Unsupervised model for grounding multimodal representations

Michal Vavrečka, Igor Farkaš
FEL, Czech Technical University Prague, FMPI, Comenius University Bratislava,
vavrecka@fel.cvut.cz, farkas@fmph.uniba.sk

Abstract: The goal of our study is to test unsupervised learning methods in the process of grounding color, shape and spatial relation of two objects in 2D space. Our model consists of several types of neural networks (Self-Organizing Map, RecSOM and Neural Gas) that constitute multimodal architecture being able to integrate information from visual and auditory inputs. Five-word sentences describing the scene (e.g. "Red box above green circle") served as auditory (linguistic) inputs with phonological encoding and the described scenes were presented as the visual inputs (to an artificial retina). The visual scene was represented in SOM and the auditory description was processed by RecSOM, a recurrent SOM-based model for processing sequences. Both these primary representations were integrated in a multimodal module (SOM or NG) in the second layer. We tested this two-layer architecture with several modifications (visual SOM representing color shape and spatial relations vs. separate SOM for spatial relations and for shape and color, imitating *what* and *where* pathways) and several conditions (scenes with varying complexity with 2-3 colors, 2-5 object shapes and 4 spatial relations). In the scenes with higher complexity we reached better results in case of using NG algorithm in the multimodal layer compared to SOM. The poorer result of multimodal SOM could be attributed to fixed neighborhood function which is relaxed in NG algorithm. The results confirm theoretical assumptions about different natures of visual and auditory coding and effeciently integrates them reflecting their specific features.

Motivation

The overall goal of our project is design of modular architecture that demonstrates sensory-motor learning, social learning and language learning. We **focus on neural network approach**, with a dominant role of **self-organizing maps** that are adopted in various parts of the model. Our emphasis on unsupervised learning is cognitively appealing and will be coupled with supervised learning only in cases when the teaching signal originates from the environment itself (e.g. linguistic labels entered via auditory modality), rather than being provided by an external teacher (designer), to avoid the grounding problem.

Main objective is to test the limitations of machine learning in the process of building representations solely from the sensory inputs to propose a hierarchical architecture that is able to represent information from different modalities and to find the mapping between unimodal representations. This approach imitates the nature of human learning capabilities within the development.

We test the radical version of embodied cognition (Varela et al., 1991), arguing that the co-occurrence of inputs from the environment is a sufficient source of information to create an intrinsic representational system. These representations preserve constant attributes of the environment. **We propose an alternative solution to the classical grounding architecture** (Harnad, 1990). The difference is the way of processing symbolic input by a separate auditory subsystem and further integration of auditory and visual information in a multimodal layer. Multimodal layer incorporates the process of identification. Our approach is similar to "grounding transfer" (Riga, Cangelosi and Greco, 2004) based on SOM maps and supervised multi-layer perceptron, but our system works in fully unsupervised manner that implies different way of symbolic level creation.

Our model is based on hierarchical processing. Top (multimodal) level is justified and modified from both modalities (the sequential "symbolic" auditory level and the parallel "conceptual" visual level). The separate multimodal level provides platform for the development of subsequent stages of the system (inference mechanisms etc.).

The models

We focus on learning spatial locations of two objects in 2D space and their linguistic description. This conceptual level is represented by the visual subsystem, the symbolic level is represented by the auditory system.

We tested two versions of the visual subsystem, keeping in mind the distinction between "*what"* and "*where"* subsystems (Ungerleider, Mishkin, 1982). The former learns to represent object features (shape and color), the latter object position. There were two models proposed for this task.

Model I: single SOM learns to capture both 'what' & 'where' information.
Model II: 'what' and 'where' systems are separated.

Visual layer

The visual subsystem is formed by artificial retina (28x28 neurons) that projects to primary (unimodal) visual layer, implemented by the unsupervised SOM, to build representations of visual scenes in topographic manner. The SOM was expected to differentiate various positions of two objects, as well as object types and their color in Model I. Model II consists of separate SOM for spatial locations (resembling *where* system) and separate SOM for color and shape of objects (resembling *what* system). The SOM was trained for 100 epochs (training data size varied from 6400 to 45000 depending on the complexity of environment) with decreasing parameter values (neighborhood radius, learning rate).

