Behaviorally Flexible Spatial Communication: Robotic Demonstrations of a Neurodynamic Framework

John Lipinski, Yulia Sandamirskaya, and Gregor Schöner

Institut für Neuroinformatik
Ruhr-Universität Bochum
Bochum, Germany
Tel.: +49-234-3224201
Fax: +49-234-3214209
2johnlipinski@gmail.com

Abstract. The ease of everyday human-human spatial communication suggests that human spatial cognitive processes may provide a model for developing artificial spatial communication systems that fluidly interact with human users. To this end, we develop a neurodynamic model of human spatial language that combines linguistic spatial and color terms with neurally-grounded scene representations. Tests of this model implemented on a robotic platform continuously linked to real-world camera input support its viability as a theoretical framework for flexible, autonomously generated spatial language behaviors in artificial agents grounded in human cognitive processes.

1 Introduction

A central goal of artificial intelligence research is to develop systems that flexibly interact with human users in human-centered environments. Effective spatial language is a necessary component of efficient human-robot interaction for one simple reason: spatial language provides a natural means by which humans communicate about and coordinate behaviors within a shared workspace. The ease and fluidity of human-human spatial communication suggests that human spatial cognitive processes themselves may provide a useful model for artificial spatial communication systems. The shear complexity and richness of human spatial cognition [4], however, is an obvious challenge to this approach. In developing such a human-based artificial system it is therefore useful to focus on a constrained but still fundamental set of characteristics underlying human spatial communication. To this end, the current work focuses on autonomy and behavioral flexibility in spatial language processing. Limiting our focus in this way provides for a theoretically manageable research agenda that still incorporates core features of human spatial cognition.

1.1 Autonomy and Flexibility

In the context of cognition, autonomy refers to the unfolding of cognitive processes continuously in time according to past and present sensory and behavioral states [7]. In agent-based systems such as robots, autonomy further implies that agent behaviors are structured according to sensory information that the agent itself acquires [9]. These aspects of autonomy draw attention to two core elements of human spatial language. First, natural spatial language comprehension and production depends on the smooth, continuous integration of both visual and linguistic information over

time, not from fixed input-output relations or rigorously constrained linguistic and visual inputs. The inherent variability of speaker timing (e.g. slow versus rapid speech), word order, and visual context demands a system that can continuously integrate this information in a manner permitting contextually adaptive behavior. Second, because spatial language often changes behavior (e.g. guiding ones attention eye-movement, or action to a specific location), behavioral changes needs to be fluidly, naturally coordinated with continuous linguistic input.

Flexibility, the second key characteristic, is embedded in the principle of autonomy. Flexibility is important for spatial language because the same spatial language system must support both the production and comprehension across a broad array of tasks, including the extraction of object features at a described location, the selection of a spatial term that describes an object's location, and the combination of the spatial and non-spatial features in processing a spatial description (i.e. "The *red* apple to the *right* of the glass"). These behaviors must also be flexibly deployed across a limitless array of visual scenes.

1.2 Dynamic Field Theory

To facilitate human-robot interaction, it is critical to identify an implementable theoretical framework that can capture these characteristics. The Dynamic Field Theory [3,8] provides this framework. Dynamical Field Theory (DFT) is a neural-dynamic approach to human cognition in which cognitive states are represented as distributions of neural activity defined over metric dimensions. These dimensions may represent perceptual features (e.g., retinal location, color, orientation), movement parameters (e.g. heading direction) or more abstract parameters (e.g. visual attributes of objects like shape or size). These continuous metric spaces represent the space of possible percepts, actions, or objects and scenes.

Spatially continuous neural networks (neural fields) are at the heart of the DFT and were originally introduced as approximate descriptions of cortical and thalamic neuroanatomy. Neural fields are recurrent neural networks, whose temporal evolution is described by iteration equations. In continuous form, these take the form of dynamical systems. The mathematics of dynamical neural fields was first analyzed by Amari [1] and much modeling has since built on the original Amari framework. Importantly, recent modeling and empirical work in spatial cognition (for review see [10]) shows how the DFT also captures core characteristics of human spatial cognition. Collecitively, this suggests that the DFT may facilitate the development of fluid human-robot spatial communication.

To rigorously test this claim we present a new DFT-based spatial language model and implement it on a robotic platform continuously linked to real-world visual images. In keeping with our goal of a "human style"cognitive approach, our model extracts the categorical, cognitive information from the low-level sensory input through the system dynamics. Our demonstrations specifically combine visual space, spatial language, and color.

