Unsupervised Grounding of Spatial Relations

Michal Vavrečka (vavrecka@fel.cvut.cz)
Department of Electrical Engineering, Czech Technical University
Prague, Czech Republic

Igor Farkaš (farkas@famph.uniba.sk)
Department of Applied Informatics, Comenius University
Bratislava, Slovak Republic

Abstract

We present an unsupervised connectionist model for grounding color, shape and spatial relations of two objects in 2D space. The model constitutes a two-layer architecture that integrates information from visual and auditory inputs. The images are presented as the visual inputs to an artificial retina and five-word sentences describing them (e.g. "Red box above green circle") serve as auditory inputs with phonological encoding. The visual scene is represented by the Self-Organizing Map(s) and the auditory description is processed by a recursive SOM (RecSOM) that learns to topographically represent sequences. Primary representations are integrated in a multimodal module (implemented by SOM or Neural Gas algorithms) in the second layer using self-organizing units with conjunctive representations. We tested this two-layer architecture in two versions (a single SOM representing color, shape and spatial relations vs. biologically inspired separate SOMs for spatial relations and for shape and color) and several conditions (scenes with varying complexity up to 3 colors, 5 object shapes and 4 spatial relations). In the scenes with higher complexity we reached better results with NG algorithm in the multimodal layer compared to SOM, which is thank to the flexible neighborhood relations in NG algorithm, relaxing topographic organization. The results confirm theoretical assumptions about the different nature of visual and auditory coding. Our model is hence able to efficiently integrate the two sources of information while reflecting their specific features.

Keywords: spatial terms; unsupervised symbol grounding; self-organization; multimodal representation

Introduction

In applied artificial intelligence, the continuing and challenging problem is the design of an adaptive system that interacts with the environment and is able to understand its internal representations. These representations should capture important attributes of the environment, store them as concepts, and connect these to the symbolic level. This idea is related to the perceptual theory of cognition where the representation is created from the perceptual inputs (sensory modalities). As the next step we should integrate representations from different modalities to the multimodal level. Barsalou (1999) introduced this principle as the Perceptual Symbol System. The integration of modalities is also described in the area of neuroscience. For instance, Damasio (1989) postulates convergence zones, which integrate the information from sensory maps and represent them. The convergence zones create hierarchical levels of associations from the specific modalities. There are similar ideas in cognitive psychology. The dual coding theory (Paivio, 1986) describes two independent but connected representational systems, which create internal representations of the external environment – the verbal and image codes.

In the mentioned theories we should identify some ideas, which are fruitful for the following discussion about the perceptually formed conceptual level and the way how to ground symbols to these concepts. Theories about integration of modalities should help us understand the symbol grounding process. The basic task for symbol grounding is to find the function and the internal mechanism to create representations that are intrinsic to the system and do not need to be interpreted by an external observer (Ziemke, 1999). Our approach presents the radical version of the symbol grounding architecture based on the theory 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 (Vavrečka, 2009). These representations preserve constant attributes of the environment. As opposed to the classical grounding architecture (Harnad, 1990), we propose an alternative solution by processing the symbolic input by a separate auditory subsystem and by further integration of auditory and visual information in a multimodal layer. In terms of Harnad's theory, the multimodal layer incorporates the process of identification. Our approach is similar to the "grounding transfer" (Riga et al., 2004) based on self-organizing maps (SOM; Kohonen 2001) and the supervised multi-layer perceptron, but our system works in a fully unsupervised manner implying a different way of incorporating the symbolic (lexical) level. In addition, our model goes beyond the single word processing, because it can process sentences. We test our model in the area of spatial orientation, similar to Regier (1996) who created a neural network model consisting of several modules to ground the spatial terms, but trained in supervised manner. Our main goal is to extend the research in this area by application of unsupervised learning algorithms where the target signal only functions as an additional input rather than being the source for error-based learning. In general, neural networks provide a natural computational framework for grounding knowledge acquisition.

