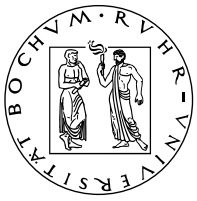
A neural dynamic model for the perceptual grounding of spatial and movement relations

Mathis Marius Richter



RUHR-UNIVERSITÄT BOCHUM
Institut für Neuroinformatik

Dissertation zur Erlangung des Grades eines Doktor-Ingenieurs der Fakultät für Elektrotechnik und Informationstechnik an der Ruhr-Universität Bochum

---

A neural dynamic model for the perceptual grounding of spatial and movement relations

Dissertation von:
Mathis Marius Richter
geboren in Bochum

2017

Colophon

This document was typeset using the XeTeX typesetting system created by the Non-Roman Script Initiative and the memoir class[1] created by Peter Wilson. It is based on the PhD thesis template[2] by Frederico Maggi but is also inspired by the work of Edward Tufte and the PhD thesis[3] of Aaron Turon. The body text is set 12pt with Adobe Caslon Pro. The bibliography was processed by Biblatex.[4]

A neural dynamic model for the perceptual grounding of spatial and movement relations
December 13, 2017, Mathis Marius Richter

[1] http://www.ctan.org/pkg/memoir

[2] https://www.latextemplates.com/template/maggi-memoir-thesis

[3] https://people.mpi-sws.org/~turon/turon-thesis.pdf

[4] http://www.ctan.org/pkg/biblatex

For my parents

Abstract

A fundamental aspect of human intelligence is our ability to express and understand relations. For example, we can easily find an object described by the phrase "the red car to the left of the silver car". Just as easily, we can describe a scene ourselves and say something like "the red car is moving toward the intersection". The processes in our brain that enable us to interpret such spatial relations and movement relations in the world may also facilitate the processing of a much richer body of abstract relations that show up most prominently in language. Understanding these processes is thus fundamental to understanding the link between cognition and language.

To capture how spatial and movement relations may be resolved, this thesis introduces a neural process model based on dynamic field theory (DFT), a mathematical and conceptual framework for modeling cognitive processes. The model receives camera input showing a white table with colored balls on it as well as input that represents a relational phrase, similar to the examples above. The task of the model is to bring the described object into the attentional foreground—to *ground* the phrase. In another type of task, it must generate a phrase itself, describing the visual scene.

Solving these tasks requires that the model is able to map discrete concepts in the phrase to continuous feature dimensions, sequentially guide attentional processes to search the scene for multiple objects, put them in working memory, adjust the reference frame of their representation, and evaluate their fit with representations of spatial relations and movement relations. A particular challenge and key aspect of this thesis is to have the model organize its own processes and perform the above tasks without human interference. Establishing the model based on neural principles additionally requires solving fundamental neural problems, such as the neural pointer problem, the binding problem, and generating discrete processing steps from processes that evolve in continuous time.

The model solves all of these problems and innovates over previous work by capturing both grounding and description tasks, spatial and movement relations, and a flexible, hierarchical organization of its processes. This is demonstrated in 104 qualitatively different tests that vary the task, the configuration of objects, and, in particular, how well the given phrase matches the scene. The model is able to correctly ground the given phrase, or generate a phrase, in all cases where this is possible.

By demonstrating how spatial and movement relations may be captured by a neural process model, the thesis brings DFT one step closer toward a comprehensive neural theory of cognition.

Short contents

List of Figures **xv**

List of Tables **xvi**

1 Introduction **1**

2 Background **7**
2.1 Grounding of language 7
2.2 Dynamic field theory 18

3 Model **39**
3.1 Perception 43
3.2 Attention 45
3.3 Spatial transformations 52
3.4 Concepts 63
3.5 Process organization 71

4 Results **83**
4.1 Grounding tasks 84
4.2 Description tasks 121

5 Discussion **135**
5.1 Core component processes 135
5.2 Specific contributions 146
5.3 Limitations 154
5.4 Further research 156

6 Conclusion **159**

Appendices **161**
A Implementation details 163
B Process organization system: equations 166
C Video data set 195

Bibliography **213**

Acronyms **221**

Glossary **223**

Detailed contents

List of Figures xv

List of Tables xvi

1 Introduction 1

2 Background 7
2.1 Grounding of language 7
2.1.1 Terminology 7
2.1.2 What is grounding? 8
2.1.3 Why is grounding necessary? 9
2.1.4 Embodied cognition 11
2.1.5 Embodied computational models 12
2.1.6 Process organization 16
2.1.7 Toward a neural theory of embodied cognition 18
2.2 Dynamic field theory 18
2.2.1 Principles of dynamic field theory 19
2.2.2 Dynamic neural fields 21
2.2.3 Architectures 25
2.2.4 Motion perception 27
2.2.5 Steerable neural mappings 29
2.2.6 Spatial language model 31
2.2.7 Behavioral organization 34
2.2.8 Numerical implementation 37

3 Model 39
3.1 Perception 43
3.1.1 Color perception 43
3.1.2 Motion perception 44
3.2 Attention 45
3.2.1 Feature attention 47
3.2.2 Spatial attention 49
3.3 Spatial transformations 52
3.3.1 Target and reference 53
3.3.2 Target inhibition-of-return 55

3.3.3 Relative position 56
3.3.4 Rotation 57
3.3.5 Inverse transformations 62
3.4 Concepts 63
3.4.1 Color concepts 65
3.4.2 Motion direction concepts 67
3.4.3 Spatial relation concepts 68
3.5 Process organization 71
3.5.1 Processes 71
3.5.2 Sequences of processes 73
3.5.3 Heterarchy of processes 73
3.5.4 Description of all processes 75
3.5.5 Sequentiality 80
3.5.6 Organizational structures in the fields . . . 82

4 Results **83**
4.1 Grounding tasks 84
4.1.1 Single features 85
4.1.2 Feature conjunctions 96
4.1.3 Relations between objects 105
4.2 Description tasks 121
4.2.1 Single objects 122
4.2.2 Relations between objects 125

5 Discussion **135**
5.1 Core component processes 135
5.1.1 Language processing 136
5.1.2 Concept grounding 137
5.1.3 Role-filler binding 139
5.1.4 Attention 140
5.1.5 Working memory representations 142
5.1.6 Reference frame transformation 142
5.1.7 Matching relations 144
5.1.8 Process organization 145
5.2 Specific contributions 146
5.2.1 Grounding and describing 147
5.2.2 Movement relations 149
5.2.3 Process organization 150
5.2.4 Hypothesis testing 152
5.2.5 Matching of relational templates 152
5.2.6 Roles and role-filler binding 153
5.2.7 Extensive qualitative testing 153
5.3 Limitations 154
5.4 Further research 156

6 Conclusion **159**

Appendices **161**
A Implementation details 163
A.1 Model parameters 163
A.2 Visual preprocessing 165
A.3 Software 165
B Process organization system: equations 166
B.1 Ground object process 166
B.2 Ground relation process 167
B.3 Describe process 168
B.4 Target process 169
B.5 Reference process 171
B.6 Spatial relation process 172
B.7 Clean process 174
B.8 Reset process 175
B.9 Perceptual boost process 176
B.10 Spatial attention process 178
B.11 Feature process 179
B.12 Target IOR process 181
B.13 Target memory node process 182
B.14 Target motion field process 183
B.15 Target field process 185
B.16 Reference memory node process 186
B.17 Reference field process 187
B.18 Spatial memory node process 188
B.19 Spatial relational field process 189
B.20 Target memory node color process 190
B.21 Target memory node motion process 191
B.22 Sequentiality 192
C Video data set 195

Bibliography **213**

Acronyms **221**

Glossary **223**

List of Figures

2.1 Terminology: symbol, concept, object 8
2.2 One-dimensional dynamic neural field 21
2.3 Selective interaction kernel 22
2.4 Multi-peak interaction kernel 22
2.5 Sigmoid function 23
2.6 Two-dimensional dynamic neural field 24
2.7 Dynamic neural node 24
2.8 Neural oscillator 25
2.9 Expansion coupling (1D to 2D) 26
2.10 Transient detector 28
2.11 Steerable neural mapping 29
2.12 Spatial language model 31
2.13 Elementary behavior 34
2.14 Precondition constraint 37
2.15 Suppression constraint 37
2.16 Discrete time Euler approximation 38

3.1 Overview of the model 40
3.2 Color perception 44
3.3 Motion perception 45
3.4 Attentional system 46
3.6 Target field and reference field 53
3.5 Target object and reference object 53
3.7 Spatial transformation for shift 57
3.8 Spatial transformation for rotation 58
3.9 Spatial relation CoS and CoD 59
3.10 Memory nodes and production nodes 64
3.11 Relational concepts 70
3.12 Process 71
3.13 Hierarchical organization of processes 74
3.14 Heterarchy of all processes 75

4.1 Grounding, unique target, single feature (color) 88
4.2 Grounding, nonunique object, single feature (motion direction) 91

4.3 Grounding, nonexistent object, single feature (motion direction) 95
4.4 Grounding, unique target, feature conjunction 99
4.5 Grounding, nonunique object, feature conjunction . . . 101
4.6 Grounding, nonexistent object, partially matching feature conjunction 103
4.7 Grounding, nonexistent object, multiple partial matches of a feature conjunction 104
4.8 Grounding, unique object, spatial relation 111
4.9 Grounding, unique object (hypothesis testing), spatial relations 114
4.10 Grounding, nonunique object, movement relation . . . 117
4.11 Grounding, nonexistent object, movement relation . . . 120
4.12 Describing a single object 124
4.13 Describing a spatial relation 128
4.14 Describing a movement relation 131

List of Tables

3.1 Overview of fields and nodes of the model 42
3.2 Overview of mathematical variables 43

4.1 Grounding tests: single feature (color) 86
4.2 Grounding tests: single feature (motion direction) . . . 87
4.3 Grounding tests: feature conjunctions 97
4.4 Grounding tests: spatial relations, unique target 106
4.5 Grounding tests: spatial relations, nonunique objects . 107
4.6 Grounding tests: movement relations, unique target . . 108
4.7 Grounding tests: movement relations, nonunique objects 109
4.8 Description tests: single objects 122
4.9 Description tests: multiple objects 126

C.1 Videos used for each test 196

Introduction 1

We are on the brink of a paradigm shift in computing. For decades we have interacted with technology by typing on keyboards, pressing mouse buttons, and touching screens. This will be forever changed by the advent of conversational interfaces. They will enable us to interact with technology by speech alone, which will have a major impact on our daily lives. No longer will we have to clumsily type on the tiny screens of our smartphones. No longer will we have to learn new interfaces with every new generation of an app. And no longer will we have to switch back and forth between multiple apps and websites just to schedule a meeting with someone. Conversational interfaces will enable us to control all of our software on all of our devices; not only the apps on our phone and computer, but also the navigation system in our cars, the lights and music in our homes, and the blinds and elevators in our buildings. All we will have to do is talk, which comes naturally to us.

This vision has been portrayed as science fiction for a long time but it is now quickly becoming reality. From "Siri", to the "Google Assistant", "Cortana", and "Alexa", all of the big technology companies are offering a conversational interface. And the interfaces are quite impressive. You can say to your phone "How do I get from here to Hamburg International Airport?" and you will be guided to the airport along the fastest route from your current position. You can ask "What is a grizzly bear?" and you will instantly be shown a photo and read a description. However, conversational interfaces neither understand what it means for a human to have to navigate nor do they understand what a grizzly bear is. This is a critical point. These interfaces can only put into words information that is made available by apps, databases, or websites on the internet. What is missing is the connection of that information to experiences in the

real world. Until we dig deeper, conversational interfaces will remain just that—interfaces.

The grand vision that is driving researchers is much more elusive. In that vision, we can hold an actual conversation with a machine. It understands the meaning of words, is able to solve problems for us, and anticipates what we may need next. It infers information from background knowledge and understands underlying problems as well as constraints that we have not explicitly stated. It is the vision of a general artificial intelligence—a virtual assistant.

But how can we get there from the conversational interfaces available today? What else is required once they reliably recognize the words we utter? How can we create a system that understands the meaning of words such as "grizzly bear" in a similar way that humans do, for instance recognizing a grizzly bear in a video or imagining one roaming in the wild? How can we get such a system to describe what it is perceiving or imagining, be it a static scene, an unfolding event, or an abstract conceptual structure; again, not by simply reading off information but by analyzing images, videos, and documents? And how can we have the system do all of this autonomously, without being explicitly guided by us, while ultimately acting in our interest? In essence, it is the question of the *grounding of language*; the question of how the production and comprehension of language is rooted in more general cognitive processes, in perception, motor behaviors, reasoning—ultimately, in experiences and memories of the real world.

We know that the human brain holds the answers to all of these questions. Unfortunately, bringing these answers to light has proven to be a serious endeavor. Despite tremendous progress in understanding the anatomy and detailed mechanisms of the human brain, the processes that give rise to cognition and language are not yet well understood. What is lacking is a general theory that explains cognitive processes in a formal way, based strictly on what we currently know about the human brain; a theory that encompasses all processes from a basic sensorimotor level up to the abstract level of language. A candidate for this is *dynamic field theory (DFT)*, a mathematical modeling framework that is deeply rooted in our understanding of the human brain. DFT has established itself in a broad range of academic fields and is able to capture cognitive processes ranging from the simple detection and selection of objects, to processes of attention, scene representation, and sequence generation (Schöner et al., 2015). This thesis builds on these foundations to capture the representations and cognitive processes underlying the grounding of language; it thus brings DFT one step closer toward a neural theory of language.

Schöner, G., Spencer, J. P., & the DFT Research Group. (2015). *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York: Oxford University Press

DFT posits that cognition is based on continuous perceptual

representations that are close to the sensorimotor layer, for instance a representation of the spatial position of objects on a table surface. Language, on the other hand, rests upon discrete symbolic representations (e.g., words and concepts such as RED or LEFT). Thus, in its most basic form, the grounding of language requires that a connection is established between discrete and continuous representations, based on the neural principles of DFT. Furthermore, it requires that a fundamental aspect of human intelligence is addressed that presents itself most prominently in language: being able to acquire and express systematic relations among multiple elements (Hummel & Holyoak, 2005; Halford et al., 2010). While relations can be expressed particularly well in language and symbolic representations in general (Hummel, 2011), it is a challenge to do the same based on continuous representations. A starting point may be *spatial relations*, where the relational information is embedded within the continuous space that objects occupy. Two forms of spatial relations can be distinguished: *basic spatial relations* simply express that an individual object is at a certain position within a continuous space, while *deictic spatial relations* express the spatial relation between multiple objects in a scene[1] (e.g., the relation TO THE LEFT OF) (Logan & Sadler, 1996). In scenes with moving objects, *deictic movement relations*, such as MOVING TOWARD, form the basis for expressing the meaning of actions. This is highly relevant for language as a majority of verbs refer to actions (Pulvermüller, 2005). Understanding the neural processes that resolve spatial and movement relations is thus fundamental to understanding the link between language and cognition.