2 Modeling neurons and dynamical neural fields

2.1 Dynamical fields

The dynamical neural fields are mathematical models first used to describe cortical and subcortical neural activation dynamics [1]. The dynamic field equation Eq. (1) is a differential equation describing the evolution of activation u defined over a neural variable(s) $\mathbf{x}$. These neural variables

represent continuous perceptual (e.g. color) or behavioral (e.g. reaching amplitude) dimensions of interest that can be naturally defined along a continuous metric.

$$\tau \dot{u}(\mathbf{x},t) = -u(\mathbf{x},t) + h + \int f(u(\mathbf{x}',t))\omega(\Delta\mathbf{x})d\mathbf{x}' + I(\mathbf{x},t) \quad (1)$$

Here, $h < 0$ is the resting level of the field; the sigmoid non-linearity $f(u) = 1/(1 + e^{-\beta u})$ determines the field's output at suprathreshold cites with $f(u) > 0$. The field is quiescent at subthreshold cites with $f(u) < 0$. The homogeneous interaction kernel $\omega(\Delta x) = c_{exc}e^{\frac{-(\Delta x)^2}{2\sigma^2}} - c_{inh}$ depends only on the distance between the interacting cites $\Delta x = \mathbf{x} - \mathbf{x}'$. This interaction kernel is a Bell-shaped, local excitation/lateral inhibition function. The short-range excitation is of amplitude c_{exc} and spread σ. The long-range inhibition is of amplitude c_{inh}. $I(\mathbf{x},t)$ is the summed external input to the field; τ is the time constant.

If a localized input activates the neural field at a certain location, the interaction pattern ω stabilizes a localized "peak", or "bump"solution of the field's dynamics. These activation peaks represent the particular value of the neural variable coded by the field and thus provide the representational units in the DFT. In our model, all entities having "field"in their name evolve according to Eq. (1), where $\mathbf{x}$ is a vector representing the two-dimensional visual space in Cartesian coordinates. The links between the fields are realized via the input term $I(\mathbf{x},t)$, where only cites with $f(u) > 0$ propagate activation to other fields or neurons.

2.2 Discrete neurons

The discrete (localist) neurons in the model representing linguistic terms can be flexibly used for either user input or response output and evolve according to the dynamic equation (2).

$$\tau_d \dot{d}(t) = -d(t) + h_d + f(d(t)) + I(t). \quad (2)$$

Here, d is the activity level of a neuron; the sigmoidal non-linearity term $f(d)$ shapes the self-excitatory connection for each discrete neuron and provides for self-stabilizing activation. The resting level is defined by h_d. The $I(t)$ term represents the sum of all external inputs into the given neuron. This summed input is determined by the input coming from the connected neural field, the user interface specifying the language input, and the competitive, inhibitory inputs from the other discrete neurons defined for that same feature group (color or space); τ is the time constant of the dynamics.

3 The spatial language framework

In this section we outline the overall structure (see Fig. 1) and functionality of the integrative model. The color-space fields (Fig. 1A) are an array of several dynamical fields representing the visual scene. Each of the fields is sensitive to a hue range which corresponds to a basic color. The resolution of color was low in the presented examples because only a few colors were needed to represent the used objects. In principle, the color (hue) is a continuous variable and can be resolved more finely. The stack of color-space fields is therefore a three-dimensional dynamic field that represents colors and locations on the sensor surface. The camera provides visual input to the color-space field, which is below the activation threshold before the task is defined. The field is thus quiescent to this point.

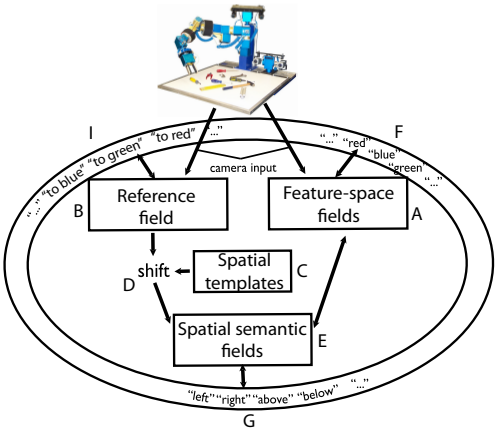


Fig. 1. Overview of the architecture

Once the language input specifies the *color* of the object, however, the resting levels of all cites of the corresponding color-space field are raised homogeneously. Because the color-space fields receive localized camera input, this uniform activation increase is summed with that input to enable the development of an instability and, ultimately, the formation of a single-peak solution. This peak is centered over the position of the object with that specified color.The *spatial* language input also influences the color-space fields' dynamics through the aligned spatial semantic fields (see below).

