Richter, M., Lins, J., Schneegans, S., Schöner, G. (2014). A neural dynamic architecture resolves phrases about spatial relations in visual scenes. In S. Wermter et al. (Eds.): Artificial Neural Networks and Machine Learning — ICANN 2014, 24th International Conference on Artificial Neural Network, Lecture Notes in Computer Science 8681, pp. 201–208, 2014, Springer Heidelberg. Best Paper Award.

A neural dynamic architecture resolves phrases about spatial relations in visual scenes

Mathis Richter*, Jonas Lins, Sebastian Schneegans, and Gregor Schöner

Institut für Neuroinformatik, Ruhr-Universität Bochum, 44870 Bochum, Germany
{mathis.richter,jonas.lins,sebastian.schneegans,
gregor.schoener}@ini.rub.de
http://www.ini.rub.de

Abstract. How spatial language, important to both cognitive science and robotics, is mapped to real-world scenes by neural processes is not understood. We present an autonomous neural dynamics that achieves this mapping flexibly. Neural activation fields represent and spatially transform perceptual information. An architecture of dynamic nodes interacts with these perceptual fields to instantiate categorical concepts. Discrete time processing steps emerge from instabilities of the time-continuous neural dynamics and are organized sequentially by these nodes. These steps include the attentional selection of individual objects in a scene, mapping locations to an object-centered reference frame, and evaluating matches to relational spatial terms. The architecture can respond to queries specified by setting the state of discrete nodes. It autonomously generates a response based on visual input about a scene.
Keywords: spatial language; sequence generation; autonomy; neural dynamics; Dynamic Field Theory

1 Introduction

Spatial language helps point people to objects in the world. For instance, in Fig. 1 the kaki fruit can be singled out by asking "What kind of fruit is to the right of the lime?", while referring to its orange color would not have been sufficient. A relational spatial phrase like this consists of a target (the orange kaki), a reference (the green lime), and a relational term (to the right). Humans employ a series of processing steps when interpreting such a phrase [7] that include binding each object to its role, centering the reference frame on the reference object, mapping the relational term onto this frame, and assessing the match of the target object with the spatial term. These processes have been addressed in both psychological (e.g., [3]) and robotic contexts (e.g., [4]). Roboticists tend to use ad-hoc algorithms to organize the perceptual processing, while psychologists have typically invoked concepts of information processing. A neural process account for how flexible spatial language may tie to perception is lacking to date.

* The authors gratefully acknowledge the financial support of the European Union Seventh Framework Programme FP7-ICT-2009-6 under Grant Agreement no. 270247—NeuralDynamics.

Fig. 1. Visual scenario affording the use of spatial language.

The first challenge for such a neural approach is that spatial language involves two types of representation. A relational phrase consists of a set of amodal, discrete symbols, signifying reference object, target object, and the relational term. The referent of such a phrase, say the pair of objects in a visual scene, is provided in a modal, subsymbolic format. Resolving a spatial phrase consists of establishing a coherent mapping between these two types of representations. We address this challenge using Dynamic Field Theory (DFT) [12] in which neural population activity is described by activation fields, defined over metric feature dimensions, that evolve continuously in time through a neural dynamics. In DFT, modal representations are captured as dynamic neural fields, while amodal, categorical representations are modeled by activation nodes, that share the same neural dynamics. The shared dynamics and mutual coupling enable integrating modal and amodal representations.

People use spatial language flexibly, in that they are able to (1) direct their attention to an object guided by a relational phrase, (2) generate a relational phrase to describe a visual scene, and (3) answer questions about spatial relations between objects in a visual scene. These tasks differ in how much information is already provided in the phrase and how much must be extracted from the sensory representation. Achieving this flexibility is a second challenge for a neural account. This requires that the elementary processing steps must be recombined depending on the context. We address this problem by exploiting an analogy with a DFT-based approach to the autonomous generation of behavioral sequences [11, 10]. In that approach, a neural representation of an *intention* activates the neural processes that execute the intention. A *condition of satisfaction* detects predicted changes in activation states that indicate that the processing steps have been successfully completed. A *condition of dissatisfaction* indicates failure to do so. Bifurcations of the neural dynamics create these events and trigger the transition to the next processing step.