The Models

We focus on learning spatial locations (of two objects in 2D space) which is a cognitive domain where perceptual information can be linked to the linguistic description. The conceptual level is represented by the visual subsystem and 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* pathways (Ungerleider & Mishkin, 1982). The former learns to represent object features (shape and color), the latter focuses on object position. To appreciate the importance of individual representations in the modules, we tested two models for this task. Model I contains a single SOM that learns to capture both *what* and *where* information. Model II is an extended version of Model I with separate SOMs for processing *what* information (color and shape) and *where* information. Figure 1 provides the sketch of Model II.

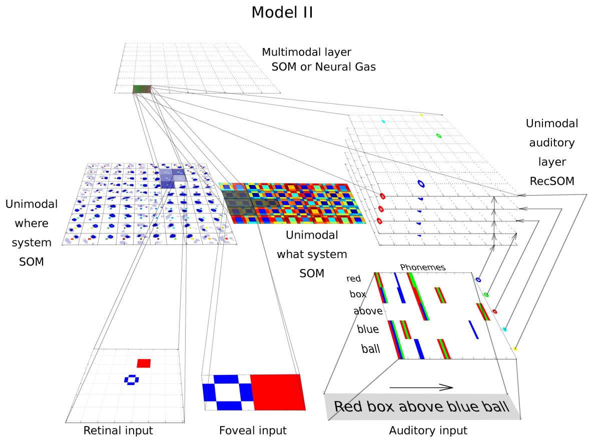

Figure 1: Visualization of information processing in Model II. There is a separate layer (unimodal *what* system) for representing shape and color, as opposed to Model I (where the two parts are merged).

Training data

The input scenes consist of the trajector and the base object in different spatial configurations. The position of the base is fixed in the center of the scene and the trajector is located in one of the spatial quadrants relative to the base. These are linguistically referred to as *up, down, left,* and *right* but perceptually, the trajector position is fuzzy. We trained the models using scenes with increasing complexity, starting with simple inputs with 2 colors, 2 object types and 4 spatial relations, up to more complex inputs consisting of 3 colors (red, green, blue), 5 object types (box, ball, table, cup, bed) and 4 spatial relations (above, below, left, right). The training data size varied from 6400 to 42000 examples (50-100 examples of each spatial configuration) depending on the complexity of the environment.

Visual layer

The sensory input of the visual subsystem is formed by an artificial retina with 28×28 receptors that projects to the primary visual layer. Visual layer consists of the SOM(s) that learn the nonlinear mapping of input vectors to output units in the topography preserving manner (i.e. similar inputs are mapped to neighboring units in the map). 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 a separate SOM for spatial locations (resembling *where* system) and for color and shape of objects (resembling *what* system). The *what* system stands for simplified attentional mechanisms and the foveal input. There are two visual fields (with 4×4 receptors each) that project visual information about the trajector and the base in fixed position to the unimodal what system SOM. The color of each pixel was encoded by the activity level normalized to values between 0 and 1. Both maps were trained for 100 epochs with decreasing parameter values (unit neighborhood radius, learning rate).

Auditory layer

Auditory inputs (English sentences) were encoded as phonological patterns representing word forms using PatPho, a generic phonological pattern generator that fits every word (up to trisyllables) onto a template according to its vowel-consonant structure (Li & McWhinney, 2002). It uses the concept of syllabic template: a word representation is formed by combinations of syllables in a metrical grid, and the slots in each grid are made up by bundles of features that correspond to consonants and vowels. In our case, each sentence consists of five 54-dimensional vectors with component values in the interval (0,1). These vectors are sequentially fed (one word a time) to the RecSOM (Voegtlin, 2002), a recurrent SOM architecture, that uses a detailed representation of the context information (the whole output map activation) and has been demonstrated to be able to learn to represent much richer dynamical behavior (Tiňo et al., 2006), compared to other recurrent SOM models (Hammer et al., 2004). RecSOM learns to represent inputs (words) in the temporal context (hence capturing sequential information). 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-like algorithm. As a property of RecSOM, its units become sequence (sentence) detectors after training, topographically organized according to the suffix (most recent words).