Auditory layer

Auditory input (English sentence), in the form phonetic feature vectors feeds into the primary RecSOM (Voegtlin, 2002), a recurrent SOM-based architecture, that learns to represent inputs (words) in temporal contexts (hence capturing sequential information). The sentences are presented one word at a time. RecSOM output, in terms of map activation, feeds to the multimodal layer, to be integrated with the visual pathway. Like SOM, RecSOM is trained by competitive, Hebbian-type of learning.

Multimodal layer

In agreement with the theory of perceptual symbol systems (Barsalou, 1999), a multimodal layer forms the core of the system. Its role is to identify unique categories and represent them by merging different sources of information.

We tested two implementations of the multimodal layer: a SOM and Neural Gas. The layer size is set to allow the unique localist representation of all combinations of objects in the considered scenario (29x29 = 841 neurons).

Inputs for this layer are taken as unimodal SOM activations (from both modalities) using the *k*-WTA mechanism (*k* most active units are proportionally turned on, all other units are reset to zero). The output representation in the multimodal layer is localist for better interpretation of results.

Training data

There were inputs with increasing complexity presented to the system starting from 2 colors, 2 objects and 4 spatial locations. The most difficult task involves 3 colors (red, green, blue), 5 object types (box, ball, table, cup, bed,) and 4 spatial relations. Unlike color, spatial relations and are fuzzy in the visual (conceptual) domain (with prototypes *above, below, left, right,* for which linguistic labels exist).