The first goal of this thesis is thus to capture the neural processes that govern how both basic and deictic spatial relations, as well as deictic movement relations, are extracted from and expressed in continuous perceptual representations. The second goal is to establish the neural processes that enable a mapping between these continuous representations and discrete representations that may interface with language. The third goal, and a particular focus of this thesis, is to capture the principles by which neural processes may perform a grounding *autonomously*, that is, without additional algorithmic approaches or human intervention.

This thesis addresses these challenges by introducing a concrete neural process model based on DFT that can solve basic language tasks. The model continuously receives real-world sensory information from a camera, showing a white table with a few colored balls on it—this feeds a continuous representation of the world. The model also receives input that corresponds to a phrase such as "Find the red object to the left of the green object" or "Find the red object moving toward the green object"—this feeds a discrete representa-

Hummel, J. E. & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture. An overview of the LISA project. *Current Directions in Psychological Science*, *14*(3), 153–157; Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, *14*(11), 497–505

Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, *23*(2), 109–118

[1] A third form, *intrinsic relations*, is a variant of deictic relations that takes into account the intrinsic reference frame of objects. This will not be covered in this thesis but has been addressed previously (van Hengel, Sandamirskaya, Schneegans, & Schöner, 2012).

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13, pp. 493–529). Cambridge, MA, USA: MIT Press

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(July), 576–582

tion of concepts. The model is able to ground the objects described in the phrase, that is, it is able to match the language description to the corresponding objects in the visual scene. In a second type of task, it is able to generate a phrase itself, in essence describing parts of the visual scene.

What is required of the model to solve these tasks? First, it requires that the discrete concepts the phrase consists of, such as RED and GREEN, are mapped onto continuous feature representations. These feature representations must guide visual search to bring matching objects into the attentional foreground (Logan, 1994). If the phrase refers to multiple objects, each object must be attended to individually and in sequence (Franconeri et al., 2012), while maintaining the binding of each object to its role (Logan & Sadler, 1996). This attentional process must lead to stable mental representations of the objects in working memory, holding all their feature values, including their spatial position. While these representations must persist over time, they must also allow to be updated whenever relevant changes occur in the scene. The spatial positions of the objects represent their relational information only implicitly (Franconeri et al., 2012). The relative position between the objects must thus be constructed and explicitly represented by adjusting the reference frame of the spatial representation (Logan & Sadler, 1996). An additional adjustment of the reference frame is required to extract movement relations. To compare how well the relative position of the objects fits the specified spatial relation (e.g., TO THE LEFT OF), a continuous representation of the spatial relation, the spatial template, must be imposed on the representation of the relative position (Logan & Sadler, 1996). If it does not fit well, other objects have to be selected and the process repeated. Most importantly, depending on the task and the visual scene, different subsets of the above processes must be organized to become active in a sequential order that solves the task. It is crucial that this organization unfolds solely on the basis of the internal dynamics of the model, the continuous sensory input, and the initial task input. This precludes control inputs by an additional algorithm or by a human user—the organization must come from within the model. Only then can we consider it autonomous.

Achieving all of this in a *neural* process model requires that the following fundamental problems are solved.[2] These problems are specific to neural models and do not present themselves in algorithmic approaches.

First, unlike in a computer program, neural populations cannot define pointers to arbitrary parts of memory and thereby access the information stored there. They can only have an influence on other neural populations if they are connected—and connectivity is fixed,

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1015–1036

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13, pp. 493–529). Cambridge, MA, USA: MIT Press

[2]The description of these fundamental problems is based on published material (Richter, Lins, & Schöner, 2016, 2017), which arose as collaboration between myself (MR), Jonas Lins (JL), and Gregor Schöner (GS). JL and MR developed the model together and collaborated in conceptual thinking. MR implemented the model, and generated results. All authors participated in writing the papers.

at least on short time scales. Applying a neural operator to a location that is represented by a neural population is thus possible only if it is connected to that location. However, connecting operators to every location in a neural population would require unrealistic neural resources. The alternative is to connect the operator to only one default location and shift the representations of objects to that location. This is analogous to the concept of an attentional neural pointer of Ballard et al. (1997) and is achieved here by steerable neural mappings (Schneegans & Schöner, 2012).

Second, for similar reasons of limiting the required neural resources, the nervous system represents high-dimensional visual information in multiple low-dimensional neural feature maps. To refer to any particular object, corresponding representational pieces must be bound together. The model employs a neural implementation of the classical idea of binding through space (Treisman & Gelade, 1980), where every feature map is endowed with a spatial dimension that is shared across maps (Schneegans, Spencer, & Schöner, 2015). The shared spatial dimension then requires that multiple objects are processed sequentially in time.

Third, the discrete processing steps this implies and that are critical to all of cognition are natural in algorithmic accounts but hard to achieve in neural process models, where neural activation evolves continuously in time under the influence of input and recurrent connectivity. In this model, discrete events emerge from continuous neural dynamics through dynamic instabilities, at which the match between neural representations of intentional states and their conditions of satisfaction are detected (Sandamirskaya & Schöner, 2010).

All of the problems above are solved by the model introduced in this thesis. It captures a variety of different tasks and visual scenes autonomously with a single set of parameters. This thesis thus shows that a basic grounding of language can be accomplished by a neurally plausible cognitive architecture.

The remainder of the thesis is organized as follows. Section 2 covers the required background knowledge. It summarizes different notions of grounding and how they are addressed in state-of-the-art computational models. It goes on to review the concepts and mathematical foundation of DFT that the rest of the thesis is built upon. Section 3 contains a conceptual and mathematical description of the model that this thesis introduces. Section 4 demonstrates the capabilities of the model in 104 experiments, out of which 14 are explained in detail. Section 5 discusses both the conceptual contributions of this thesis as well as the specific novel implementations found in the model. The thesis is briefly concluded in Section 6.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–767

Schneegans, S. & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136

Schneegans, S., Spencer, J. P., & Schöner, G. (2015). Integrating "what" and "where": Visual working memory for objects in a scene. In G. Schöner & J. P. Spencer (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (Chap. 8, pp. 197–226). New York: Oxford University Press

Sandamirskaya, Y. & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179

Previously published material

This thesis draws on material previously published with various coauthors:

- Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Science*, *9*(1), 35–47

- Richter, M., Lins, J., & Schöner, G. (2016). A neural dynamic model parses object-oriented actions. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1931–1936). Austin, TX: Cognitive Science Society

- Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2847–2852). Austin, TX: Cognitive Science Society

- Richter, M., Lins, J., Schneegans, S., & Schöner, G. (2014). A neural dynamic architecture resolves phrases about spatial relations in visual scenes. In S. Wermter (Ed.), *Artificial Neural Networks and Machine Learning: ICANN 2014, 24th International Conference on Artificial Neural Networks, Lecture Notes in Computer Science 8681* (pp. 201–208)

Background 2

This chapter contains background information on both the grounding of language and on dynamic field theory (DFT). Section 2.1 summarizes what the *grounding of language* is, why grounding is necessary to enable a cognitive system to form an understanding of the world, and how the problem of grounding is approached in state-of-the-art computational models. These models range from purely algorithmic solutions to approaches based on neural principles. They lead to the conclusion that human cognition can be understood most directly based on models that capture neural processes. The mathematical framework of DFT enables building such neural process models. Section 2.2 reviews all the concepts and mathematical foundations of DFT that are the foundation for the model introduced in Section 3.

2.1 Grounding of language

The grounding of language refers to the connection between language and the physical world, the connection to colors and shapes, to objects and people, to scenes that are motionless or those that are full of life. Before defining the grounding of language more clearly, it serves to establish a common terminology.

2.1.1 Terminology

The literature on the grounding of language suffers from the abstractness of the subject matter, which sometimes results in unclear descriptions. To clear up the most important terminology, I will use an example to refer to the definitions brought forth by Steels (2008).

Steels, L. (2008). The symbol grounding problem has been solved. So what's next? In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and Embodiment: Debates on Meaning and Cognition* (pp. 223–244). New York: Oxford University Press

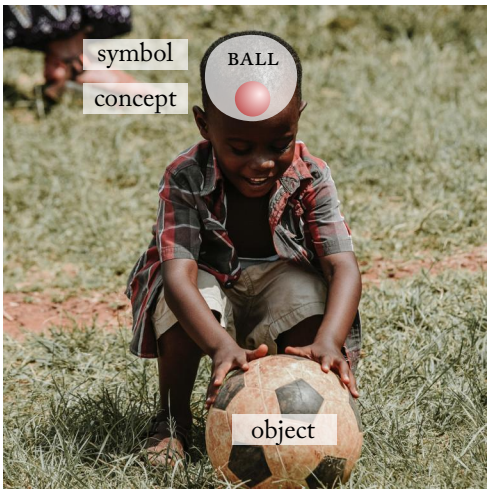


FIGURE 2.1: The actual ball in the physical world is the *object*, the mental representation (red sphere) of the ball is the *concept*, and the name (BALL) for both the concept and the object is the *symbol*.

Suppose you are looking at a soccer ball that is lying on the ground in front of you (Figure 2.1). The actual ball in the physical world that you can see, touch, and kick, is an *object*. You perceive it through your sensory system and form a perceptual (modal) representation of it in your mind, based on continuous feature spaces. This representation embodies your *concept* of a ball (denoted as a red sphere in Figure 2.1) that you have learned from previous perceptual experiences. It enables you to recognize the soccer ball as a ball but also to imagine and think about other balls. The perceptual representation of a single concept may be distributed over multiple modalities and features, which are linked together in a discrete amodal representation of the concept. Finally, the word BALL that refers to both the concept and the object is a *symbol*. It is symbolic because the form of the symbol is independent of the features of the concept or object; its form is arbitrary. That is, we could choose to refer to the concept or object by a different symbol, as is done in other languages, and it would not have an influence on our understanding of the concept. Bear in mind that symbolic representations do not necessarily have to be based on words. In fact, if the discrete amodal representation of a concept were operated on without taking into account the perceptual representation it links together, it could also be viewed as a symbol.

2.1.2 What is grounding?

Given these definitions, we can make the different meanings of *grounding* more explicit. First, there is a notion that grounding is a static property of a cognitive system. In this view, a symbol is regarded as grounded when a connection exists between the symbol and a concept (Harnad, 1990; Cangelosi & Riga, 2006; Cangelosi, 2010). Similarly, a concept is regarded as grounded when a method exists to establish a connection between the concept and an object in the world (Steels, 2008). But grounding is also understood as a *process*. This includes the process by which a connection between a concept and a concrete object is established (Gorniak & Roy, 2004; Roy, 2005b), for instance by identifying a familiar object or category. It also includes the process by which a concept (and its symbol) are learned in the first place (Cangelosi & Harnad, 2001; Harnad, 1990). Finally, there is the notion of a symbol being "socially grounded", that is, it is learned and distributed among an entire population of communicating individuals (Steels & Kaplan, 2002; Steels, 2003).

While all of these notions of grounding are complementary and valid, in this thesis, grounding is primarily understood as the process by which a concept is connected to an actual object in the world.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346; Cangelosi, A. & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, *30*(4), 673–689; Cangelosi, A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, *7*(2), 139–151

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470; Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205

Cangelosi, A. & Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, *4*(1); Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346

This can happen in two directions. First, a concept can be activated through its symbol by language input, which then drives attentional processes that single out the matching object in the world. Second, a salient object in the world can attract attentional focus and activate a matching concept, which may turn activate a corresponding symbol. In this thesis, the latter process is also referred to as "describing" but it is a form of grounding nonetheless. After a grounding process is successfully finished, the symbol and concept are regarded as grounded.

Steels, L. & Kaplan, F. (2002). AIBO's first words. The social learning of language and meaning. *Evolution of Communication*, *4*(1), 3–32; Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308–312

This thesis does not address the processes by which concepts and symbols are learned. Within the presented DFT model, all concepts (i.e., concepts of color, motion direction, and spatial relations) are built in by hand. However, please note that the substrate in which concepts are represented is open to learning through established neural learning mechanisms, for instance Hebbian learning. The notion of "social grounding" is also not addressed in this thesis. The model introduced in this thesis only comprises an *individual* cognitive system that autonomously perceives its surrounding, describes it, and grounds language input.

2.1.3 Why is grounding necessary?

For a long time, cognition was thought to be a process taking place all but removed from the physical world surrounding us. Classical research in cognition was built on the conviction that language and cognition can best be explained as the processing of abstract symbols (Fodor & Pylyshyn, 1988). The characteristic properties of cognition that the research tried to capture include productivity, systematicity, compositionality, and the coherence of inference,[1] all of which are aspects of a feature of cognition that enables us to combine and generalize knowledge in systematic ways. These properties present themselves most prominently in language but are also thought to be hallmarks of general cognition. Symbolic approaches have been successful because they capture these properties.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71

The symbols upon which these approaches of cognition are built as well as the rules that define their processing are denoted using arbitrarily chosen words or variable names. For example, the statement "John, Mary, and Alice went to the city." can be represented by the symbolic statement $J\&M\&A$, where the symbols J, M, and A respectively represent that John, Mary, and Alice each went to the city. Based on such a symbolic representation, inferences can be made using explicit rules. For instance, the rule $J\&M\&A \rightarrow J$ enables the inference that if it is true that "John, Mary, and Alice went to the city" ($J\&M\&A$), then it is also true that "John went to the city" (J). Such rules and the inferences on them do not nec-

[1] The properties of language and cognition that are deemed characteristic by classic symbolic accounts are explained by Fodor and Pylyshyn (1988). Here is a short summary:

Productivity We are able to produce a seemingly unlimited number of distinct sentences based on a finite vocabulary.

Systematicity If we understand a sentence, we also always understand other sentences that have the same structure but different content.

Compositionality For systematicity to work, the content that is different must fulfill a similar semantic function.

Coherence of inferences Inferences must be applicable across all possible statements.

essarily have to be implemented algorithmically based on variable names and strings, but can also be expressed in neural terms, where concepts and relations are represented by networks of nodes (Shastri, 1999; Bergen & Chang, 2005). This approach may seem to be more closely connected to human cognition because operations can be expressed by connectionist networks (Shastri et al., 2002), but the representation remains symbolic whether the representational format is neurally plausible or not.