The reference field (Fig. 1B) is a spatially-tuned dynamic field which also receives visual input. When the user specifies the reference object color, the corresponding "reference-color" neuron becomes active and specifies the color in the camera image that provides input into the reference field. A peak of activation in the reference field specifies the location of the reference object. The reference field continuously tracks the position of the reference object. It also filters out irrelevant inputs and camera noise and thus stabilizes the reference object representation. Having a stable, but updatable reference object representation allows the spatial semantics to be continuously aligned with the visual scene.

The spatial semantic templates (Fig. 1C) are represented as a set of synaptic weights that connect spatial terms to an abstract, "retinotopic"space. The particular functions defining "left", "right", "below", and "above"here were two-dimensional Gaussians in polar coordinates and are based on a neurally-inspired approach to English spatial semantic representation [5]. When viewed in Cartesian coordinates, they take on a tear-drop shape and correspond to prototypical neural representations of spatial relations in animals.

The shift mechanism (Fig. 1D) aligns these retinotopically defined spatial semantics with the current task space. The shift is done by convolving the "egocentric"weight matrices with the outcome of the reference field. Because the single reference object is represented as a localized activation peak in the reference field, the convolution simply centers the semantics over the reference object. The spatial terms thus become defined relative to the specified reference object location (for related method see [6]).

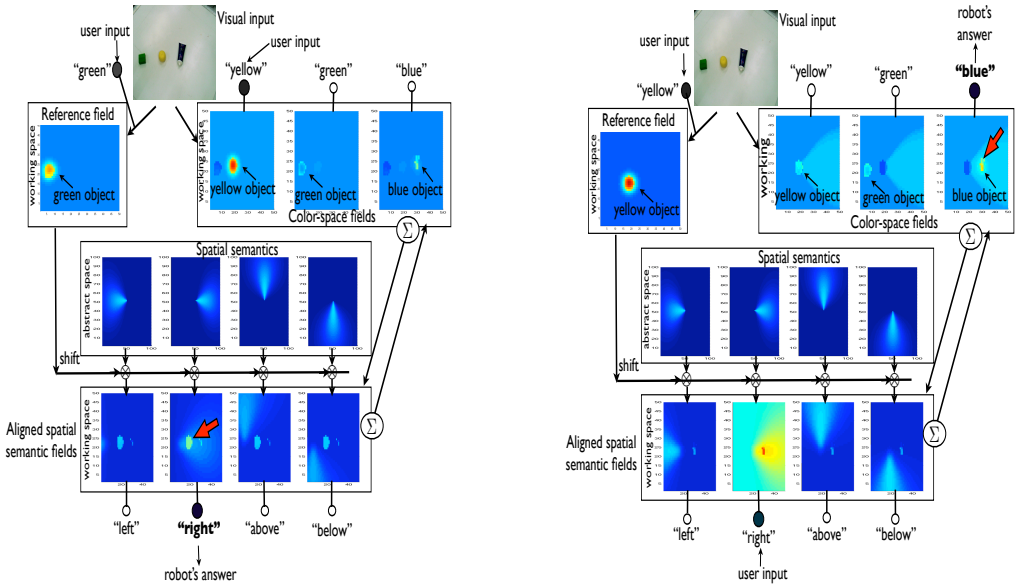
The aligned spatial semantic fields (Fig. 1E) are arrays of dynamical neurons with weak lateral interaction. They receive input from the spatial alignment or "shift"mechanism which maps the

spatial semantics onto the current scene by "shifting"the semantic representation of the spatial terms to the reference object position. The aligned spatial semantic fields integrate the spatial semantic input with the summed outcome of the color-space fields and interact reciprocally with the spatial-term nodes. Thus, a positive activation in an aligned spatial semantic field increases the activation of the associated spatial term node and vice versa.

4 Demonstrations

In the presented scenarios, everyday objects (e.g. a red plastic apple, a blue tube of sunscreen) were placed in front of the robot. The visual input was formed from the camera image and sent to the reference and color-space fields. The color-space field input was formed by extracting hue value ("color") for each pixel in the image and assigning that pixel's intensity value to the corresponding location in the matching color-space field. The input for the reference field was formed in an analogous fashion according to the user-specified reference object color. When the objects are present in the camera image, the reference and color-space fields receive localized inputs, corresponding to the three objects in view (marked with arrows, see Fig. 2(a) and Fig. 2(b)). This was the state of the system before the particular task was set.

4.1 Demonstration 1: Describing "Where"



(a) Demonstration 1 activations just before answering "Where".

(b) Demonstration 2 activations just before answering "Which".