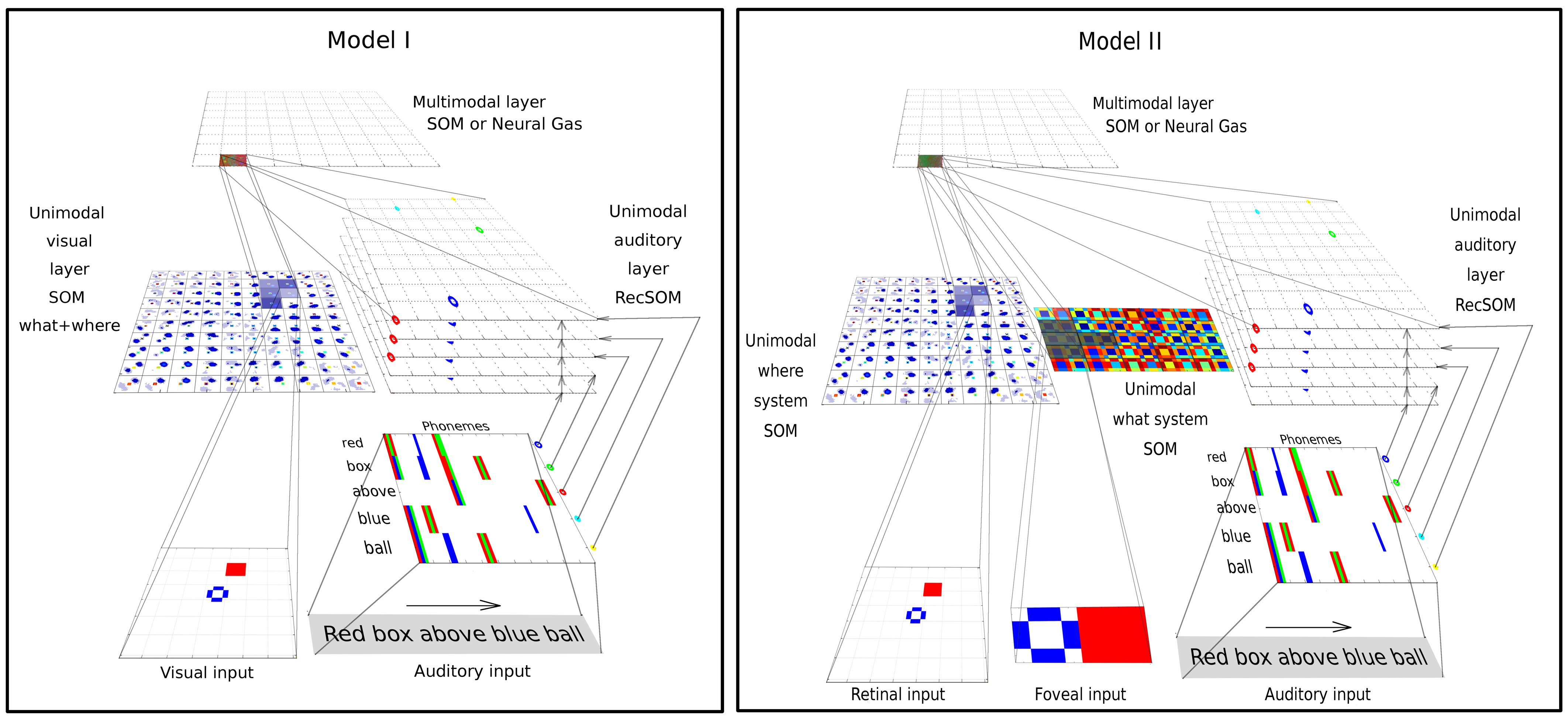

Fig 1. Comparison of the Model I and II. There is the separate layer for representing shape and color in Model II

Results

We trained the system for each size of unimodal SOM for 100 epochs and tested it using a novel set of inputs. Then we measured the effectiveness of this system, based on the percentage of correctly classified test inputs. The error rate was lower for the auditory layer compared to the visual subsystem in the Model I. This should be attributed to the smaller variability of the inputs because there is the same sentence, describing the spatial location of the trajector that varies in the different position in the specific area. There is also a difference between the *what* and *where* system effectiveness in the Model II. The *where* system is more accurate in the representation of spatial locations and the *what* system represents color and shape of trajector with smaller error.

We obtain best results for the Model II and Neural Gas in the multimodal layer.. There is a visualization for the environment with 3 colors, 3 shapes and 4 spatial terms in Fig.2. In the most complex scenario the error rate for the Neural Gas was 20% on average compared to 70% for the SOM map. The poorer result of multimodal SOM could be attributed to fixed neighborhood function which is relaxed in NG algorithm. The poorer result of multimodal SOM could be attributed to fixed neighborhood function which is relaxed in NG algorithm.

Discussion

Our system is able to map the words in the sentence with fixed grammar to the objects in the environment without any prior knowledge in fully unsupervised manner. In this task the system was able to solve the **binding problem.** The system learns to unambiguously map the pairs of output vectors (from the unimodal layers) to single units. This mapping is actually a clustering process. If (at least) one perceptual input source creates discrete clusters, successful clustering learning can be achieved. In case of two fuzzy sources of information, it is difficult to create a system that is able (without any additional information) to provide a successful mapping, e.g. to learn a new meaning of spatial position in the middle of two spatial areas (below and right) and the auditory information, i.e. "beright."