The important question is whether these symbols have any meaning apart from what we prescribe to them. Does the system itself *understand* that if *J* is true, this represents that John went to the city? For instance, can the system imagine John walking along streets, looking at windows, and eating ice cream? Can such an understanding of the world arise based solely on arbitrary symbols and their relations?

Harnad (1990) prominently argued that symbolic systems can never acquire such an understanding. He called this the *symbol grounding problem*, which he illustrated with the following example.[2] Suppose you had to learn Chinese and all you had to go on was a Chinese-Chinese dictionary. Learning the new language would be immensely difficult because the dictionary would "explain" one meaningless and arbitrary symbol using more meaningless and arbitrary symbols. No matter how many of the symbols you tried to look up, you would not be able to find any meaning in them. You may be able to extract meaning by analyzing the frequency of certain symbols and thereby making assumptions about what certain symbols could be referring to. However, decoding the meaning of the symbols this way only connects symbols to meaningful concepts that you already have, concepts that are meaningful in another language that you speak. The meaning would be "parasitic" in your first language, instead of intrinsic in the Chinese symbols. If the task was to learn Chinese as a *first* language, from the dictionary alone, the task would be impossible. Harnad (1990) thus concluded that at least some elementary symbols need to be grounded, that is, they need to be connected to representations of objects in the real world. These *iconic representations* must be non-symbolic insofar as they are transformed projections of the objects we perceive. From these iconic representations, we can form *categorical representations*, which are also non-symbolic but enable us to discriminate qualitatively different representations and identify certain ones. Iconic and categorical representations thus correspond to the respective notions of modal and amodal representations of a concept, as defined above and used in this thesis. For both representations, Harnad (1990) suggests connectionist approaches to be a good candidate. In his view, elementary symbols would then connect to categorical

Shastri, L. (1999). Advances in SHRUTI: A neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence*, *11*, 79–108; Bergen, B. K. & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (Eds.), *Construction Grammars: Cognitive Grounding and Theoretical Extensions* (Chap. 6, pp. 147–190). Amsterdam/Philadelphia: John Benjamins Publishing Company

Shastri, L., Grannes, D., Narayanan, S. S., & Feldman, J. (2002). *A Connectionist Encoding of Parameterized Schemas and Reactive Plans* (tech. rep. No. TR-02-008). International Computer Science Institute. Berkeley

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346

[2]The example is based on the "Chinese room argument" by Searle (1980), which was given in a more general context to show that formal AI systems do not understand the subject matter they are processing. The reasoning and conclusion are the same as in Harnad's example: without grounding, formal symbol systems cannot have intentionality (be about something) and therefore cannot understand anything.

representations, giving them a name. As we will see in Section 3, the DFT model introduced in this thesis is largely consistent with this suggested solution for the symbol grounding problem.

2.1.4 Embodied cognition

The insight that classical symbolic theories of cognition are deficient because they lack grounding goes along the mounting empirical evidence that various sensory-motor areas in the brain are active during cognitive tasks (for review, see Pulvermüller, 2005), as well as behavioral evidence that shows graded effects of language on motor tasks (e.g., Glenberg & Kaschak, 2002; for review, see Kaschak & Jones, 2014). This evidence suggests that there is no clear divide between cognitive processes and the underlying grounding in perception (Gallese & Lakoff, 2005). The focus of research thus shifted to include the connection between cognition, language, and the human body in the physical world—a shift toward an *embodied cognition* (Clark, 1999; M. Wilson, 2002).

Barsalou (1999) prominently formulated a model of how the connection between cognition, language, and the world could be established. He formulates a perceptual theory of knowledge and cognition, in which patterns in sensory-motor areas are captured in *perceptual symbols*. Importantly, these perceptual symbols are not holistic images or recordings of the content in sensory-motor areas but consist only of a small subset of distributed perceptual components. Related perceptual symbols are organized into *simulators* that are able to produce infinitely many simulations of a certain entity. In this view, the notion of a simulator corresponds to that of a *concept*, as defined above. For instance, a simulator (or concept) of a chair contains all the aspects of different chairs we have encountered; it is able to produce simulations of simple wooden chairs, adjustable office chairs, camping chairs, and so forth, all with their individual shape, color, texture, and feel. Barsalou (1999) shows that such a perceptual symbol system supports basic tasks like categorization, identification, and categorical inferences. He furthermore shows that it also supports productivity, the formation of propositions, and the representation of abstract concepts (e.g., truth and negation), properties that are deemed critical to a fully functional conceptual system and that were believed to be incompatible with perceptual theories. His theory thus shows that grounded accounts of cognition can, in fact, capture a similar functionality as classical symbolic accounts—without their inherent problems.

His idea of a perceptual symbol system was criticized for being only a verbal theory, just a vision, instead of a concrete model (Dennett & Viger, 1999). The same critique applies to similar ver-

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(July), 576–582

Glenberg, A. M. & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*(3), 558–565; Kaschak, M. P. & Jones, J. L. (2014). Grounding language in our bodies and the world. In Thomas Holtgraves (Ed.), *The Oxford Handbook of Language and Social Psychology* (Chap. 20, pp. 317–329). London: Oxford University Press

Gallese, V. & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3), 455–479

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, *3*(9), 345–351; Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–36

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609, 577–609

Dennett, D. C. & Viger, C. D. (1999). Sort-of symbols? [Peer commentary on "Perceptual symbol systems" by Lawrence W. Barsalou]. *Behavioral and Brain Sciences*, *22*(4), 613

bal theories brought forth in the community of cognitive linguistics that resonate with Barsalou's perceptual symbol system. For instance, the ideas of cognitive grammar (Langacker, 1986), grammatical construal (Talmy, 1988), and image schemas (Gibbs & Colston, 1995) explain characteristics of cognition and language on a similar intuitive level but do not address how they could be implemented.

What is clearly missing is a computational theory, a theory that explains the representations and processes that give rise to cognition.

Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive Science*, *10*, 1–40

Talmy, L. (1988). The relation of grammar to cognition. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics* (pp. 165–205). Amsterdam/Philadelphia: John Benjamins

Gibbs, R. W. & Colston, H. L. (1995). The cognitive psychological reality of image schemas and their transformations. *Cognitive Linguistics*, *6*(4), 347–378

2.1.5 Embodied computational models

Computational models often uncover problems that would remain hidden in purely theoretical accounts. This is because they force modelers to think about and implement all the processes that are required to solve a particular problem. This can lead to a deeper understanding of the problem.

An example of computational models is the work of Deb Roy, whose goal is closely aligned with this thesis, at least from a functional point of view. He aims at building conversational robots that are able to solve language grounding tasks (Roy, 2005b, 2008). The robotic implementation requires that mechanistic models are built of all the processes involved. Instead of modeling isolated parts of language, he advocates for building a holistic model of language that covers all of its layers, from auditory speech input, to structured symbolic representations, to representations that are grounded in and connected to concrete sensors and motors. To keep this project manageable in scope, he only includes a small subset of linguistic features in his models. He compares this to language at a child's level, with only a rudimentary grammar and a small vocabulary. Additionally, the language focuses on describing objects, relationships, and actions in the here-and-now and is void of abstract concepts, metaphors, and past and future events. His concrete robotic models cover an impressive number of different tasks. These include grounding descriptive language in the environment (Gorniak & Roy, 2004), as well as generating descriptions of the environment, answering questions, and mental imagery (Mavridis & Roy, 2006). In fact, the system is able to pass parts of the "Token test", an assessment of language abilities in children ages 3 to 12.

The computational models by Yiannis Aloimonos are focused on generating descriptions of actions in videos. He is inspired by Chomsky's Minimalist framework for language and extracts a syntactic structure of actions based on a generative grammar (Pastra & Aloimonos, 2012). The constituent parts of the grammar (e.g.,

Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205; Roy, D. (2008). A mechanistic model of three facets of meaning. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and Embodiment: Debates on Meaning and Cognition* (pp. 1–32). Oxford, UK: Oxford University Press

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470

Mavridis, N. & Roy, D. (2006). Grounded situation models for robots: Where words and percepts meet. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on* (pp. 4690–4697). IEEE

Pastra, K. & Aloimonos, Y. (2012). The minimalist grammar of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 103–117

terminals) are defined in the sensorimotor domain and are recognized in input videos. They are learned from motion capture data of real humans that is algorithmically segmented and transformed into a sequence of symbolically represented segments, which are then grouped into more macroscopic structures (Guerra-Filho & Aloimonos, 2012). In addition to recognizing body parts and their motor primitives, the model can detect contact between body parts and other objects. This in particular enables it to infer relations between body parts and objects, for instance when an object is being used as a tool. The parsing of action is also addressed in a preliminary model based on DFT (Lobato et al., 2015). It receives input from a three-dimensional camera (Kinect) and is able to parse hand actions that are directed at objects, for instance reaching, grasping, and dropping. While the representation of objects and some of the model's process organization is based on neural dynamics, crucial parts of the problem are solved algorithmically, including how relations between objects in the scene are computed.

Guerra-Filho, G. & Aloimonos, Y. (2012). The syntax of human actions and interactions. *Journal of Neurolinguistics*, *25*(5), 500–514

Lobato, D., Sandamirskaya, Y., Richter, M., & Schöner, G. (2015). Parsing of action sequences: A neural dynamics approach. *Paladyn, Journal of Behavioral Robotics*, *6*, 119–135

The computational models such as those described above are impressive and go beyond verbal theories in the sense that they may uncover problems at the computational level that may not have been apparent before. However, the models are implemented without constraints on the computational operations. This is unlikely to lead to deep insights about human cognition, because the neural operations that the human brain employs to solve a problem may differ significantly from the computational operations used to solve the problem in the model. In order to gain insights about human cognition, it is therefore imperative that computational models rest upon established principles of neural processing. While this is a demanding in its own right, the combinatorial structures prevalent in language are a particular challenge to represent in neural systems, leading to theoretical problems like the binding problem, the problem of 2, or the problem of variables (van der Velde & de Kamps, 2006; Jackendoff, 2002). In computational models, these problems are, so far, often ignored. The models instead focus on the more concrete problems of grounding concepts and relations in the real world.

van der Velde, F. & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*(1), 37–108; Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. New York: Oxford University Press

An example is an approach developed by Peter Dominey and colleagues that is closer to neural accounts of cognition. It consists of an architecture that is able to learn grammatical constructions[3] from observing a scene in which objects interact, while also listening to a spoken language description of what is happening in the scene (Dominey & Boucher, 2005a). The architecture parses the spoken language, separating it into open-class and closed-class words.[4] From a visual scene, it extracts a representation of an ongoing action, as well as the objects that are engaged in that action. This

[3] *Grammatical constructions* are the units of language in theories of construction grammar (e.g., Goldberg, 1992). Constructions are syntactic templates that express the correspondence between form and meaning. They can range in complexity from morphemes, to words, up to entire sentences.

Dominey, P. F. & Boucher, J. D. (2005a). Developmental stages of perception and language acquisition in a perceptually grounded robot. *Cognitive Systems Research*, *6*(3), 243–259

[4] *Open-class* words are the category of content words (i.e., nouns, lexical verbs, adjectives, and adverbs) that are open to new members. Examples are "grizzly bear", "run", and "big". *Closed-class* words are the category of function words that does not accept new members. Among others, this category includes conjunctions (e.g., "and"), articles (e.g., "the"), and prepositions (e.g., "from").

is done algorithmically based on predetermined sequences of contact between objects. The architecture then learns the connections between each word, its grammatical role in the sentence, and its corresponding representation by relying on the structure of closed-class words in the sentence. This structure is regarded as a grammatical construction, which can be learned and enables the architecture to generalize over different content. The architecture is able to ground sentences in the active and passive form for simple actions, for instance "The block pushed the triangle", complex actions like "The block that pushed the triangle touched the moon", as well as spatial relations. The architecture was later extended to enable it to produce language about observed videos, essentially describing the scene, and to answer questions about it in conversational form (Dominey & Boucher, 2005b). Dominey (2007) sketched how the idea of constructions could be generalized to represent physical events as well as social aspects of interaction. The representations of objects and words used within the core of the architecture are based on binary vectors that lend themselves to neurally inspired learning mechanisms. These vectors are in fact arbitrary symbols; what is missing is a direct connection to the sensorimotor system. Madden et al. (2010) present a hybrid architecture that extended the symbolic core with an embodied simulator. Compared to the work of Deb Roy, Dominey's work is closer to neural approaches for two reasons. First, the internal representations could, in theory, be implemented with neurons. Second, Dominey often explicitly states in which areas of the brain certain mechanisms would be located (e.g., Madden et al., 2010).

Dominey, P. F. & Boucher, J. D. (2005b). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1-2), 31–61

Dominey, P. F. (2007). Towards a construction-based framework for development of language, event perception and social cognition: Insights from grounded robotics and simulation. *Neurocomputing*, *70*(13-15), 2288–2302

Madden, C., Hoen, M., & Dominey, P. F. (2010). A cognitive neuroscience perspective on embodied language for human-robot cooperation. *Brain and Language*, *112*(3), 180–188

The work of Angelo Cangelosi focuses on modeling how grounded concepts are *acquired* both through evolutionary and developmental processes (for review, see Cangelosi, 2010). Moreover, he addresses a common critique of embodied approaches to language: that some concepts, for instance "goodness", "truth", or "beauty", are too abstract to have a grounding in sensorimotor modalities. He shows that once a basic repertoire of grounded concepts has been learned, new concepts can be acquired by composing them from previously learned ones. Importantly, the learning is based only on language input and does not ground the new concepts in perceptual representations. Instead, only an amodal representation of the new concept is connected to the amodal representations of the previously learned concepts. This shows that more abstract concepts can arise without a need to ground them directly. Moreover, he shows that this form of concept learning may be more adaptive than learning concepts by grounding them completely. For this, Cangelosi and Harnad (2001) simulated a population of organisms that forage for mushrooms in a two-dimensional grid-world. Each organism is

Cangelosi, A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, *7*(2), 139–151

Cangelosi, A. & Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, *4*(1)

controlled by a three-layer feedforward neural network. As input, the network receives a binary code that corresponds to feature values describing the mushroom as well as linguistic input describing actions. The output of the network controls the movement of the organism, action like eating, marking, or returning to a mushroom, and it issues linguistic calls to other organisms. Organisms that learn the action "return" only from calls of other organisms (and without feature input) return more mushrooms than organisms that learn the action by grounding it in feature input (but without language input). Cangelosi and Riga (2006) build upon this work and show in a simulated robotic scenario that with additional phases of learning even more abstract concepts can be learned. Stramandinoli et al. (2012) further extend this paradigm to a simulated iCub robot that learns concepts from a human instructor. Based on a recurrent neural network, this model can learn an action that is composed of a specific *sequence* of other actions. Similar to the models by Peter Dominey, the input to the neural networks in Cangelosi's models is based on a binary code, where each input node represents the presence of a certain feature. Again, what is missing is the direct connection to the sensorimotor system.