Since RecSOM, unlike SOM, is not common, we provide its mathematical description for interested readers here. Each neuron $i \in \{1,2,...,N\}$ in RecSOM has two weight vectors associated with it: $\mathbf{w}_i \in \mathbb{R}^n$ – linked with an n-dimensional input $\mathbf{s}(t)$ (in our case, the current word, $n = 54$) feeding the network at time t and $\mathbf{c}_i \in \mathbb{R}^N$ – linked with the context $\mathbf{y}(t-1) = [y_1(t-1), y_2(t-1), ..., y_N(t-1)]$ containing map activations $y_i(t-1)$ from the previous time step.

The output of a unit i at time t is $y_i(t) = \exp(-d_i(t))$, where

$$d_i(t) = \alpha \cdot \|\mathbf{s}(t) - \mathbf{w}_i\|^2 + \beta \cdot \|\mathbf{y}(t-1) - \mathbf{c}_i\|^2.$$

Here, $\|\cdot\|$ denotes the Euclidean norm, $\alpha > 0$ and $\beta > 0$ are model parameters that respectively influence the effect of the input and the context upon neuron's profile. Their suitable values are usually found heuristically (in our model, we used $\alpha = \beta = 0.1$). Both weight vectors are updated using the same form of SOM learning rule:

$$\begin{aligned} \Delta \mathbf{w}_i &= \gamma \cdot h_{ik} \cdot (\mathbf{s}(t) - \mathbf{w}_i), \\ \Delta \mathbf{c}_i &= \gamma \cdot h_{ik} \cdot (\mathbf{y}(t-1) - \mathbf{c}_i), \end{aligned}$$

where k is an index of the best matching unit at time t, $k = \text{argmin}_i\{d_i(t)\}$, and $0 < \gamma < 1$ is the learning rate. Note that the best matching ('winner') unit can be equivalently defined as the unit k of the highest activation $y_k(t)$: $k = \text{argmax}_i\{y_i(t)\}$. Neighborhood function h_{ik} is a Gaussian (of width σ) on the distance $d(i,k)$ of units i and k in the map: $h_{ik} = \exp(-d(i,k)^2/\sigma^2)$. The 'neighborhood width', σ, linearly decreases in time to allow for forming topographic representation of input sequences. After training, all RecSOM units become sensitive to particular sentences, ordered topographically according to sentence endings.

Multimodal layer

The units in the multimodal layer identify and represent unique categories (if successful). In agreement with the theory of perceptual symbol systems (Barsalou, 1999), the main task for the multimodal layer is to process the output from the unimodal layers and to find and learn the categories by mapping and merging different sources of information (visual and auditory) that refer to the same object in the external world.

The multimodal layer is formed exploiting the concept of self-organized *conjunctive representations* that have been hypothesized to exist in the brain with the purpose of binding the features such as various perceptual properties of objects Mel & Fiser (2000). Here we extend this concept by linking the subsymbolic and symbolic information. Hence, each output units learns to represent a unique combination of perceptual and symbolic information (that could be forwarded to another, higher module).

We tested two versions of the multimodal layer, namely SOM and Neural Gas (NG) algorithms (Martinetz & Schulten, 1991). Both algorithms are unsupervised, based on the competition among units and the same cooperative learning rule, but NG uses a flexible neighborhood function, hence relaxing topographic organization, as opposed to fixed (2D) neighborhood relations in SOM. The size of the multimodal layer was set to allow a distinct localist representation of all 840 object combinations (3 colors, 5 object types, 2 objects in scene, 4 spatial terms) in the most complex data set, so we used 841 neurons (arranged in a 29×29 grid in case of SOM). [1] Inputs for this layer are outputs of unimodal first-layer modules (concatenated from both modalities) using the k-WTA (i.e. winner-takes-all) mechanism, where k most active units are proportionally turned on (with the activity of the best matching unit rescaled to 1), and all other units are reset to zero (in the models, we used $k = 6$). The motivation for this type of output representation consists in introducing some overlaps between similar patterns to facilitate generalization. On the other hand, the output representation in the multimodal layer is chosen to be localist for better interpretation of results and the calculation of error rate.