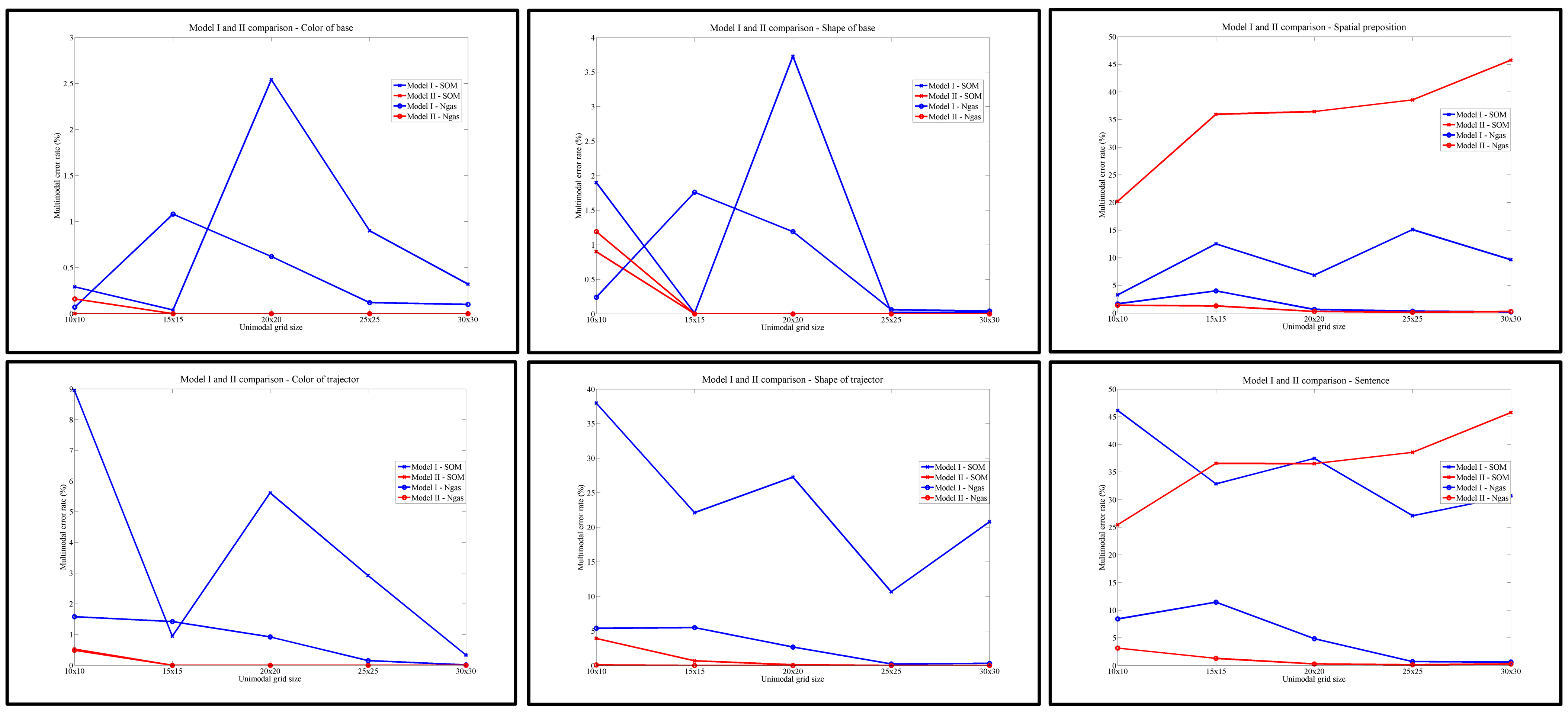

Fig 2. Error rates of the multimodal layer. Model I and II are compared with respect to the specific algorithms (SOM and Neural Gas). Each window shows the error rate for the specific part of the sentence (trajector color, trajector shape, spatial term, base color, base shape) and also for the whole sentence. These are results for the medium-complex environment with 3 colors, 3 shapes and 4 spatial terms.

Future steps

- Representation of causal relations both in visual (spatiotemporal) and auditory (linguistic) domains
- Inference processes based on multimodal layer
- Backward transfer from the symbolic to the conceptual level
- Representation of reference frames and frame-dependent linguistic descriptions of the scene
- Representation of homonymic words (see Fig. 3)
- Hierarchical representation of abstract terms
- Implementation into a robotic simulator and real robot for testing in the real environment (see Fig 4.)
-

We are open for cooperation and looking for partners.

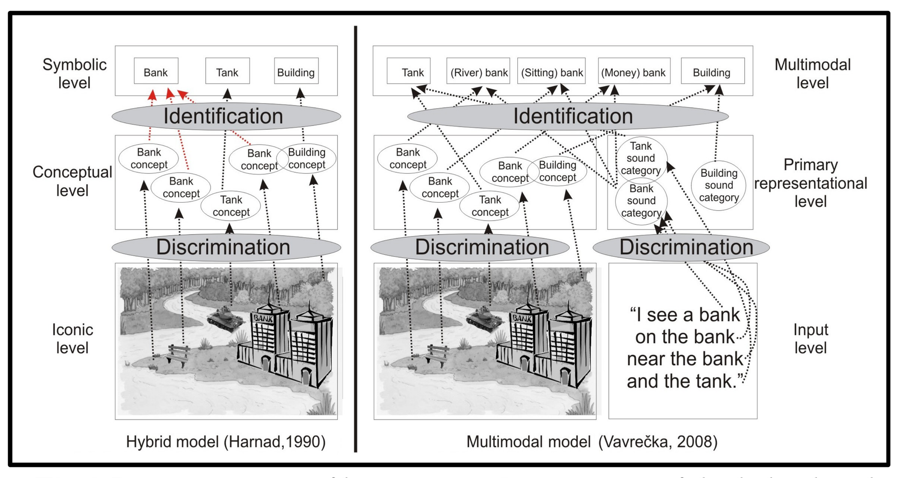

Fig 3.Representation of homonyms. Comparison of the hybrid and the multimodal architectures.