Cangelosi, A. & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, *30*(4), 673–689

Stramandinoli, F., Marocco, D., & Cangelosi, A. (2012). The grounding of higher order concepts in action and language: A cognitive robotics model. *Neural Networks*, *32*, 165–173

Since language is spoken within a population and is thus a social act, one may understand the process of grounding perceptual categories as a social process as well. That is, additionally to the grounding that happens within the individual, the interaction with other individuals and the culture of the population should have an influence on the grounding. Luc Steels and collaborators explore the hypothesis that from a very early stage, categories are formed and words are learned under a strong influence of culture (Steels & Kaplan, 2002). They focus on how the perceptual grounding of concepts is influenced by language interaction and how grounded concepts are coordinated within a population. Steels (2003) investigates these questions in "language games", where a population of embodied agents interacts in a shared physical (or simulated) environment and communicates via their sensorimotor system. One such game is the "Talking Heads experiment", in which two out of a population of thousands of agents "talk about" an object in a shared environment. Over time, the population of agents develops a coherent set of categories for objects as well as a coherent vocabulary connected to the categories. In fact, throughout the population, the categories are significantly more similar than in similar games that do not incorporate language interaction. Other simulations that investigate category formation by comparing the mechanisms of evolution, learning (based only based on visual input), and social learning (with language), show similar results (Steels & Belpaeme, 2005). While all mechanisms produce adequate sets of categories

Steels, L. & Kaplan, F. (2002). AIBO's first words. The social learning of language and meaning. *Evolution of Communication*, *4*(1), 3–32

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308–312

Steels, L. & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–489, 469–489

for the given (color) stimuli, these are only shared between individuals that interact through language. Similarly, language games between the AIBO robot and a human mediator show that the robot learns words and categories better the more the mediator interacts via language (Steels & Kaplan, 2002). These findings suggest that social interaction via language has a strong influence on the formation of categories and thus on the way the words are grounded. Luc Steel's work is of interest here less because of the notion of social grounding, which is not addressed in this thesis, but because, in contrast to the approaches discussed earlier, his models have a direct connection to the sensorimotor system. For instance, colors are categorized with adaptive networks (a modification to radial basis function networks) and the names of colors are modeled with associative memory networks (Steels & Belpaeme, 2005).

Steels, L. & Kaplan, F. (2002). AIBO's first words. The social learning of language and meaning. *Evolution of Communication*, *4*(1), 3–32

2.1.6 Process organization

Cognitive architectures often ignore or do not explicitly model the problem of *process organization*, that is, how individual cognitive operations or processes are organized in time to form a coherent overall behavior. These processes may govern anything inside of a cognitive architecture, from bringing the attentional focus to a particular location in space, making a selection decision between multiple options, to activating and deactivating entire behaviors. The problem of process organization includes the question of how such discrete processes are represented in the first place: how is their beginning triggered, how is their state represented during execution, and how and under what conditions are they terminated? How is a multitude of processes organized in such a way that only those are activated that are relevant to the current situation? How may some processes become active simultaneously, while others are constrained to not be active at the same time? Finally, how are processes organized to be executed in a sequence or even in a particular serial order (Henson & Burgess, 1997)?

Henson, R. N. A. & Burgess, N. (1997). Representations of serial order. In J. A. Bullinaria, D. W. Glasspool, & G. Houghton (Eds.), *4th Neural Computation and Psychology Workshop, London 9-11 April 1997: Connectionist Representations* (pp. 283–300). London, UK: Springer

In computational and robotic models the problems named above may not even appear to be a challenge. They are either solved directly through algorithmic tools, for instance `if`-statements, `while`-loops, and the sequential execution of program code—tools that were developed to solve that same problem for computer program code. Or they are solved based on structural principles, for example schema theory in the work of Roy (2005b). However, underneath, such structural principles are implemented by algorithmic tools as well.

Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205

Classical symbolic architectures, for instance ACT-R, SOAR, or ICARUS (for review, see Langley et al., 2009) organize their

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141–160

operations based on symbolic production rules. These production rules are akin to `if`-statements, as they express conditions under which appropriate actions should be executed. More importantly, however, how production rules are organized in memory slots and processed is also determined by common algorithms that are often not discussed.

Using algorithmic approaches may be reasonable if one is interested in solving problems that require cognitive operations or if one believes that human cognition can be explained on an abstraction level that is detached from the neural reality of the human brain. However, even many computational models that are in part based on neural principles (e.g., Cangelosi & Harnad, 2001; Steels & Belpaeme, 2005; Dominey & Boucher, 2005b) rely on algorithmic approaches to control different behaviors or generate sequences. But by doing so, they sidestep the problem of process organization, which quickly becomes complex when based on neural principles (e.g., Shastri, 1999; Shastri et al., 2002).

In the human brain, processes are believed to be organized by a combination of two mechanisms. First, organization occurs distributed throughout cortex by properties emergent from the underlying neural dynamics. Second, as a central device for action selection, the basal ganglia are believed to moderate between distant regions of cortex (Redgrave et al., 1999). The basal ganglia have connections to and from an extensive number of cortical regions. By default, their connections to cortex are believed to inhibit any behavior or process until it becomes relevant. When it does, the inhibition is removed and the behavior or process can become active (Gurney et al., 2001).

A neural model whose processes are organized by a model of the basal ganglia is the Semantic Pointer Architecture Unified Network (SPAUN; Eliasmith et al., 2012). The entire model is based on neural principles, in particular, spiking neurons. It is able to perform an impressive multitude of different tasks, including image recognition, reinforcement learning, counting, and question answering; all without intervention by the modeler and without modification of the model. The model has a direct connection to visual input, which it also uses to differentiate between different task settings, and to motor output, where it uses a simulated physical arm to respond to queries. The model is still simplified and has many limitations. Most importantly, its perceptual and conceptual representations are restricted to digits, precluding it from reasoning about objects in real environments. This is a prerequisite to building up a representation of a scene and to ground relations and language about real environments. Nevertheless, Semantic Pointer Architecture Unified Network (SPAUN) is an impressive neural process model. That all

Cangelosi, A. & Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, *4*(1); Steels, L. & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–489, 469–489; Dominey, P. F. & Boucher, J. D. (2005b). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1-2), 31–61

Shastri, L. (1999). Advances in SHRUTI: A neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence*, *11*, 79–108; Shastri, L., Grannes, D., Narayanan, S. S., & Feldman, J. (2002). *A Connectionist Encoding of Parameterized Schemas and Reactive Plans* (tech. rep. No. TR-02-008). International Computer Science Institute. Berkeley

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, *89*(4), 1009–1023

Gurney, K., Prescott, T. J., & Redgrave, P. (2001, June). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, *84*(6), 401–10

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205

its processes and behaviors are organized based on a model of the basal ganglia make it all the more remarkable.

Neural models do not necessarily have to incorporate a model of the basal ganglia. It may suffice to model the organization of processes based on similar principles, for instance the removal of inhibition. But neural models must explain *all* of its parts on the basis of neural principles. Only then can they claim to be committed to explaining human cognitive processes that are fully autonomous.

2.1.7 Toward a neural theory of embodied cognition

What is thus missing is a comprehensive neural theory of embodied cognition and embodied language. In order to be comprehensive, it must be a concrete, formal process model that is able to explain cognitive processes ranging from the sensorimotor level all the way up to abstract cognition. In order for it to be neural, all processes must be explained as emergent properties of basic neural principles that are consistent with current empirical data. The theory must explain basic cognitive processes such as detection, and selection of objects, categorization, and identification. Furthermore, in order to explain characteristic properties of cognition and language such as productivity, and propositions, it must support structured representations.

A candidate for such a theory is dynamic field theory (DFT), a modern variant of neural dynamics. The model introduced in this thesis (Section 3) is based on this work. The next section covers the conceptual and mathematical foundation of DFT.

2.2 Dynamic field theory

Schöner, G., Spencer, J. P., & the DFT Research Group. (2015). *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York: Oxford University Press

Dynamic field theory (DFT) is a mathematical framework for modeling cognitive processes. It has been applied to capture and explain many of the issues discussed above (Schöner et al., 2015). Ultimately, it aims at explaining cognition as a whole, in a single, coherent framework.

This section summarizes both the conceptual principles on which DFT is founded as well as its mathematical framework. It focuses on the aspects of DFT that are relevant to the model presented in this thesis. To get a broader overview of DFT, in particular its empirical foundations, please refer to the textbook by Schöner et al. (2015).

2.2.1 Principles of dynamic field theory

DFT models of cognition are *process models*. This means that they are not only aimed at describing what is empirically found in human cognition but that they additionally explain the underlying mechanisms of *how* these findings come about. In other words, they not only model the outcome of cognitive processes but capture the processes themselves. Since human cognitive processes unfold in the nervous system, most prominently the cortex, process models need to be informed by established principles of neural organization. DFT models apply this idea rigorously and require that *all* functionality that is part of a model must be explained on the basis of these neurally plausible principles. This approach wields explanatory power but at the same time produces additional challenges. Some problems that are trivial to solve with conventional computer algorithms require complex solutions when resorting to neurally plausible mechanisms alone.

Many basic cognitive processes have already been addressed by DFT, for instance the detection of an object, the selection between multiple objects, and the build-up of working memory. Starting from these basic processes, DFT enables to build models of increasingly complex cognitive processes. All of it is based on a mathematical framework that captures fundamental neural principles on a level that most directly impacts human behavior. These principles are as follows.

First, DFT is based on the hypothesis that throughout cortex, behaviorally relevant parameters are coded for by populations of neurons, rather than individual neurons. For instance, the movement direction of an arm movement can be accurately predicted using the response of an ensemble of broadly tuned neurons (Georgopolous et al., 1986). In DFT, this idea of *population code* is so pervasive that the models do not include individual neurons. In fact, the models do not even explicitly specify populations of neurons. Instead, they are based on *dynamic neural fields*, activation variables that are defined directly over feature spaces, for instance movement direction, that neural populations represent. This may seem like an abstraction that is quite removed from neural reality. However, the activation of a neural population with respect to some feature dimension, the *distribution of population activation (DPA)*, can be computed by incorporating the entire tuning curve of each neuron within the population (Jancke et al., 1999). Thus, even though the representations that DFT models are based on are not explicitly expressed in terms of neurons, they can be mapped onto neural populations and express what they represent.

Second, DFT models are characterized by their *gradedness*. All

Georgopolous, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal Population Coding of Movement Direction. *Science*, *233*(March), 1416–1419

Jancke, D., Erlhagen, W., Dinse, H. R., Akhavan, A. C., Giese, M., Steinhage, A., & Schöner, G. (1999). Parametric population representation of retinal location: neuronal interaction dynamics in cat primary visual cortex. *Journal of Neuroscience*, *19*(20), 9016–9028

processes are graded in their state, that is, they are based on activation variables that have continuous values and are defined over continuous feature spaces. This means that both discrete events, such as the detection of an object, as well as the formation and existence of discrete categories, requires explanation. Similarly, all processes evolve in continuous time. This means that processes are not inherently discretized into processing steps and that their formation needs to be explained.

Third, the function of a model is determined by its internal *connectivity* alone. This means that an activation distribution, defined over some feature space, evolves in time only due to input that it receives through connections from other such distributions as well as input through recurrent connections (from itself). These connections are analogous to synaptic connections between individual neurons and can only be changed through slow learning processes. Over short time scales, they can thus be thought of as fixed. This is a constraint for models since different behaviors cannot be explained by rapidly changing connections. It also means that whenever one part of a model should have an influence on another part, they already need to be physically connected. This raises the question how a neural operation can be applied to many different locations in a neural map without requiring unrealistic neural resources. However, even relatively rigid connectivity can still exhibit flexible behavior, in part due to its connection to sensors like the eye that can be directed at many locations in space. Similarly, attentional processes enable focusing on certain aspects over others and allow shifting the representation of an input to a location that a neural operator is connected to (Ballard et al., 1997). Of course, such processes also require an explanation based on neural principles.

Fourth, DFT takes the position of *embodied cognition* seriously (Clark, 1999; Barsalou, 1999; M. Wilson, 2002). Many DFT models are *situated*, that is, they are placed in a real-world environment (e.g., Knips et al., 2017). This means that they have to deal with perception and action simultaneously. DFT models do so in a closed loop: sensory input influences the internal dynamics of the model; this shapes the model's motor output, which in turn may have an influence on the sensor input. All of these processes are tightly coupled to the real time in which the actions unfold. However, the aspect of embodiment that is most relevant to this thesis is the hypothesis that there is no division between primitive cognitive processes at the sensorimotor level and higher cognitive processes. They only differ in their distance to the sensorimotor surfaces. This means that higher cognitive processes are based on the same neural principles as the processes close to the sensorimotor surfaces. In particular, the embodiment position precludes that cognition is based on purely

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–767

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, *3*(9), 345–351; Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609, 577–609; Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–36

Knips, G., Zibner, S. K. U., Reimann, H., Popova, I., & Schöner, G. (2017). A neural dynamics architecture for grasping that integrates perception and movement generation and enables on-line updating. *Frontiers in Neurorobotics*, *11*(9)

abstract symbolic mechanisms (Fodor & Pylyshyn, 1988) that are disconnected from the processes at the sensorimotor level.

Fifth, both processes close to the sensorimotor system as well as representational states about the world require *stability* (Spencer & Schöner, 2003). The sensorimotor system, like the rest of the human nervous system, is inherently noisy. In order to perceive the world and execute actions, cognitive processes must be stable against such noisy fluctuations. Similarly, human cognition works in a broad range of quickly changing, chaotic environments. The representations that we build about the world need to be invariant to all the irrelevant events that happen around us. At the same time, our cognitive processes still need to enable us to continuously update our representations as we become aware of important changes in our surroundings. The notion of stability is deeply ingrained in DFT. Its mathematical framework is built on continuous time differential equations that are in stable attractor states most of the time. While in those states, the model resists noise. Significant changes in input that reflect important changes in the environment or elsewhere in the model may lead to instabilities—points at which the system changes into a different stable state.