Results

We trained the system with fixed size of the multimodal layer and varying size of the unimodal maps (ranging from 10×10 to 35×35 neurons) for 100 epochs and tested the models using a novel set of inputs. The size of the testing set was 30% of the overall data set. To calculate the accuracy of neuron responses, we applied a voting algorithm after training to label each neuron in the layer based on its most frequent response. Then we measured the accuracy of this system, based on the percentage of correctly classified test inputs.

Regarding primary modules, there were low error rates in the auditory unimodal layer in both Models I and II, and also in *what* unimodal system in Model II. This was thank to the smaller variability of the inputs because the same sentence describes the spatial location presented to the primary auditory system and the objects are presented to the *what* system without spatial variability. On the other hand, we observed a high error rate in *where* system for the the trajector shape that varied in the different positions of the specific area. There was also a difference between the *what* and *where* system accuracy in Model II. The *where* system was more accurate in the representation of spatial locations and the *what* system represents color and shape of trajector with a smaller error, because there are specific inputs projected to the particular subsystems.

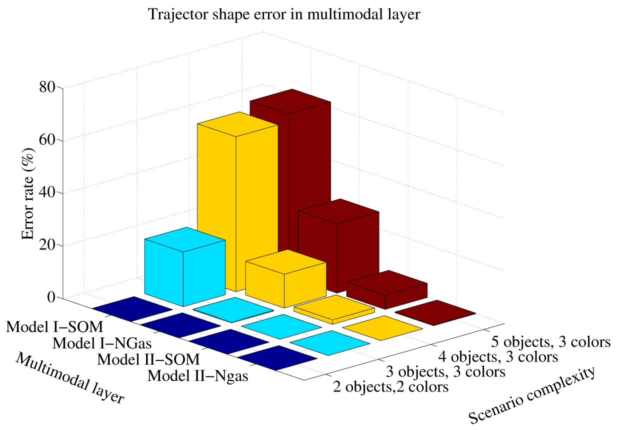

Figure 2: Error rates in the multimodal layer for the trajector shape representation. Model II based on Neural Gas performs best.

The analysis of the model behavior at the multimodal layer revealed that the trajector shape and the spatial term representations are the most difficult tasks for visual unimodal systems, caused by the variability and fuzziness of these inputs. The results are depicted in Figures 2 and 3. These refer to

[1] It could be argued that we used high-level knowledge about data for choosing the suitable number of neurons in the multimodal layer. That is true, we wanted to avoid unnecessarily a large number of units and save the training time; any larger layer would have worked fine as well.

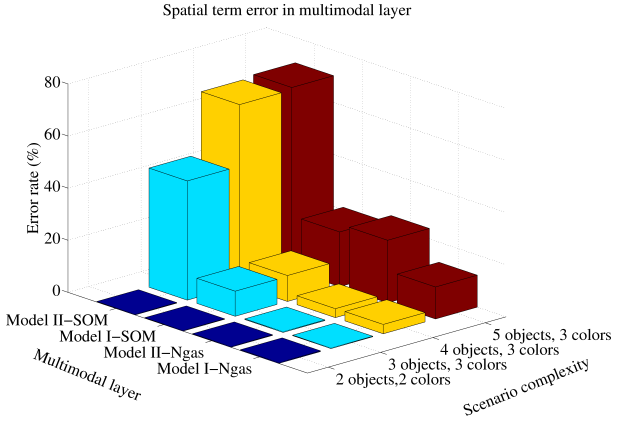

Figure 3: Error rates in the multimodal layer for the spatial term representation. Model I based on Neural Gas performs best.

the model with unimodal maps having 30×30 neurons and the multimodal layer consisting of 841 neurons. We compare two Models (I and II), using one of the two algorithms (SOM and NG) compared in case of scenes with increasing complexity. The error rate for the representation of trajector color, base color and base shape was very low in all scenarios and the results were similar for both models and algorithms.