We build on an earlier DFT model of spatial language [6] that provided the key processes for resolving spatial phrases, including the attentional selection of target and reference objects and the transformation of target locations into a frame centered on the reference object. The processing sequence was externally controlled in that earlier model, while we will demonstrate the emergence of the discrete steps from continuous neural dynamics here. The resulting neural architecture is able to autonomously resolve relational phrases and answer questions about real visual scenes.

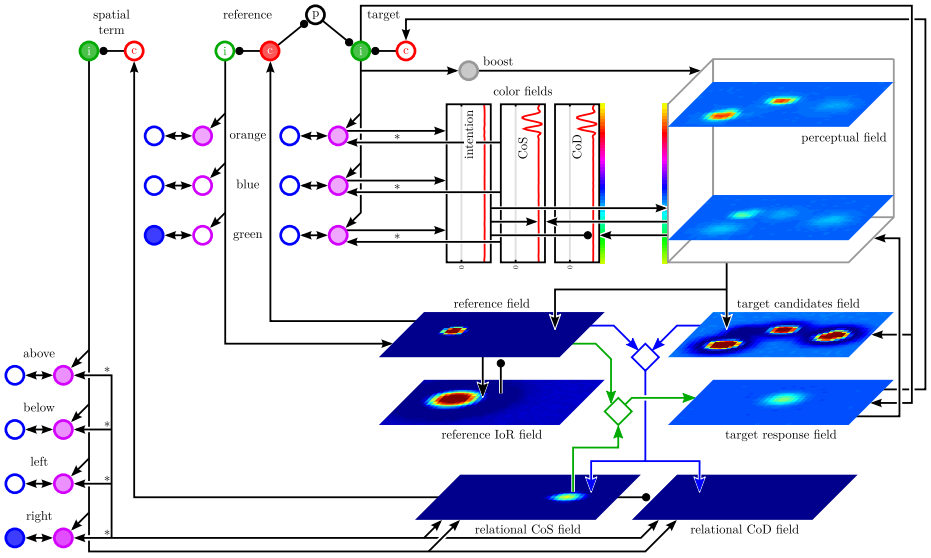

Fig. 2. Overview of the architecture, showing the activation state when answering the question "What is to the right of the green object?" on the scene in Fig. 1. On the right, dynamic fields are shown as color-coded activation patterns (blue for lowest, red for highest activation). On the left, dynamic nodes are denoted as circles with activation levels indicated by fill color opacity. The three-dimensional perceptual field is shown as slices through the activation pattern for the colors orange and green. Excitatory synaptic connections are denoted by arrows, inhibitory connections by lines ending in circles. Arrows marked with stars are patterned connections that encode concepts.

2 Methods

The DFT architecture shown in Fig. 2 can be viewed as one integrated dynamical system, that combines coupled dynamics fields (DFs) supporting perception with coupled dynamic nodes that instantiate concepts and organize sequential processing.

2.1 Dynamic fields and dynamic nodes

DFs can be thought of as a temporally and spatially continuous form of neural networks. Activation fields, $u(x,t)$, over a continuous feature dimension x (e.g., hue or spatial position) evolve over time t according to

$$\tau \dot{u}(x,t) = -u(x,t) + h + S(x,t) + \int f(u(x',t))\, w(x-x')\, dx',$$

where τ is a time constant, $h < 0$ is a resting level, and $S(x,t)$ is external input. Lateral interactions in the field are homogeneous and can be described as a

convolution of the interaction kernel w and the field output $f(u(x,t))$, where f is a sigmoid function with threshold at zero [1]. Local excitatory and surround inhibitory interaction leads to stable localized peaks of activation that are the units of representation. Fields may either support multiple peaks or select a single peak from multiple inputs through competitive lateral interactions. For strong self-excitation, peaks may be sustained when local input is removed, a model of working memory.

Discrete activation nodes governed by the same equation can be viewed as zero-dimensional DFs. The nodes can switch between an 'on' state, stabilized by self-excitation, and a sub-threshold 'off' state. The transitions between different peak configurations in a DF or different states of a node constitute instabilities in the neural dynamics that create the discrete events that are critical for the autonomous organization of sequential processing steps.