These are the neural principles on which DFT is founded. They are expressed as process models in a mathematical framework that I will summarize in the remainder of this section.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71

Spencer, J. P. & Schöner, G. (2003). Bridging the representational gap in the dynamic systems approach to development. *Developmental Science*, *6*(4), 392–412

2.2.2 Dynamic neural fields

The core element of dynamic field theory is a *dynamic neural field* (or simply "field"), an activation distribution $u(x,t)$ that is defined over one or more continuous feature dimensions x, for instance color or space. Figure 2.2 shows a plot of an exemplary dynamic neural field that is defined over a single such feature dimension.

The activation $u(x,t)$ of such a field evolves in time t based on the following differential equation (H. R. Wilson & Cowan, 1973) that was analyzed by Amari (1977)

$$\tau \dot{u}(x,t) = -u(x,t) + h + s(x,t) + w_\xi \cdot \xi(x,t) + \int dx'\, k(x-x')\, g(u(x',t)). \quad (2.1)$$

This equation determines the rate of change $\dot{u}(x,t)$ of the activation $u(x,t)$ at position x and at time t. The change depends on the current activation level $u(x,t)$ of the field, a negative resting level $h < 0$ that brings the activation below a threshold of zero, external input $s(x,t)$ from sensors or other fields, Gaussian white noise $\xi(x,t)$ with strength w_ξ, and a time constant τ that determines the time scale on which the change happens. The last term

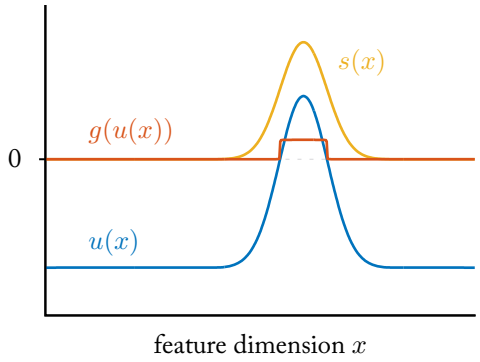


FIGURE 2.2: Dynamic neural field that is defined over a single feature dimension x. This plot shows the activation $u(x)$ of the field (blue line), the localized input $s(x)$ into the field (yellow line), and the sigmoided output $g(u(x))$ of the field (red line).

Wilson, H. R. & Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Biological Cybernetics*, *13*(2), 55–80

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87

of the equation, the integral, determines how different positions in the field interact. Each position in the field has an influence on and is influenced by all other positions in the field, including itself. The interaction has a homogeneous structure throughout the field and depends solely on the distance $\Delta x = x - x'$ between two positions in the field along the feature dimension: positions that are close together excite each other whereas positions that are farther apart inhibit each other. This type of interaction structure draws on neural wiring structures found throughout the human cortex (Jancke et al., 1999). In the dynamics in Equation 2.1 it is formalized by the interaction kernel

$$k(\Delta x) = w_{\text{exc}} \cdot \varphi(\Delta x, \mu, \sigma) - w_{\text{inh}}, \tag{2.2}$$

which consists of an excitatory center, a Gaussian function φ scaled by a positive scalar w_{exc}, and an inhibitory perimeter, determined by w_{inh} (Figure 2.3). The Gaussian function follows the equation

$$\varphi(\Delta x, \mu, \sigma) = a \cdot \exp\left(-\frac{(\Delta x - \mu)^2}{2\sigma^2}\right), \tag{2.3}$$

where μ is the mean value and center of the Gaussian curve, which is usually zero, σ is the standard-deviation around that mean, and a determines the amplitude. When the kernel is defined with a strong inhibition that is effective across the entire feature dimension, a single stable peaks will suppress the formation of additional peaks.

Alternatively, an interaction kernel can be used whose inhibition is only effective over a medium range. This enables the field to form multiple stable peaks. Such an interaction kernel can have the form of a sum of two Gaussian functions

$$k(\Delta x) = w_{\text{exc}} \cdot \varphi(\Delta x, \mu, \sigma_{\text{exc}}) - w_{\text{inh}} \cdot \varphi(\Delta x, \mu, \sigma_{\text{inh}}), \tag{2.4}$$

where the inhibitory part is broader, $\sigma_{\text{inh}} > \sigma_{\text{exc}}$, but smaller in amplitude, $w_{\text{inh}} < w_{\text{exc}}$ (Figure 2.4).

Independently of the interaction kernel, a position in the field can only have an influence on other positions if its activation value is above a threshold u_0, which is also usually zero. In Equation 2.1 this is expressed by the logistic function

$$g(u) = \frac{1}{1 + \exp(-\beta(u - u_0))}, \tag{2.5}$$

where u_0 determines the inflection point of the sigmoid along the input variable and β controls the steepness at this point (Figure 2.5). The logistic function formalizes the output of the field (to other fields as well as to itself). For activation values below the threshold

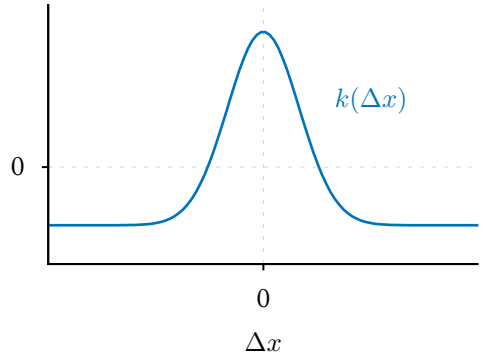


Figure 2.3: Interaction kernel as formalized in Equation 2.2. It is defined over the distance Δx between two positions along the feature dimension x. Values above zero yield excitatory interaction; values below zero create inhibitory interaction. This type of interaction kernel facilitates a selective behavior of a field.

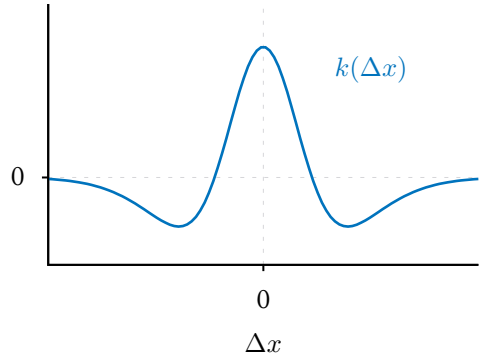


Figure 2.4: An interaction kernel that facilitates the formation of multiple peaks, as formalized in Equation 2.4. It is defined over the distance Δx between two positions along the feature dimension x. Values above and below zero yield excitatory and inhibitory interaction, respectively.

of zero this output function produces zero output, for values above the threshold it produces an output of one; in between, there is a smooth transition.

The first term of Equation 2.1, $-u(x,t)$, creates an attractor at zero. If we disregard all other terms of the equation, the activation will relax toward zero over time, regardless of its initial value. The negative resting level shifts the attractor to h, below the threshold of the output function. This means that without external input $s(x,t)$, the field neither has an influence on other fields, nor does it have any interaction.

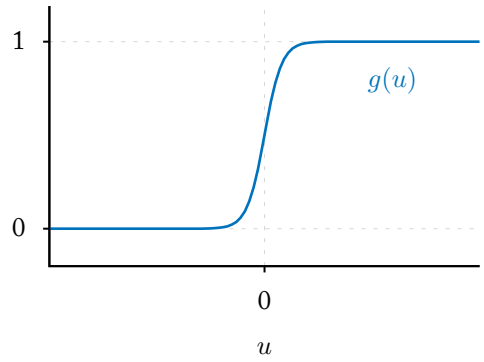


FIGURE 2.5: Exemplary logistic function $g(u)$ following Equation 2.5 with $u_0 = 0$.

Instabilities

Adding external input $s(x,t)$ to a field shifts its attractor, possibly such that the activation rises above the threshold. Most commonly, input is localized along the feature dimension, such that only a region of the field is affected. As soon as a region of the field is pushed above the threshold, interaction within the field becomes active: the positions in the local region of the input mutually excite each other while regions farther away are inhibited. This interaction forms a stable peak of activation above the threshold that is larger than the initial input that brought it about (see Figure 2.2). In doing so, the field goes through the *detection instability*, a bifurcation in which the attractor below the threshold disappears and the system relaxes to an attractor above the threshold. The *reverse detection instability* occurs in the opposite case, when a peak disappears, for instance when there is no longer input to the field.

If multiple localized regions of input appear simultaneously, a field can make a selection decision: it forms a peak at the position of the strongest input and suppresses the other inputs. In doing so, the field goes through the *selection instability*. For a field to be selective, it requires an interaction kernel with strong global inhibition (refer back to Equation 2.2 and Figure 2.3).

If the local excitation of the interaction kernel is strong enough, the field forms a self-sustained peak. That is, the self-excitation is strong enough to keep the peak stable even after the initial input is removed, forming a *memory* of the input. This works both for interaction kernels that facilitate selection as well as for those allowing for multiple peaks.

In dynamic field theory, the processes of detection, selection, and memory are regarded as the most basic cognitive processes. They all emerge from the dynamics of a single neural field. More complex cognitive processes can be modeled by coupling multiple fields into architectures, as will be explained later, in Section 2.2.3.

Higher dimensionalities

The equations and exemplary plots have so far only shown dynamic neural fields defined over a single feature dimension. However, the field equation (Equation 2.1) can be generalized to multiple dimensions:

$$\tau \dot{u}(\vec{x}, t) = -u(\vec{x}, t) + h + s(\vec{x}, t) + w_\xi \cdot \xi(\vec{x}, t) + \int \cdots \int dx'_1 \ldots dx'_n \, k(\vec{x} - \vec{x}') \, g(u(\vec{x}', t)), \quad (2.6)$$

where $\vec{x}$ is a vector of all feature dimensions $x_1, \ldots, x_n$. Figure 2.6 shows a plot of an exemplary two-dimensional field, which could for instance represent the position of an object on a table plane. DFT architectures are commonly limited to low dimensional spaces due to the large amount of neurons and synaptic connections that higher dimensional spaces would require (Schneegans, Lins, & Spencer, 2015).

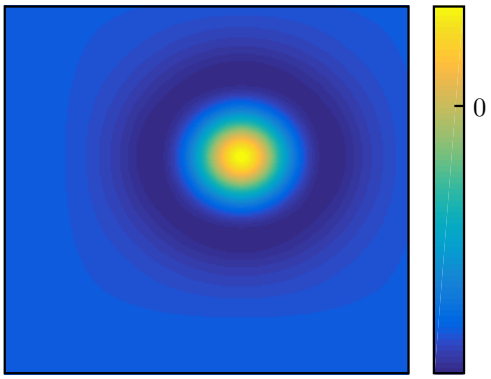


Figure 2.6: Activation of a two-dimensional neural field. The activation is illustrated in a color code (see color bar). The yellow region is a stable peak of activation.

Schneegans, S., Lins, J., & Spencer, J. P. (2015). Integration and selection in multidimensional neural fields. In G. Schöner & J. P. Spencer (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (Chap. 5, pp. 121–150). New York: Oxford University Press

Dynamic neural nodes

A special case of the dynamic neural field is a *dynamic neural node*. It follows the same type of differential equation but is not defined over a continuous dimension; the activation $u(t)$ of such a node is only a scalar value. For this case, Equation 2.1 reduces to

$$\tau \dot{u}(t) = -u(t) + h + w_{\text{se}} \cdot g(u(t)) + s(t) + w_\xi \cdot \xi(t). \quad (2.7)$$

It uses the same output function $g(u(t))$ (Equation 2.5) as the field and can thus represent only two states, the "off" state, where it produces output of around zero and the "on" state where it produces an output of roughly one. The interaction kernel is replaced by a single recurrent connection, the *self excitation*, which projects the output $g(u(t))$ of the node onto itself, weighted by a factor $w_{\text{se}} \in \mathbb{R}_+$ (Figure 2.7). This recurrent connection makes the node bistable and creates behavior analogous to what we have seen in the case of a field. The node goes through the detection instability when presented with sufficiently strong input and it will go through the reverse detection instability if that input is removed. If the self-excitation is sufficiently strong, the node exhibits the same memory property as the field. Since a single node can only represent the state of a single entity, it cannot exhibit selection by itself. However, selective behavior can be implemented by coupling multiple nodes.

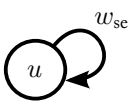


Figure 2.7: Neural node (circle) with activation variable u. The arrow denotes self-excitation.

2.2.3 Architectures

We have seen that the primitive cognitive processes of detection, selection, and memory can be modeled with a single dynamic neural field. More complex cognitive processes may be explained by combining fields and nodes into interconnected architectures. A simple example of such an architecture are two dynamic neural nodes with activation variables u_1 and u_2 that are connected such that node 1 excites node 2 and node 2 inhibits node 1 (Figure 2.8). The equations for both nodes follow the general form of Equation 2.7 with the following external inputs s_1, s_2 for nodes 1 and 2, respectively

$$s_1(t) = w_{1,2} \cdot g(u_2(t)), \tag{2.8}$$
$$s_2(t) = w_{2,1} \cdot g(u_1(t)). \tag{2.9}$$

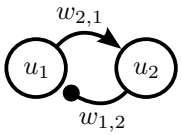


FIGURE 2.8: Neural oscillator consisting of two neural nodes with activation variables u_1 and u_2.

The excitatory connection from node 1 to node 2 is formalized by $w_{2,1} > 0$, while the inhibitory connection from node 2 to node 1 is given by $w_{1,2} < 0$. Incidentally, this simple model produces a rhythmic pattern of activation; it is an oscillator (Amari, 1977).

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87

Coupling dynamic neural fields to model their interaction is done analogously. However, not all fields are of the same dimensionality. And even if they are, they may not be defined over the same feature dimensions. In the following, I will formalize three different coupling schemes,[5] in all cases referring to a source field A with activation $u_\mathrm{A}(\vec{x}, t)$ of dimensionality a and a target field B with activation $u_\mathrm{B}(\vec{x}, t)$ of dimensionality b. The activation of both fields evolves in time based on dynamics analogous to Equation 2.6. I will further assume that the dimensions of the two fields are aligned, such that they are defined over the same vector $\vec{x}$.