We also tested the error rate of the sentence representation based on the correct representation of all five labels (Figure 4). We obtain the best results for Model II with NG algorithm in the multimodal layer. In the most complex scenario, the error rate for the NGs was 20% on average compared to 70% for the SOM. The poorer result of multimodal SOM can be attributed to the fixed neighborhood function which imposes constraints on the learned nonlinear mapping and hampers unit differentiation.

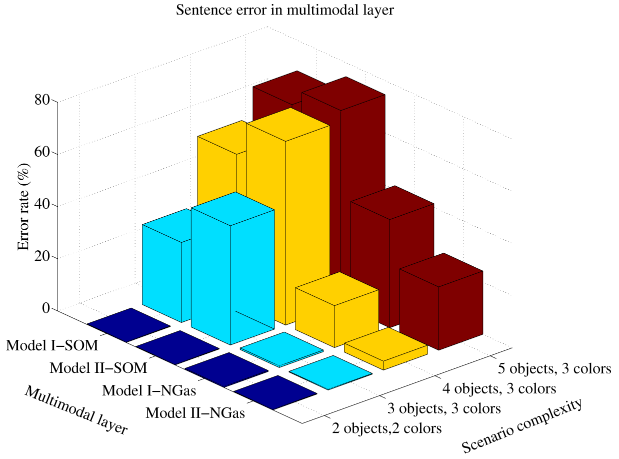

Figure 4: Error rate in the multimodal layer for the whole scene/sentence representation. Model II based on Neural Gas performs best.

Discussion

The results presented in Figures 2-4 reveal some differences in the tested models. The correct representation of the trajector shape in the multimodal layer (Figure 2) requires the separate unimodal *what* system. The error rates for Model I and II confirm the necessity of the *what* system in the complex environment because there is a 60% increase of errors in the model without separate *what* system. There is also a difference in the effectiveness of the SOM and NG in the multimodal layer. In both models the NG algorithm yields lower error rates. The higher error rate in SOM is probably caused by its fixed neighborhood function that imposes additional constraints on the learned mapping. There is a different type of clustering in unimodal layers that are transferred to multimodal layer. The SOM algorithm based on the fixed neighborhood function is not able to adapt to the joint outputs from unimodal layers.

On the other hand, there is a problem to represent spatial terms both for NG and SOM algorithms in Model II. The inputs are taken from 3 unimodal layers and we observe a higher error rate for both algorithms in most complex scenario compared to Model I. This is caused by the missing information about the spatial term in *what* system that projects to the multimodal layer. This results in poor categorization of Model II. These are contradictory results, because Model II performs better for trajector shape but yields higher error rates for the spatial term representation.

To compare the overall accuracy of both models we needed to analyze the whole sentence processing (Figure 4). The best results were obtained for Model II using NG algorithm in the multimodal layer. These results are consistent with our expectation about the advantages of the relaxed neighborhood function of NG. Hence, the better results are achieved with NG by sacrificing the topographicity of responses in the multimodal layer. From the biological perspective, it remains an open question, whether brain organizes its responses topographically also at higher levels of organization.

We can also conclude that it is advantageous to follow the biologically inspired hypothesis about processing of visual information in separate subsystems. The question for the future research is to find a proper way of output coding from unimodal layers to increase system accuracy.

The mapping in our models is actually a clustering process. If (at least) one perceptual input source creates discrete clusters, successful learning can be achieved. In case of all 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." In other words, the successful clustering presumes that at least one modality provides distinct activation vectors for different classes to drive the clustering process (i.e. the classes are well separable in the corresponding input subspace).

Some theoretical aspects

Let us also look at some theoretical aspects of the presented architectures. The representation process takes advantage of the two or three unimodal layers of units. The auditory layer represents unique labels (linguistic terms), the *where* system represents fuzzy information about the spatial locations of objects in the external world and *what* system captures shapes and colors of objects in fixed foveal position. The multimodal level integrates the outputs of these unimodal layers. In contrast to the classical approaches that postulate the abstract symbolic level as fixed and prior (defined by the designer), in our model it is possible to learn and modify all levels of representation. The meaning is simultaneously represented by all layers (auditory, visual and multimodal), making this approach resemble the theory of Peirce (1931) who defined three components of a sign – representamen, interpretant and the sign itself. This contrasts with the classical symbolic approach, where interpretant (concept) is made of arbitrary symbols (e.g. frames, semantic networks) or it is missing. Our proposed architecture satisfies the requirement that the artificial system (agent) should learn its own functions and representations (Ziemke, 1999). In contrast to the classical top-down approach, our bottom-up approach restricts the designers intervention into the representational system to a minimum. Representations are learned from the external environmental inputs in a completely unsupervised manner.