2.2 Perceptual system and feature attention

The *perceptual field* (top right in Fig. 2) represents the current visual scene as a distribution of salient colors. Activation along two spatial and one color (hue) dimension is driven by input from a camera image. The perceptual field is coupled to three color fields (top middle in Fig. 2): In the *color intention field*, input from color nodes (explained below) induces a peak at the currently relevant hue value. That peak projects to the perceptual field where it enables peaks to form at locations in the scene that match this color. At the same time, it pre-activates the corresponding color value in the *color condition-of-satisfaction (CoS) field* and suppresses that value in the *color condition-of-dissatisfaction (CoD) field*. These two fields receive excitatory input also from the perceptual field. A match of the two inputs leads to a peak in the CoS field, a mismatch to a peak in the CoD field.

2.3 Representing spatial relations

The *reference field* holds the location of a single reference object, the *target candidates field* holds the locations of one or more potential targets. Both fields are defined over the two-dimensional space of the camera image and receive input from the perceptual field. The reference field drives the *reference inhibition-of-return (IoR) field*, in which all locations that have previously been selected are stored, sending back inhibition to the reference field.

The outputs of the target and reference field are combined in a coordinate transformation (blue diamond in Fig. 2), to determine the relative position of the target candidates with respect to the selected reference object. Neurally, this can be realized using a four-dimensional DF [6], but for performance reasons we use a convolution here. The result is fed into a pair of *relational fields*, organized as one CoS and one CoD field. In these fields, the match of the transformed locations with the relational term is evaluated based on a spatial template for that term that is provided to the two fields as additional input. The input is excitatory for the relational CoS field to detect match, inhibitory for the relational CoD field, to

detect mismatch. An inhibitory projection from the CoS to the CoD field ensures that only the CoS field forms a peak if both matching and mismatching target candidates are present. From the relational CoS field, a reverse transformation back into image coordinates (green diamond in Fig. 2) projects to the *target response field* using the stored location of the reference object.

2.4 Processing spatial phrases

Dynamic nodes represent concepts, here colors and spatial terms, in an amodal form. These concepts can fill different roles in a relational spatial expression, namely target object, reference object, and spatial relation. To realize role-filler binding, we employ conjunctive coding, such that a copy of each concept representation exists for every role it can take (e.g., 'target: red' and 'reference: red'). Each concept-role conjunction is represented by one pair of nodes, a *production* and a *memory* node (purple and blue circles in Fig. 2). Reciprocal patterned connections between the production nodes and the perceptual feature spaces of different fields instantiate concepts. The connection weights encoding these concepts are handcoded here but should eventually be learned using neurally realistic methods. When the production nodes are activated, they induce specific activation patterns in the color intention field or the relational CoS and CoD fields. The memory nodes, coupled bidirectionally to their corresponding production nodes, remain active by self-excitation to encode a relational phrase while it is processed.

Production nodes are excited by their memory nodes, but become fully active only when receiving input from an additional set of nodes that control the sequential processing of a phrase. These nodes are organized in pairs of an *intention* and a *CoS* node (green and red circles in Fig. 2), with one such pair for each of the three roles in a relational phrase. The intention node for a specific role projects to all its production nodes and provides homogeneous input to fields associated with that role (e.g., the reference field). The corresponding CoS node in turn receives input from these fields and from the intention node. It is turned on when activation peaks form in the fields while the matching intention node is active. On activation, a CoS node suppresses its intention node. Which intention nodes can be active simultaneously is controlled by *precondition constraints* (black circles) which suppress the intention node of one role until the CoS node of another role becomes active.

3 Results

We illustrate the core functionality of the architecture by showing the dynamic processes that unfold when answering the question "What is to the right of the green object?" about the scene shown in Fig. 1. To simplify the visual object recognition, the scene consists of uniformly colored objects on a white background. We implemented and simulated the architecture in real time using *cedar* [8], an open source library for DFT.