When the two fields have the same dimensionality ($a = b$), the coupling is a *one-to-one coupling*. In this case, the input $s_\mathrm{B,A}$ from field A to field B is determined by

$$s_\mathrm{B,A}(\vec{x}, t) = g(u_\mathrm{A}(\vec{x}, t)). \tag{2.10}$$

When field A is defined over less metric dimensions than field B ($a < b$), the coupling is an *expansion*. In this case, the vector $\vec{x}_\mathrm{B}$, which describes the dimensions that field B is defined over, contains all of the entries of $\vec{x}_\mathrm{A}$, the dimensions of field A, as well as some additional entries. For such a coupling the input to the target field is

$$s_\mathrm{B,A}(\vec{x}_\mathrm{B}, t) = g(u_\mathrm{A}(\vec{x}_\mathrm{A}, t)). \tag{2.11}$$

The input is constant within the additional dimensions of field B, which leads to characteristic shapes in the input. For instance, if

[5] The description of these coupling schemes is based on published material (Lomp, Richter, Zibner, & Schöner, 2016), which arose as a collaboration between Oliver Lomp (OL), Mathis Richter (MR), Stephan Zibner (SZ), and Gregor Schöner (GS). MR and SZ implemented the exemplary model shown in the paper, MR produced results. OL, SZ, and MR developed the software described in this paper. All authors participated in writing the paper.

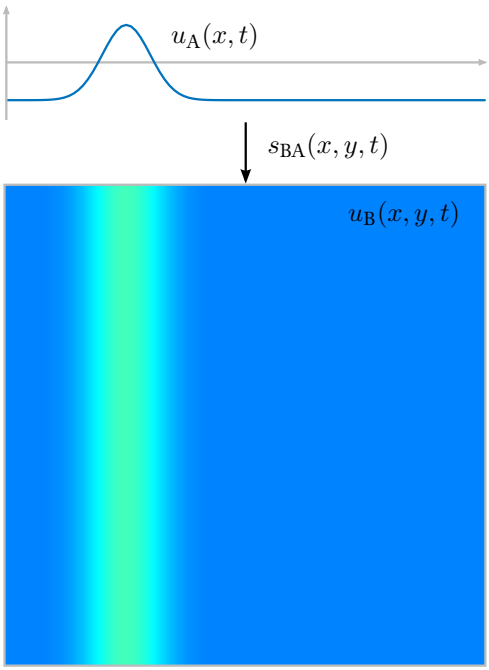


FIGURE 2.9: Expansion coupling from a one-dimensional field to a two-dimensional field.

the source field is one-dimensional and the target field is two-dimensional, the input forms a ridge (Figure 2.9); if the source is one-dimensional and the target is three-dimensional, the input forms a cylinder.

When the source field A is defined over more metric dimensions than the target field B ($a > b$), the coupling is a *contraction*. I assume here that the extra dimensions that are represented in A but not in B are the last $(a - b)$ entries $x_{b+1}, \ldots, x_a$ of $\vec{x}_A$. One way to contract these dimensions is to integrate them

$$s_{B,A}(\vec{x}_B, t) = \int \cdots \int dx_{b+1} \ldots dx_a \, g(u_A(\vec{x}_A, t)), \tag{2.12}$$

which has the disadvantage that the input to the receiving field varies in strength depending on the number of objects that are represented along the contracted dimension. Parameterizing the receiving field may thus require some form of normalization of the input strength. To simplify the process of parameterizing the model introduced in this thesis, I use the maximum function to contract dimensions

$$\max_{x_i,\ldots,x_j} (f(\vec{x})). \tag{2.13}$$

It takes the maximum over the dimensions $x_i, \ldots, x_j$ and projects it onto the remaining dimensions. The function thus reduces the dimensionality of the input by the number of contracted dimensions. Using the maximum function, the exemplary contraction coupling from field A to field B would be

$$s_{B,A}(\vec{x}_B, t) = \max_{x_{b+1},\ldots,x_a} (g(u_A(\vec{x}_A, t))), \tag{2.14}$$

where the maximum is taken over the dimensions $x_{b+1}, \ldots, x_a$. Although uncommon in DFT models, the maximum function is an essential part of the "standard" neural model of object recognition, the HMAX model (Riesenhuber & Poggio, 1999). It is employed there in the early stages of visual processing, where the responses of simple cells are pooled to give input to complex cells. Compared to a more traditional pooling by summation of the incoming synaptic connections, the maximum function provides a more robust response when the visual input is cluttered. Furthermore, the maximum function is more invariant to changes in size of the stimulus, since it only projects the best-matching response; this is similar to the reason it is used in this thesis. The maximum function can be approximated by the *softmax* function

$$w_i = \sum_j \frac{\exp(p \cdot |s_j|)}{\sum_k \exp(p \cdot |s_k|)} s_j, \tag{2.15}$$

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–25

which pools the responses of simple cells s_j to produce a response w_i of a complex cell. The strength of the non-linearity is determined by the parameter p. For $p = 0$ the softmax function yields a linear sum of the responses of all simple cells, whereas for $p \to \infty$ it approximates the maximum function (Riesenhuber & Poggio, 1999).

In all three coupling schemes, one-to-one, expansion, and contraction, the output of field A is commonly also convolved with a Gaussian kernel $k_{\mathrm{B,A}}(\vec{x})$, a point-spread function. For instance, expressing the one-to-one coupling in this way results in

$$s_{\mathrm{B,A}}(\vec{x}, t) = \int d\vec{x}\; k_{\mathrm{B,A}}(\vec{x} - \vec{x}')\, g(u_{\mathrm{A}}(\vec{x}', t)). \tag{2.16}$$

The equations for the other coupling schemes have an analogous form.

Since equations with many couplings that each contain a convolution quickly become long, I introduce the following notational shorthand for convolutions

$$[k * g(u)](\vec{x}, t) = \int_{R^n} d\vec{x}'\, k(\vec{x} - \vec{x}')\, g(u(\vec{x}', t)) \tag{2.17}$$

and use it where it helps to shorten equations throughout the rest of the thesis. Please note that the kernel and the output of the field are only convolved along the dimensions $\vec{x}$, not over time t. Only the activation $u(\vec{x}, t)$ is dependent on time, the kernel $k(\vec{x})$ is not.

By default, the synaptic weights between fields are homogeneous and act either excitatorily or inhibitorily upon the target field. In some cases we employ synaptic weights that have a pattern, for instance in the connection between certain nodes and fields. This will be further explored in Section 3.4.

2.2.4 Motion perception

In order to perceive objects in the world, DFT architectures are connected to sensors, most importantly to cameras. Based on their input, the features of objects can be represented in dynamic neural fields. The motion of objects in a scene can be detected and their movement direction estimated through a neural-dynamic version (Berger et al., 2012) of the counterchange model for motion perception (Hock et al., 2009). The counterchange model of motion perception asserts that we perceive motion when an intensity change (e.g., in luminance) at one location coincides with an inverse intensity change at a nearby location. This model is able to explain the perception of both real motion as well as apparent motion. It is built around arrays of transient detectors that react to local intensity changes.

Berger, M., Faubel, C., Norman, J., Hock, H., & Schöner, G. (2012). The counter-change model of motion perception: An account based on dynamic field theory. In A. E. P. Villa (Ed.), *ICANN 2012, Part I, LNCS 7552* (pp. 579–586). Berlin Heidelberg: Springer

Hock, H. S., Schöner, G., & Gilroy, L. (2009). A counterchange mechanism for the perception of motion. *Acta Psychologica*, *132*(1), 1–21

A transient detector can be modeled by a dynamic variable u, which interacts with a dynamic variable v according to the following dynamics

$$\tau_u \dot{u}(t) = -u(t) + w \cdot s(t) - v(t), \tag{2.18}$$
$$\tau_v \dot{v}(t) = -v(t) + w \cdot s(t). \tag{2.19}$$

Both equations have a stabilizing term ($-u$ and $-v$, respectively), which, in absence of other input, creates an attractor at zero. The dynamics continuously receive input $s(t)$ (e.g., luminance intensity) with strength $w > 0$, which shifts the attractor to the input and makes the dynamics follow it. Since the second equation is inhibitorily coupled to the first (through the $-v(t)$ term), it cancels that input in the first equation. If both equations used the same time constant, the first dynamics would always have an attractor at zero while the second followed the input. However, if the time constants are set such that $\tau_v > \tau_u$, the variable v relaxes slower than u and allows it to follow the input for some time before canceling it out. This makes the dynamics act as a transient detector: for every change in the input s, the activation u will follow the change for a short period of time and then return to zero until the input changes again (Figure 2.10).

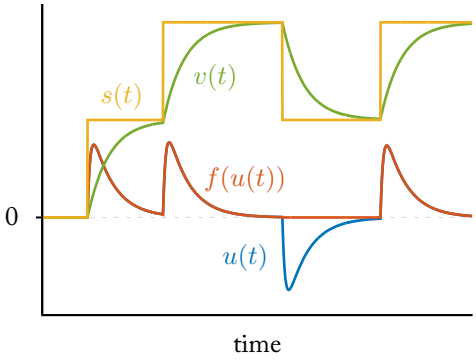


Figure 2.10: Transient detector

Equations 2.18 and 2.19 define a transient detector that reacts to any change in the input, be it negative or positive. These two types of change can be distinguished by defining two different detectors that only react to either the positive or the negative changes in the input. Positive changes can be detected by using the semi-linear rectifier function $f(u) = \max(0, u)$ as output of the transient detector. Negative changes can be detected if, in addition, we choose a constant $w < 0$ as strength for the input of the detector. Both detectors yield a positive response whenever they detect change.

To implement the counterchange model of motion perception, transient detectors for positive and negative change at different positions have to be compared. This requires that two arrays of transient detectors are defined that span every spatial position $\vec{x}$; one array, $u_p(\vec{x}, t)$ reacts to positive changes, the other, $u_n(\vec{x}, t)$ to negative changes. One can then compare a pair of detectors, one positive and one negative, at different spatial positions $\vec{x}_0$ and $\vec{x}_1$

$$m = f(u_p(\vec{x}_0, t)) \cdot f(u_n(\vec{x}_1, t)). \tag{2.20}$$

If a change is detected by both detectors, m is close to one, otherwise it is close to zero. The direction of movement ϕ is determined by the angle of the vector $\vec{x}_0 - \vec{x}_1$.

These equations can be generalized to compare the signals of all combinations of transient detectors, from which the movement

direction of all objects in the scene can be deduced. This is explored further as part of the model in Section 3.1.2.

2.2.5 Steerable neural mappings

A perceptual front-end like the one described above enables architectures to form representations of the spatial position of objects. When fed by camera input, these representations are in camera coordinates, that is, their reference frame is determined by the position and orientation of the camera. *Steerable neural mappings* enable us to flexibly change the reference frame of a representation. That is, they yield a representation in a field in which the peak positions are shifted, centering them on an arbitrary reference position that can be specified. This mechanism can for instance explain how a retinocentric representation that changes with every saccade can be mapped onto a more stationary allocentric representation (Schneegans & Schöner, 2012). As explored in the next section, it is also central to explaining how spatial relations between objects can be extracted.

Schneegans, S. & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109

To demonstrate the mechanism here, I use an example where the representation is defined over a one-dimensional feature space (Figure 2.11). Please note that I am choosing a one-dimensional space for clarity of illustration only; the method can be extended to higher dimensional spaces.

Let us assume that we are representing the position of objects in a one-dimensional space x within a neural field A. The activation $u_\mathrm{A}(x,t)$ of that field follows the generic field equation (Equation 2.1)

$$\begin{aligned}\tau \dot{u}_\mathrm{A}(x,t) = &- u_\mathrm{A}(x,t) + h_\mathrm{A} \\ &+ [k_\mathrm{A,A} * g(u_\mathrm{A})](x,t) \\ &+ s_\mathrm{A}(x,t),\end{aligned} \tag{2.21}$$

where the external input $s_\mathrm{A}(x,t)$ consists of sensory input that leads to the formation of two peaks in the field. The location of peaks reflects the position of two objects in space, relative to some reference frame.

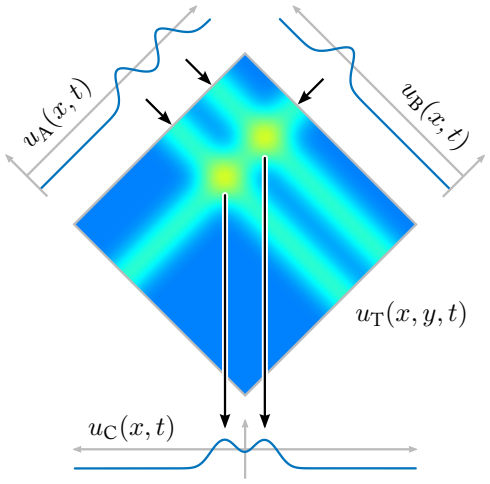


FIGURE 2.11: Steerable neural mapping

Let us assume further that we have another field B that holds a peak at some reference position. The field is defined over the same space x and its activation $u_\mathrm{B}(x,t)$ follows the dynamics

$$\begin{aligned}\tau \dot{u}_\mathrm{B}(x,t) = &- u_\mathrm{B}(x,t) + h_\mathrm{B} \\ &+ [k_\mathrm{B,B} * g(u_\mathrm{B})](x,t) \\ &+ s_\mathrm{B}(x,t).\end{aligned} \tag{2.22}$$

The external input $s_{\mathrm{B}}(x,t)$ leads to the formation of a peak at the reference position.

We can then define a transformation field T, which produces a representation of the objects (the ones represented in field A) that is centered on that reference position. The transformation field T expresses a mapping from the original feature space onto itself. It thus has twice the number of dimensions as the original feature space. In our example, its activation $u_{\mathrm{T}}(x,y,t)$ evolves in time based on the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{T}}(x,y,t) = &- u_{\mathrm{T}}(x,y,t) + h_{\mathrm{T}} \\ &+ [k_{\mathrm{T,T}} * g(u_{\mathrm{T}})](x,y,t) \\ &+ [k_{\mathrm{T,A}} * g(u_{\mathrm{A}})](x,t) \\ &+ [k_{\mathrm{T,B}} * g(u_{\mathrm{B}})](y,t).\end{aligned} \tag{2.23}$$

Along x, it receives input from field A, along y it receives input from field B. These inputs are projected onto the two-dimensional space of T, leading to ridges of activation. At the position of that crossing, the transformation field will form a peak (Figure 2.11).

The output of the transformation field feeds into a field C, which holds the representation of the shifted object positions. The transformation happens through a special coupling between the transformation field T and field C, in which the activation of the transformation field is "read out diagonally". The activation u_{C} of field C has the dynamics

$$\begin{aligned}\tau \dot{u}_{\mathrm{C}}(x,t) = &- u_{\mathrm{C}}(x,t) + h_{\mathrm{C}} \\ &+ [k_{\mathrm{C,C}} * g(u_{\mathrm{C}})](x,t) \\ &+ s_{\mathrm{C,T}}(x,t)\end{aligned} \tag{2.24}$$

where the input from the transformation field T is given by

$$s_{\mathrm{C,T}}(x,t) = [k_{\mathrm{C,T}} * G_{\mathrm{C}}(u_{\mathrm{T}})](x,t), \tag{2.25}$$

with the diagonal read-out

$$G_{\mathrm{C}}(u_{\mathrm{T}})(x,t) = \int dp\, g(u_{\mathrm{T}}(x-p,p,t)). \tag{2.26}$$

With this transformation, the positions of the peaks in field C reflect their position relative to the reference position held by field B.

