integrated in a knowledge–centered reasoning framework. Future work will be devoted to enrich the knowledge component of the system and using reasoning and statistical inferences to derive and predict goals of agents in performing a given action. Finally, we are exploring the possibility of implementing a core mechanism of abductive reasoning to enable information selection from complex visual streams based on saliency. As we began this article standing on the shoulders of a giant, Alan Turing, no better conclusion could come than from him: *"We can only see a short distance ahead but we can see plenty there that needs to be done"*.

Acknowledgments

This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0061. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Laird, J.E.: The SOAR Cognitive Architecture. The MIT Press (2012)
2. Turing, A.M.: Computing machinery and intelligence. MIND **59**(236) (oct 1950) 433–460
3. Kant, I.: Critique of judgment / Immanuel Kant ; translated, with an introduction, by Werner S. Pluhar ; with a foreword by Mary Gregor. Hackett Pub. Co., Indianapolis, Ind. : (1987)
4. Stich, S.: From folk psychology to cognitive science. The MIT Press, Cambridge, MA (1983)
5. Oltramari, A., Lebiere, C.: Mechanism meet content: Integrating cognitive architectures and ontologies. In: Proceedings of AAAI 2011 Fall Symposium of "Advances in Cognitive Systems". (2011)
6. Anderson, J.: How Can the Human Mind Occur in the Physical Universe? Oxford University Press (2007)
7. Anderson, J., Lebiere, C.: The Atomic Components of Thought. Erlbaum (1998)
8. Fahlman, S.: Using scones multiple-context mechanism to emulate human-like reasoning. In: First International Conference on Knowledge Science, Engineering and Management (KSEM'06), Guilin, China, Springer–Verlag (Lecture Notes in AI) (2006)
9. Baader, F., Calvanese, D., Mcguinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: The Description Logic Handbook : Theory, Implementation and Applications. Cambridge University Press (2003)
10. Maitikanen, P., Sukthankar, R., Hebert, M.: Feature seeding for action recognition. In: Proceedings of International Conference on Computer Vision. (2011)
11. Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C.: Framenet: Theory and practice (June 2005)
12. Gildea, D., Jurafsky, D.: Automatic labelling of semantic roles. In: Proceedings of 38 th Annual Conference of the Association for Computational Linguistics (ACL-00). (2000) 512–520
13. Pedersen, T., Patwardhan, S.J., Michelizzi, M.: Wordnet :: Similarity: Measuring the relatedness of concepts. In: Demonstration Papers at HLT-NAACL. (2004) 38–41
14. Schooler, L., Anderson, J.: Reflections of the environment in memory. Psychological Science **2** (1991) 396–408
15. Casati, R., Varzi, A., eds.: Events. Dartmouth, Aldershots, USA (1996)
16. Simons, P., ed.: Parts: a Study in Ontology. Clarendon Press, Oxford (1987)
17. Casati, R., Varzi, A.: Parts and Places. The Structure of Spatial Representation. MIT Press, Cambridge, MA (1999)