The question is encoded by activating specific memory nodes, here 'reference: green' and 'spatial term: right of'. Since the question asks for information about the target object, we enable the architecture to provide homogeneous input to the perceptual field and the target production nodes, when the intention node for the target object becomes active. We will now follow along Fig. 3, which shows how the activation patterns of the architecture evolve over time.

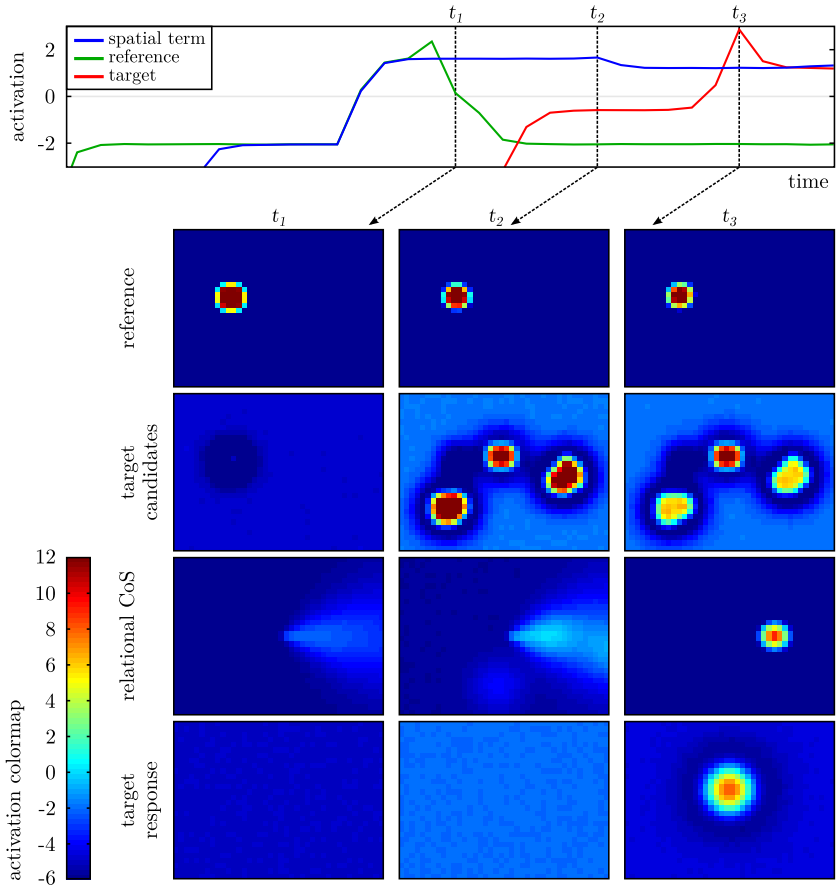

Fig. 3. Evolution of activation patterns for answering a question about the scene in Fig. 1. On top we show continuous activation time courses for the production nodes 'reference: green', 'target: orange', and 'spatial term: right of'. On the bottom, we show activation snapshots at three time steps for relevant fields using a color code (see colorbar; the threshold is at zero).

We initiate the processing by giving a task input that activates all intention nodes as well as the precondition node. The architecture works autonomously

from this point on. The intention nodes for reference and spatial term become active, while the one for target is inhibited by the precondition node. The reference intention node boosts the reference field while the spatial term intention node boosts the relational CoS and CoD fields. Additionally, both intention nodes homogeneously boost their associated production nodes, activating those nodes that receive input from their respective memory nodes (i.e., 'reference: green' and 'spatial term: right of'). This can be seen by the green and blue curves rising above the treshold in Fig. 3. The production node for reference induces a peak at the color green in the color intention field. This brings green objects into the attentional foreground, forming a peak in the perceptual field at the location of the lime. That position is projected to the reference field, which is visible at time t_1 in Fig. 3. At the same time, the production node for the spatial term 'right of' projects activation into the relational CoS and CoD fields.