Fig. 2. The basic behaviors of the architecture

Demonstration 1 asks "Where is the yellow object relative to the green one?"To respond correctly, the robot must select "Right". Fig. 2(a) shows the neural field activation just before the answer is given. The task input first activates two discrete neurons, one representing "green"for the user-specified reference object color and the other "yellow"for the user-specified object color (see user inputs, top Fig. 2(a)). The reference object specification "green"leads to the propagation of the green camera input into the reference field, creating an activation bump in the reference field at the location of the green item (see Reference Field, Fig. 2(a)). The specification of the target color "yellow"increases the activation for the yellow node linked to the yellow color-space field, which raises the resting level of the associated yellow color-space field. This uniform activation boost coupled with the camera input from the yellow object induces an activation peak in the field (see "yellow"color-space field, Fig. 2(a)).

This localized target object activation is then transfered to the aligned semantic fields. In addition to receiving this target-specific input, the aligned semantic fields also receive input from spatial term semantic units. Critically, these semantic profiles are shifted to align with the reference object position. In the current case, the yellow target object activation therefore overlaps with the aligned "right"semantic field (see red arrow in the "right"aligned spatial semantic field, Fig. 2(a)). This overlap ultimately drives the activation and selection of the "right"node.

4.2 Demonstration 2: Describing "Which"

Demonstration 2 asks "Which object is to the right of the yellow one?". To respond correctly, the robot must select "Blue". As indicated in Fig. 2(b), the task input first activates two discrete neurons, one representing the reference object color "yellow"and the other representing "right".

The reference object specification "yellow"creates an activation bump in the reference field location matching that of the yellow item (see reference field, Fig. 2(b)). The specification of "right", in its turn, increases the activation for that spatial-term node, creating a homogeneous activation boost to the "right"semantic field. This activation boost creates a positive activation in the field to the right of the yellow reference object (see "right"aligned spatial semantic field, Fig. 2(b)). This spatially-specific activation is then input into the color-space fields and subsequently raises activation at all those color-space field locations to the right of the reference object (see lighter blue color-space field regions, Fig. 2(b)). This region overlaps with the localized input of the blue object in the "blue"color-space field and an activation peak develops in that field (see red arrow in the "blue"object color-space field, Fig. 2(b)). This increases the activation of the associated "blue"color-term node, triggering selection of the correct answer, "blue".

4.3 Demonstration 3: Dynamically driven sensor movement

The previous demonstrations highlight our architecture's flexibility and robustness in the face of varying scenes and linguistic input. Movement presents an additional set of behavioral challenges. First, movements (gaze, orienting, reaching, etc) can be driven by internal cognitive states shaped by spatial language [2]. Linking internal decision dynamics to bodily movement is thus an important benchmark for capturing key aspects of natural spatial communication. Second, when that movement involves the sensor providing the spatial information (e.g. eyes) the changing visual input can disrupt the dynamics supporting the peaks driving cognitive behaviors. Robustly adaptive behavior in the context of such movement is thus an additional test the dynamic approach to spatial communication.