Our model assumes the existence of the higher layer that integrates the information from two primary modalities. This assumption makes the units in the higher layer bimodal (i.e. they can be stimulated by any of the primary layers) and their activation can be forwarded for further processing. Bimodal (and multimodal) neurons are known to be ubiquitous in the association areas of the brain (Stein & Meredith, 1993).

Relation to other connectionist architectures

Interestingly, the bimodal layer with conjunctive units is also used in generative probabilistic models that can be designed to link information from two (or more) modalities. For example, the deep belief net (DBN) is a stochastic generative model (a multi-layer neural network with the bidirectional connections) that learns to approximate the complex joint probability distributions of high-dimensional data in a hierarchical way. DBN was trained to classify the isolated handwritten digits into 10 categories, so the visual inputs (28×28 pixel images) were to be linked with categorical labels (Hinton et al., 2006). The linking was established via the training on image-label pairs, using the higher (bimodal) layer (with 2000 units) that learned the joint distribution of those pairs. DBN was shown to be superior to various other (discriminatory) models in this classification task. From the perspective of the representations formed in the multimodal units, their goal was the same as ours (although our units are deterministic rather than stochastic).

Our model also shares some similarities with the DevLex model of early lexical acquisition (Li et al., 2004). DevLex, originally inspired by the DISLEX model (Miikkulainen, 1997) also consists of self-organizing maps, but these are directly interconnected, rather than projecting their outputs to a higher, multimodal layer. DevLex was proposed to learn the form-meaning associations (phonological word forms and meanings) via Hebbian updating the (bidirectional) connection links, aiming to model the processes of lexical comprehension and production. DevLex does not contain a higher (e.g. multimodal) layer that integrates the modalities, as other grounding models (Riga et al., 2004; Roy, 2005). Instead, the overall representation of the meaning is in DevLex taken as the joint co-activation in the two maps. At the same time, each map has a capacity to activate the other map, yielding the overall representation. However, direct linking of the sensory modalities is also based on neuroscientific rationale, because the brain is known to have these direct connections as well (see e.g. Allman et al. 2009).

Our model is also very similar to the model of Dorffner et al. (1996). They created a connectionist system consisting of two primary levels (symbolic and conceptual) connected to one central layer. There is a linking layer (the counterpart of our multimodal layer) interconnecting the two primary layers via localist units that link both representations (i.e. one unit connects one word-concept pair of primary representations). First, one set of links (weights to the linking layer) is trained using a competitive mechanism exploiting the winner-take-all approach. Then, the winners weights towards the other layer are updated according to the outstar rule (Grossberg, 1987). Hence, the purpose is to learn form-concept mapping, mediated by the linking layer. Regarding DevLex, the similar mapping was obtained by connecting the two SOMs directly. In both models, these mappings were aimed at simulating the word comprehension (form-to-meaning) and the word production (meaning-to-form).

The mentioned models deal with lexical level but our model goes beyond words because it is able to represent sentences with fixed grammar via RecSOM map. It finds the mapping of the particular words to the concepts in the multimodal layer without any prior knowledge, so the system proposes the solution to the binding problem.

Conclusion

Our model proposes a solution to the binding problem by establishing a self-organized mapping between the concept and the symbol. The system design allows us in principle to append other modalities into this system and still represent discrete multimodal categories. Our current version of the model does not provide the direct association, but it could be implemented via the multimodal layer by adding the top-down links from it to the unimodal layers.

The important advantage of our model is the hierarchical representation of the sign components. It guarantees better processing and storing of representations because the sign (multimodal level) is modifiable from both modalities (the sequential “symbolic” auditory level and the parallel “conceptual" visual level). The separate multimodal level provides a platform for the development of subsequent stages of this system (e.g. inference mechanisms).