As will be explained later (Section 3.3), in my thesis model I use transformations on two-dimensional representations. This requires four-dimensional transformation fields, which are a serious performance bottleneck when simulated on a computer. In the implementation of the model, I thus opted for approximating steerable

neural mappings by convolutions, as their computation can be optimized for performance by using the fast Fourier transform. Using a convolution to replace the transformation field T, the input $s_{\mathrm{C}}(x,t)$ from the spatial transformation into field C would be

$$s_{\mathrm{C}}(x,t) = \int dx'\, g(u_{\mathrm{B}}(x-x'))\, g(u_{\mathrm{A}}(x')), \tag{2.27}$$

replacing $s_{\mathrm{C,T}}$ in the third line of Equation 2.24. Even though the model uses convolutions for spatial transformations, it remains neurally plausible since all convolutions can be mapped onto steerable neural mappings.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511

2.2.6 Spatial language model

Steerable neural mappings are most prominently employed in the DFT model of spatial language. This line of research aims at explaining the cognitive processes that are required to resolve spatial relations between objects in a scene, for instance following a question such as: "What is to the left of the cup?" The most extensive version of the DFT model of spatial language is reported by (Lipinski et al., 2012) and lays the foundation for the work presented in this thesis. This section covers both the structure of the model as well as some of the tasks it is able to capture. I will forgo a mathematical description of the model, as this would go beyond the scope of this thesis. While Lipinski et al. (2012) do not offer a mathematical description either, they do for a previous version (Lipinski, Sandamirskaya, & Schöner, 2009).

Model

An overview diagram of the model of spatial language is shown in Figure 2.12. The model receives real visual input from a camera, which feeds into a set of three *color-space fields* that are all defined over the two-dimensional image space of the camera. The three fields represent the spatial position of objects in the camera image, where each of the fields represents only objects of a single color: red, green, and blue, respectively.[6] Each field is connected reciprocally to a neural *color term node*. The nodes represent the colors RED, GREEN, and BLUE. The connections between each node and its field are homogeneous across the entire field. If the node is active, it brings the field into a dynamic range where it can form peaks from the localized input it receives from the camera. The field can also form a peak without the node being active if the user manually raises the resting level of the field. When a peak forms, this activates the node.

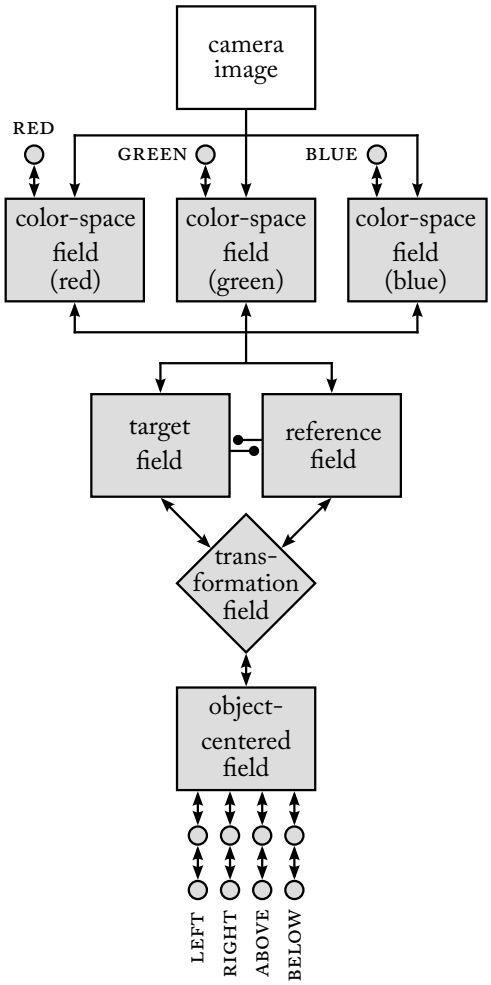


FIGURE 2.12: Overview of the DFT model for spatial language. Dynamic neural nodes are depicted as gray circles, dynamic neural fields are shown as gray rectangles. Diagram adapted from Lipinski et al. (2012).

All three fields are coupled reciprocally to the *target field* and the *reference field*. Both of these fields are also defined over the two-dimensional image space of the camera. The target field holds a representation of the spatial position of the target object, the object that is being referred to in a phrase such as "the red object to the left of the green object".[7] Analogously, the reference field holds a representation of the spatial position of the reference object, the object used as a reference in order to point out the target object (the green object in the example above). The target field and the reference field are reciprocally coupled with inhibitory connections. This ensures that an object is never represented as both the target object and the reference object.

The target field and the reference field are both reciprocally coupled to the *transformation field* with excitatory connections. The transformation field is a steerable neural mapping[8] that is defined over all combinations of spatial positions of both the target field and the reference field, both of which are two-dimensional; it is thus four-dimensional.

It is reciprocally connected to the *object-centered field*, which is defined over a two-dimensional space that is centered on the spatial position of the reference object.[9] The output of the transformation field that feeds into the object-centered field represents the spatial position of the target object relative to that of the reference object.

The object-centered field is reciprocally connected to four *spatial relation nodes*, which represent the spatial relations TO THE LEFT OF, TO THE RIGHT OF, ABOVE, and BELOW. Each of these relations is encoded in patterned synaptic weights between a spatial relation node and the object-centered field. For instance, the pattern for the relation TO THE LEFT OF has excitatory connections on the left side of the field and neutral connections on the right side of the field. The excitatory strength is highest to the left of the center of the field and diminishes both with increasing distance from the center as well as increasing angular distance from true left. Multiple spatial relation nodes can be active at the same time. In addition, there is a *spatial term node* for each of the spatial relation nodes, forming pairs that are reciprocally coupled with excitatory connections. Inhibitory connections between all spatial term nodes ensure that only one such node can be active at any given time. This enables the model to make a selection decision about the spatial relation it perceives between the target object and the reference object.

Demonstrations

Lipinski et al. (2012) show that the model can exhibit flexible spatial language behaviors. Different types of questions can be posed

[6] As a set, the three color-space fields approximate a three-dimensional color-space field that is defined over the two-dimensional image space as well as the color dimension (where the color dimension has only a coarse resolution).

[7] In this example, the red object is the target object and the green object is the reference object.

[8] See Section 2.2.5: the transformation field defined here corresponds to the field with activation u_T defined there. The target field and reference field correspond to the fields A and B, respectively.

[9] The object-centered field defined here corresponds to field C in Section 2.2.5.

to the model in the form of inputs to various parts of the model. For instance, the user can give an input that corresponds to the question "Where is the green flashlight relative to the red tape dispenser?" in a visual scene that contains both of these objects as well as a blue box cutter (the flashlight is to the right of the tape dispenser). Over the course of the demonstration, the model continuously receives input from the camera. The user manually gives input to the model, activating the color term node that represents GREEN while at the same time raising the resting level of the target field. The model forms a peak in the color-space field for the color GREEN as well as in the target field. The user removes the input from the node, which leads to the decay of the peak in the color-space field. He also removes input from the target field, which lowers its resting level. However, this peak remains due to large local excitatory interaction within the target field. The peak is a working memory representation of the spatial position of the target object, the green flashlight. To build up a representation of the reference object, the user analogously activates the color term node that represents RED while at the same time raising the resting level of the reference field. After a peak has formed in the reference field at the position of the red tape dispenser, the user removes both inputs. As for the target field, the peak in the reference field remains due to large local excitatory interaction within the field. With activation both in the target field and the reference field, the transformation field forms a peak, projecting its activation into the object-centered field. Since the green flashlight (target) is to the right of the red tape dispenser (reference) the peak appears to the right of the center of the field. This overlaps most with the patterned synaptic connections to the spatial relation node for the relation TO THE RIGHT OF and activates this node. To force a definitive answer to the question, the user gives a last input into the model, homogeneously raising the resting level of all spatial term nodes. Since the spatial term node for the relation TO THE RIGHT OF also receives input from its corresponding spatial relation node, it becomes active. This activation represents the answer of the model to the question it was given.

Similarly, one can ask a question such as "Which object is above the blue deodorant stick?" by giving input to both the color term node for the color BLUE and raising the resting level of the reference field as well as giving input to the spatial term node for the relation ABOVE and raising the resting level of the object-centered field. The model finds the correct target, a red box cutter that is above the blue deodorant stick, and forms a peak at the spatial position of the box cutter in the target field once the user raises its resting level. When the user raises the resting level of all color term nodes, the node for the color RED activates and represents the answer to the question.

In a similar manner, one can ask a question such as "Where is the green highlighter?" and the model finds both a reference object as well as a fitting spatial relation.

Additionally to demonstrating that the model can exhibit flexible spatial language behaviors, Lipinski et al. (2012) show that it captures empirical data from behavioral experiments. In particular, they capture data by Regier and Carlson (2001) by showing that the model's rating of spatial relations is influenced by two distinct measures of orientation, the *proximal orientation*—the orientation of a vector pointing toward the target object from the closest point of the reference object—and the *center-of-mass orientation*—the orientation of a similar vector that originates in the center-of-mass of the reference object. Moreover, the model captures data reported by Carlson and Hill (2008), which indicates that while selecting a reference object, humans are influenced less by its saliency and rather by its alignment with spatial relations relative to the target object.

Regier, T. & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, *130*(2), 273–298

Carlson, L. A. & Hill, P. L. (2008). Processing the presence, placement, and properties of a distractor in spatial language tasks. *Memory & Cognition*, *36*(2), 240–255

2.2.7 Behavioral organization

One of the limitations of the DFT model of spatial language described in the last section is that it depends on input from a human user at different points in time in order to initiate instabilities and select different behaviors. Modeling architectures that can generate *multiple behaviors*, *autonomously* without help from a user, and activating some behaviors in a certain *sequential order* requires the controlled activation and deactivation of the architecture's functional parts at critical moments in time. This section explains how such a *behavioral organization* can be realized within DFT.

Elementary behavior

Larger DFT architectures can most often be subdivided into elementary behaviors (EBs) that comprise independent functional parts of the architecture. An EB can be any cognitive process, including movements executed by the motor system, perceptual acts driven by the sensory system, or cognitive processes, such as memorizing the position of an object in space.

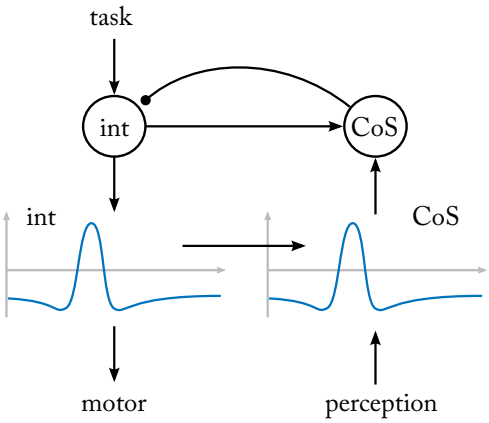


Figure 2.13: Elementary behavior

In DFT, each EB is modeled by a structure of nodes and fields (Figure 2.13) that implements two concepts: the *intention (int)* represents whether the EB is active and what its effect is going to be; the *condition of satisfaction (CoS)* checks whether the EB is successfully completed and shuts it off. For each of these two concepts, we employ both a node and a field (Richter et al., 2012). Whether the EB is active or inactive is represented by the "on" and "off" state of the *intention node*. The effect the EB has on the architecture

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2457–2464). New York, NY: Institute of Electrical and Electronics Engineers (IEEE)

is determined by the *intention field*, which is defined over feature dimensions relevant for the EB. The intention node is excitatorily connected to the intention field, such that a peak forms in the field when the node is activated by some task input. The position of the peak can be determined by patterned connection weights from the node to the field or by localized input from another source into the field. In the latter case, the connection weights from the node to the field are homogeneous. A peak in the intention field can have various effects on the rest of the architecture, from simply creating a peak in another field to having a direct impact on the motor system.

To determine the moment at which the EB is successfully completed, the *CoS field* matches the intended end-state of the EB, encoded in the connections from the intention field, with current sensory input (Sandamirskaya & Schöner, 2010). Only if the input from these two sources coincide does the condition of satisfaction field form a peak. This activates the *CoS node*, which inhibits the intention node, turning it off. When the intention node is turned off, the peak in the intention field decays as well, turning off any effect the EB may have had on the architecture—the EB is inactive. For some EBs, it may be necessary to preserve the information that the EB has been successfully completed. This can be achieved by increasing the self-excitation of the CoS node, making it self-sustained. Once activated, the CoS node will not turn off unless actively inhibited.

Sandamirskaya, Y. & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179

The following set of coupled differential equations formalizes the EB shown in Figure 2.13, where the activation variables are as follows: intention node u_{IN}; CoS node u_{CN}; intention field u_{IF}; CoS field u_{CF}. Please note that this is just an example and EBs in larger architectures may vary in structure depending on the kind of processes they control. For instance, it is assumed here that both the intention field and the CoS field are defined over the same one-dimensional feature space x; this does not have to be the case for every EB. Thus, the following equations serve as an exemplary guideline rather than a generic formalization.

The intention node with the activation variable u_{IN} evolves in time based on the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{IN}}(t) = &- u_{\text{IN}}(t) + h \\ &+ w_{\text{IN,IN}}\, g(u_{\text{IN}}(t)) \\ &- w_{\text{IN,CN}}\, g(u_{\text{CN}}(t)) \\ &+ s_{\text{IN,T}}(t),\end{aligned} \tag{2.28}$$

Both τ and $h < 0$ can have the same values for all equations of the EB. This is not required, but it simplifies finding an appropriate set of parameters.

where the first two lines correspond to the general equation (Equation 2.7) for a dynamic neural node, the second line being the self-excitation. The third line formalizes the inhibitory input from the

CoS node to the intention node. The fourth line denotes input from a task variable, which could be from another neural node or from user input. The exact formalization is left open here because it depends on how the EB is integrated into the rest of the architecture.