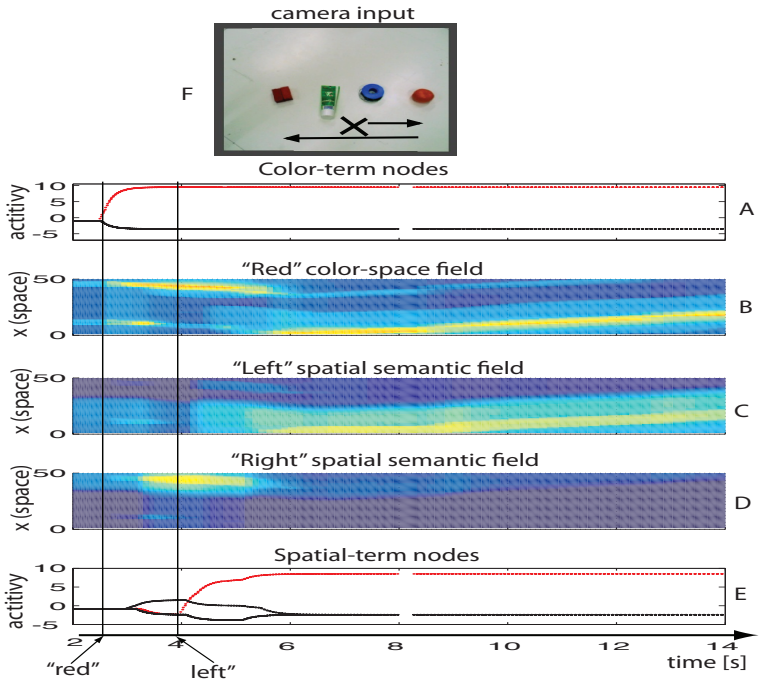


Fig. 3. Demonstration 3 time course. The horizontal axis in all panels represents time. The vertical axis in A and E represents activation, on panels B-D – the projected activation onto horizontal axis of the image. **Panels A-E (Demo. 3, "red"then "left")**: The "red"color term input activates node (red line, Panel A), creating peak in "red"color-space field at red plastic apple location (first orange ridge, Panel B); the "right"spatial semantic field (Panel D) and "right"node (black line, Panel E) also become active. The camera then moves rightward (see esp. bounded region, Panel B). When "left"spatial term input is given (Panel E), "left"node becomes active (red line, Panel E), increasing "left"semantic activity in region of leftmost red object (orange ridge, Panel B). The color-space field regions left of referent then become more active. In Panel B, first activation peak is eliminated and new peak emerges, driving camera movement towards the correct object.

Demonstration 3 addresses these challenges with the addition of dynamic motor control module that drives sensor (camera) movement. We present the sentence "The red one to the left of the blue" in the context of two red objects. The robot's task to establish a peak at the correct object location, shifting the camera accordingly.

Fig. 3 presents the time course of the task (blue reference object specified previously) along with the summary camera movements (see Fig. 3). We present the "red" color term first which uniformly boosts the "red" color-space field and creates an activation peak for the slightly larger, but *incorrect* red object (red apple) location on the right (see yellow activity in Fig. 3B). The camera then begins to center that location by shifting to the right. This in turn leads to the smearing and shift of the activation profiles across all the fields in Fig. 3(see especially Fig. 3B). Nevertheless, note that this peak is stably maintained across the camera movement, thus tracking the location of the red object. When we later specify the "left" spatial relation(Fig. 3E), however, this initial peak is extinguished

and a peak at the fully described correct location arises instead (see later portion of Fig. 3B). This new peak then shifts the camera dynamics and the camera moves in the opposite direction to center the correct object (see shifting activity profiles in Fig. 3B-D). This result demonstrates our model's ability to dynamically drive motor behaviors based on emergent, dynamic decision processes within a neurally-grounded spatial communication system.

5 Conclusion

The development of effecient, fluid human-robot communication systems is a central aim of artificial intelligence research. Given the ease and flexibility of human-human spatial communication, developing artificial systems based on human spatial cognitive processes offers one means of reaching this aim. To this end, the present work adopted a mathematically specified, systems-level neural dynamic perspective with strong links to human spatial cognition to develop an implementable spatial language framework. We then tested this model in three demonstrations with a robotics platform linked to a real-time camera image of a shared workspace. Although not comprehensive, the behavioral flexibility arising from our autonomous neurodynamic model across the three demonstrations suggests that systems-level neural dynamic theories like the DFT can aid the development of effective artificial agent spatial communication systems.

References

1. S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.
2. C.G. Chambers, M.K. Tanenhaus, K.M. Eberhard, H. Filip, and G.N. Carlson. Circumscribing referential domains during real-time language comprehension. *journal of Memory and Language*, 47:30–49, 2002.
3. W. Erlhagen and G. Schöner. Dynamic field theory of movement preparation. *Psychological Review*, 109:545–572, 2002.
4. S.C. Levinson. *Space in language and cognition: Explorations in cognitive diversity*. Cambridge University Press, Cambridge, 2003.
5. J. O'Keefe. Vector grammar, places, and the functional role of the spatial prepositions in english. In E. van der Zee and J. Slack, editors, *Representing direction in language and space*. Oxford University Press., Oxford, 2003.
6. E. Salinas. Coordinate transformations in the visual system. *Advances in Neural Population Coding*, 130:175–190, 2001.
7. Y. Sandamirskaya and G. Schöner. Dynamic field theory and embodied communication. In I. Wachsmuth and G. Knoblich, editors, *Modeling communication with robots and virtual humans*, Lecture Notes in Artificial Intelligence, Vol. 4930. Springer, 2006.
8. G. Schöner. Dynamical systems approaches to cognition. In R. Sun, editor, *The Cambridge handbook of computational psychology*, pages 101–126. Cambridge University Press, 2008.
9. G Schöner, M Dose, and C Engels. Dynamics of behavior: Theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems*, 16:213–245, 1995.
10. J. P. Spencer, V. S. Simmering, A. R. Schutte, and G. Schöner. What does theoretical neuroscience have to offer the study of behavioral development? insights from a dynamic field theory of spatial cognition. In J. M. Plumert and J. P. Spencer, editors, *Emerging landscapes of mind: Mapping the nature of change in spatial cognition*, pages 320–361. Oxford University Press, Oxford, 2007.