In our model we have created a system that is able to represent constant features of the environment and identify them with abstract symbols. Meaning is nonarbitrarily represented at the conceptual level (interpretant) that guarantees the correspondence of the internal representational system with the external environment.

Even though our model (especially Model II) was shown to perform quite well (20% error for the most complex scanario), there are ways how to increase its accuracy. For instance, the task of *where* system can be reasonably facilitated by reducing the two objects to radially symmetric blobs of activation, which will eliminate several degrees of variance. Actually, our preliminary simulations confirm this hypothesis. Another thing that we are currently investigating is the scaling of the system.

We would also like to use this representational system in the process of "thinking" (mental manipulation of grounded representations). We intend to compare systems based on the prior symbolic level (the classical grounding approach) with the system based on the symbols grounded to the nonarbitrary concepts via the multimodal layer. The goal is to confirm the advantages of the multimodal representations in the area of symbol grounding.

Acknowledgments

This research was 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. 1/0439/11 (I.F.).

References

Allman, B., Keniston, L., & Meredith, M. (2009). Not just for bimodal neurons anymore: The contribution of unimodal neurons to cortical multisensory processing. *Brain Topography*, *21*, 157-167.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*(04), 577-660.

Damasio, A. (1989). Time-locked multiregional retroactivation: A systems level proposal for the neural substrates of recall and recognition. *Cognition*, *33*, 25-62.

Dorffner, G., Hentze, M., & Thurner, G. (1996). A connectionist model of categorization and grounded word learning. In *Proceedings of the Groningen assembly on language acquisition (GALA'95).*

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, *11*(1), 23-63.

Hammer, B., Micheli, A., Sperduti, A., & Strickert, M. (2004). Recursive self-organizing network models. *Neural Networks*, *17*(8-9), 1061-1085.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 335-346.

Hinton, G., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527-1554.

Kohonen, T. (2001). *Self-organizing maps*. Springer. (3rd edition)

Li, P., Farkaš, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*(8-9), 1345-1362.

Li, P., & McWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments and Computers*.

Martinetz, T., & Schulten, K. (1991). A neural-gas network learns topologies. In *Proceeedings of the int. conference on artificial neural networks* (p. 397-402).

Mel, B., & Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Computation*, *12*, 247-278.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, *57*, 334-366.

Paivio, A. (1986). *Mental representation: A dual coding approach*. Oxford: Oxford University Press.

Peirce, C. (1931). *Collected papers of Charles Sanders Peirce* (C. Hartshorne, Ed.). Harvard University Press.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. Cambridge, MA: MIT Press.

Riga, T., Cangelosi, A., & Greco, A. (2004). Symbol grounding transfer with hybrid self-organizing/supervised neural networks. In *International joint conference on neural networks (IJCNN'04).*

Roy, D. (2005). Grounding words in perception and action: computational insights. *Trends in Cognitive Sciences*, *9*, 389-396.

Stein, B., & Meredith, M. (1993). *Merging of the senses*. Cambridge, MA: MIT Press.

Tiňo, P., Farkaš, I., & Mourik, J. van. (2006). Dynamics and topographic organization in recursive self-organizing map. *Neural Computation*, *18*, 2529–2567.

Ungerleider, L., & Mishkin, M. (1982). Two cortical visual systems. In D. Ingle et al. (Eds.), *Analysis of visual behaviour* (p. 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.

Vavrečka, M. (2009). Model for the grounding of spatial relations (in Czech). In M. Petru (Ed.), *Struny mysli: Kognice 2007* (p. 139-148). Ostrava, Czech Republic: Montanex.

Voegtlin, T. (2002). Recursive self-organizing maps. *Neural Networks*, *15*(8-9), 979-992.

Ziemke, T. (1999). Rethinking grounding. In A. Riegler, M. Peschl, & A. von Stein (Eds.), *Understanding representation in the cognitive sciences* (p. 177-190). New York: Plenum Press.