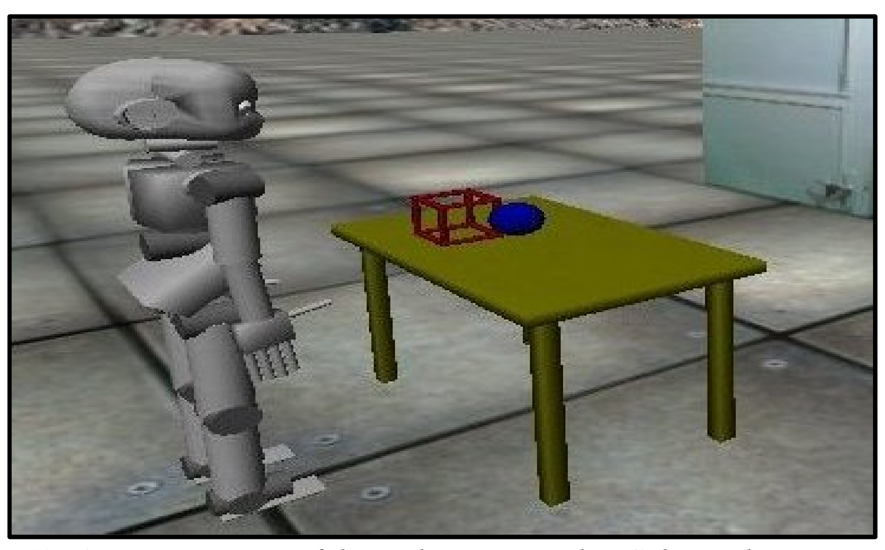

Fig 4.Impementation of the architecture to the iCub simulator

Acknowledgment

This research has been supported by the research program MSM 6840770012 of the CTU in Prague, Czech Republic and by SAIA scholarship (M.V.) and by Slovak Grant Agency for Science, no. 1/0361/08 (I.F.)

References

Barsalou, L. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22:577-609.
Harnad, S. (1990). The Symbol Grounding Problem. Physica D, 335-346.
Riga, T., Cangelosi, A., Greco, A. (2004). Symbol grounding transfer with hybrid self-organizing/supervised neural networks. In International Joint Conference on Neural Networks, Budapest, pp. 2865-2869.
Ungerleider, L., Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M.A.Goodale, & R. J. W. Mansfield (Eds.), Analysis of visual behaviour (pp. 549-586). Cambridge, MA: MIT Press.
Varela, F., Thompson, E., Rosch, E. (1991). The Embodied Mind - Cognitive Science and Human Experience. Cambridge, MA: MIT Press / Bradford Books.
Vavrečka, M. (2008). Reprezentace významu pomocí multimodálních reprezentací. Kognice a umělý život VIII , Praha: VŠE
Voegtlin, T. (2002).Recursive self-organizing maps. Neural Networks. 15(8-9):979-991.

Authors
Michal Vavrečka *is a psychologist and works as the research fellow in the Gerstner Laboratory (CTU Prague). He combines methods of psychology, neuroscience and computational modeling in the area of spatial cognition.*
More info: http://bio.felk.cvut.cz/~vavrecka/
Igor Farkaš *is a computer scientist, developing cognitive science research and education at the Faculty of mathematics, physics and informatics (CU in Bratislava). He specializes on connectionist modeling of various cognitive tasks, including language processing and recently also robotics.*
More info: http://ii.fmph.uniba.sk/~farkas/

Presented at 3rd EUCog members conference, "Multi-sensory integration", 8-9.10.2010, Palma de Mallorca, Spain