Having represented the position of the green lime in the reference field, the reference CoS node is activated, which deactivates the precondition node. This in turn enables the target intention node to become active, boosting the perceptual field and bringing all available objects into the target candidates field (in this field, the reference object is suppressed; see snapshot at t_2 in Fig. 3). The snapshot of the relational CoS field at the same time shows slightly elevated activation levels at the spatially transformed locations for the target candidates. One of the target candidates—the orange kaki to the right of the green lime—overlaps with the spatial template. This overlap leads to a peak in the relational CoS field, which is projected back to the image space in the target response field and from there into the perceptual field, highlighting the final target object (see snapshots of the target response and target candidates field at t_3 in Fig. 3). As a response to the question "What is to the right of the green object?", the architecture activates the production and memory node 'target: orange'.

4 Discussion

We have shown how autonomous processing of relational phrases can be achieved in a neural architecture. In particular, we have demonstrated how the architecture can identify target objects in a visual scene whose location is only indirectly specified via a reference object and a relational term. Due to lack of space we could only hint at further capabilities of the architecture. They emerge from its structure in a way analogous to the detailed example shown here. The architecture is, for instance, able to select a relational term as a response to a question and ground a complete relational phrase in perceptual representations.

We can extend the architecture to include other feature dimensions beyond color (e.g., shape), by adding one set of feature fields (i.e., perceptual, intention, CoS, and CoD fields) for each dimension.[1]

Combining dynamic nodes and neural fields, the architecture connects to theoretical strands that stress the modal nature of mental processes [2] as well as

[1] Note, however, that this does not address the neurally plausible extraction of features from visual scenes, which is beyond the scope of the architecture.

to traditional, amodal views on cognition [9]. The architecture has the potential to generalize to non-spatial and abstract problems. This is based on evidence [5] suggesting a pervasive role of spatial representation in reasoning. With the current architecture, we have provided a first step toward connecting these ideas and thereby grounding mechanisms of higher cognition in neural reality.

References

1. Amari, S.i.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics 27(2), 77–87 (1977)
2. Barsalou, L.W.: Perceptual symbol systems. Behavioral and Brain Sciences 22(04), 577–660 (1999)
3. Carlson, L.A., Logan, G.D.: Attention and spatial language. In: Itti, L., Rees, G., Tsotsos, J.K. (eds.) Neurobiology of Attention, chap. 54, pp. 330–336. Elsevier Academic Press (2005)
4. Guadarrama, S., Riano, L., Golland, D., Gohring, D., Jia, Y., Klein, D., Abbeel, P., Darrell, T.: Grounding spatial relations for human-robot interaction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2013)
5. Knauff, M.: Space to reason: A spatial theory of human thought. MIT Press, Cambridge, MA (2013)
6. Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J.P., Schöner, G.: A neurobehavioral model of flexible spatial language behaviors. Journal of Experimental Psychology. Learning, Memory, and Cognition 38(6) (2012)
7. Logan, G.D., Sadler, D.D.: A computational analysis of the apprehension of spatial relations. In: Bloom, P., Peterson, M., Nadel, L., Garrett, M. (eds.) Language and Space, chap. 13, pp. 493–529. MIT Press, Cambridge, MA (1996)
8. Lomp, O., Zibner, S.K.U., Richter, M., Rano, I., Schöner, G.: A software framework for cognition, embodiment, dynamics, and autonomy in robotics: cedar. In: Mladenov, V. (ed.) ICANN 2013, Lecture Notes in Computer Science 8131. pp. 475–482. No. 270247, Springer, Heidelberg (2013)
9. Pylyshyn, Z.W.: The imagery debate: Analogue media versus tacit knowledge. Psychological Review 88, 16–45 (1981)
10. Richter, M., Sandamirskaya, Y., Schöner, G.: A robotic architecture for action selection and behavioral organization inspired by human cognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 2457–2464 (2012)
11. Sandamirskaya, Y., Schöner, G.: An embodied account of serial order: How instabilities drive sequence generation. Neural Networks 23(10), 1164–1179 (2010)
12. Schneegans, S., Schöner, G.: Dynamic field theory as a framework for understanding embodied cognition. In: Calvo, P., Gomila, T. (eds.) Handbook of Cognitive Science: An Embodied Approach, pp. 241–271. Perspectives on Cognitive Science, Elsevier (2008)