The CoS node with the activation variable u_{CN} is governed by the equation

$$\begin{aligned}\tau \dot{u}_{\text{CN}}(t) = &- u_{\text{CN}}(t) + h \\ &+ w_{\text{CN,CN}}\, g(u_{\text{CN}}(t)) \\ &+ w_{\text{CN,IN}}\, g(u_{\text{IN}}(t)) \\ &+ \max_x([k_{\text{CN,CF}} * g(u_{\text{CF}})](x,t)).\end{aligned} \tag{2.29}$$

As above, the first two lines correspond to the general form for a dynamic neural node; the second line is the self-excitation of the node. The third line is the excitatory connection from the intention node to the CoS node. The fourth line formalizes the contraction coupling from the CoS field to the CoS node, in which the output $g(u_{\text{CF}}(x,t))$ of the CoS field is convolved with a kernel $k_{\text{CN,CF}}(x)$ and the feature dimension x is contracted.

The intention field with the activation variable u_{IF} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{IF}}(x,t) = &- u_{\text{IF}}(x,t) + h \\ &+ [k_{\text{IF,IF}} * g(u_{\text{IF}})](x,t) \\ &+ w_{\text{IF,IN}}(x) \cdot g(u_{\text{IN}}(t)),\end{aligned} \tag{2.30}$$

where the first two lines correspond to the general equation for a dynamic neural field (Equation 2.1) defined over a one-dimensional feature space x. The second line is the lateral interaction in the field. The third line is the input from the intention node to the intention field. In this coupling, the synaptic weights $w_{\text{IF,IN}}$ can either be homogeneous or have a pattern that depends on the feature space x, as shown here. In the latter case, the weights have an influence on the position of the peak in the intention field.

The CoS field with the activation variable u_{CF} is governed by the equation

$$\begin{aligned}\tau \dot{u}_{\text{CF}}(x,t) = &- u_{\text{CF}}(x,t) + h \\ &+ [k_{\text{CF,CF}} * g(u_{\text{CF}})](x,t) \\ &+ [k_{\text{CF,IF}} * g(u_{\text{IF}})](x,t) \\ &+ s_{\text{CF,P}}(x,t).\end{aligned} \tag{2.31}$$

As above, the first two lines correspond to the general form for a dynamic neural field, defined here over the same feature space x

as the intention field. The second line is the lateral interaction in the field. The third line formalizes the connection from the intention field to the CoS field, in this example a one-to-one coupling. The fourth line denotes perceptual input from the sensory system. Only if this input overlaps with the input from the intention field does the CoS field form a peak. This can be modeled by balancing the strength of the kernel $k_{\mathrm{CF,IF}}$ and the strength of perceptual input $s_{\mathrm{CF,P}}$.

Sequential constraints

Elementary behaviors create a more abstract layer of control within a DFT architecture. Activating and deactivating the intention nodes of different EBs may evoke qualitatively different overall behaviors from the architecture. In some cases, EBs may be active simultaneously but often they may not. In fact, meaningful overall behavior often consists of sequences of smaller actions. To organize the activation of multiple EBs in time, we activate all relevant EBs at the same time but introduce *sequential constraints* that lead to a sequential execution of EBs (Sandamirskaya et al., 2011).

For two behaviors, EB1 and EB2, the *precondition constraint* expresses that EB1 has to be successfully completed before EB2 can be activated. In DFT, the constraint is represented by a *precondition node*, which is activated together with both EBs and inhibits the intention node of EB2 (Figure 2.14). The precondition node is in turn inhibited by the CoS node of EB1. As soon as EB1 has reached its condition of satisfaction and is completed, the precondition node is turned off, releasing inhibition from EB2. This leads to the sequential activation of EB1 and EB2.

If what matters is not a particular sequential order of EB1 and EB2 but only that they are not active at the same time, this can be expressed by a *suppression constraint*. It is represented by a *suppression node*, which inhibits the intention node of EB2 but is only activated while the intention node of EB1 is active (Figure 2.15). Set up in this way, EB1 will always suppress EB2 when it becomes active. Adding an additional suppression node in the other direction (inhibiting the intention node of EB1) will create competition between the behaviors; whichever EB becomes active first will suppress the other.

2.2.8 Numerical implementation

DFT models explain how cognitive processes unfold in continuous time based on continuous representations of feature spaces. To show what these processes look like, how they are influenced by sensory

Sandamirskaya, Y., Richter, M., & Schöner, G. (2011). A neural-dynamic architecture for behavioral organization of an embodied agent. In *IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL EPIROB 2011)* (pp. 1–7). IEEE

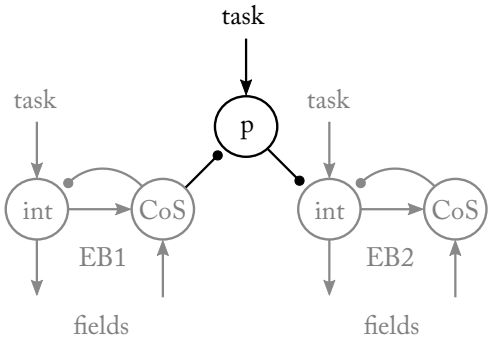


Figure 2.14: Precondition constraint. The precondition node is labeled "p".

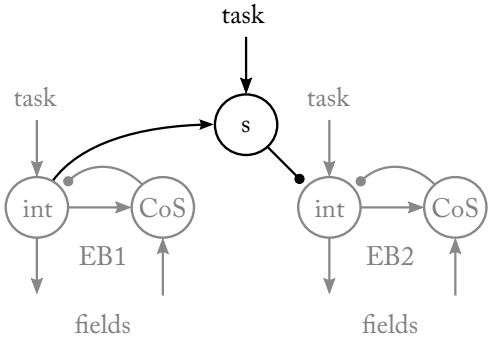


Figure 2.15: Suppression constraint. The suppression node is labeled "s".

input and how they can control motor output, the models can be simulated on computers. With the coupling to sensory input and the stochastic properties inherent to the models, analytic solutions of the underlying dynamics cannot be determined. Instead, the dynamics are solved numerically. A stochastic differential equation like the neural field equation (Equation 2.1) of the form

$$\tau \dot{u} = f(u) + w_\xi \cdot \xi(t), \tag{2.32}$$

with a deterministic term $f(u)$ and a stochastic term $\xi(t)$ can be solved numerically with the stochastic forward Euler method (Ascher & Petzold, 1998)

$$u_i = u_{i-1} + \frac{1}{\tau}\left(\Delta t_i \, f(u_{i-1}) + \sqrt{\Delta t_i}\, w_\xi \, \xi_{i-1}\right). \tag{2.33}$$

Given an initial activation value $u_0 = c$, this yields an approximation u_i of the continuous time activation $u(t)$ at discrete time points t_i ($i = 1, 2, 3, \ldots$) (see Figure 2.16 for an example). The sampling of the time points is approximately equidistant with $\Delta t_i = t_i - t_{i-1}$. Choosing Δt_i is a trade-off between precision and performance; smaller values make the approximation more precise but also lead to a higher computational load.

Building larger DFT models requires that many such simulations of fields and nodes are instantiated, coupled among each other, and possibly connected to artificial or real sensors and motors. A major effort in building models is finding a suitable parameter set for all dynamics that brings all fields into the desired dynamic regimes. For larger DFT models, this requires close monitoring of the activation of multiple fields and nodes, while at the same time systematically changing parameters as well as varying sensory input. The software framework *cedar*,[10] which I co-developed, enables users to build, parameterize, simulate, analyze, and document DFT architectures using a graphical user interface. The model introduced in the next section was developed and simulated using this software framework. Please refer to Lomp et al. (2016) for further details on *cedar* and the numerical implementation of DFT models.

Ascher, U. M. & Petzold, L. R. (1998). *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics

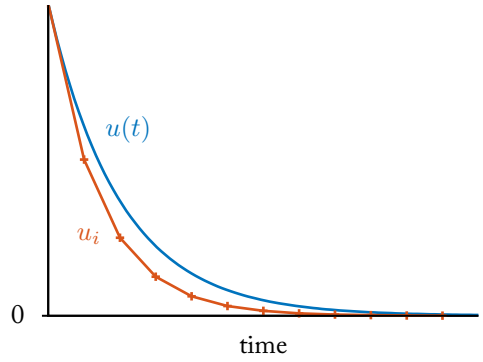


Figure 2.16: Discrete time Euler approximation u_i (red line, $i = 0, 1, 2, \ldots$) of a continuous time activation $u(t)$ (blue line) that is governed by the differential equation $\dot{u}(t) = -u$, which relaxes to zero.

[10] *cedar* is an open-source C++ library that is freely available under the LGPL license (version 3). The source code and documentation can be accessed at http://cedar.ini.rub.de.

Lomp, O., Richter, M., Zibner, S. K. U., & Schöner, G. (2016). Developing dynamic field theory architectures for embodied cognitive systems with cedar. *Frontiers in Neurorobotics*, *10*, 1–18

Model 3

This chapter introduces a neural dynamic model for the perceptual grounding of spatial and movement relations. It can solve two related tasks: first, given real camera input of a scene with colored objects, it can *ground* a spatial phrase like "the red object to the left of the green object" by directing attentional focus to the corresponding object in the scene. Second, given just the camera input, it can *generate a description* of the scene in the form of a spatial phrase like the one above. In doing so, the model extracts the relevant features of objects as well as their spatial relationship. The model is able to solve these two tasks for stationary scenes, where the relations between objects correspond to relative spatial positions like TO THE LEFT OF or ABOVE, as well as for dynamic scenes that feature movement relations like TOWARD or AWAY FROM. In solving these tasks, the model acts *autonomously*, that is, without user intervention during processing.

The entire model is based on concepts of dynamic field theory (DFT) and does not use algorithmic solutions.[1] The model is thus a single dynamical system that consists of a large number of coupled differential equations. On a functional level, it can be subdivided into five parts. Figure 3.1 shows an overview diagram of the model, in which these five parts are highlighted by white boxes with blue labels. The parts are as follows:

[1] The only exception to this is the algorithmic preprocessing of the camera images. This is used here to avoid adding unnecessary complexity to the model in the form of a neurally realistic vision system.

Perception The perceptual system represents the spatial position and feature values of all objects in the camera input. The representation consists of peaks in two three-dimensional dynamic neural fields that are both defined over a shared two-dimensional visual space and the respective additional feature dimensions color and motion direction. Activation in these

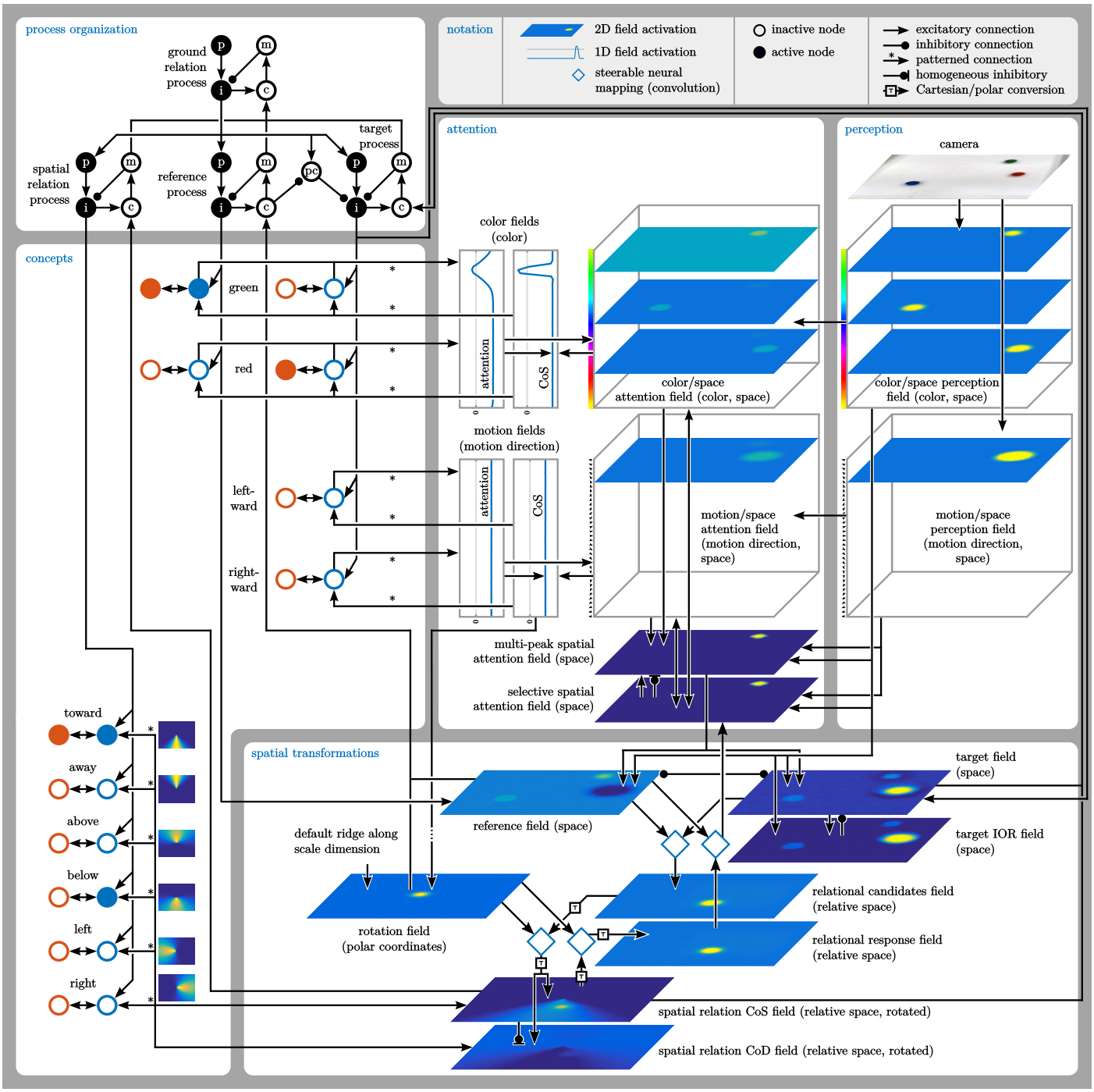


FIGURE 3.1: Diagram overview of the entire model ("process organization" and "concepts" are not shown in full). The figure shows an activation snapshot during grounding the phrase "the red object moving toward the green object". The graphical notation of the diagram is explained in the "notation" box (top right). For three-dimensional fields (in "perception" and "attention" box), two-dimensional slices of activation are shown.

fields is driven by camera input and is continuously updated as objects move in the scene.

Attention The attentional system enables guiding the attentional focus along all feature dimensions (here: color, motion direction, and visual space) as well as bind the representation of attended objects across all feature dimensions. In absence of guided feature attention, the attentional system has a simple saliency mechanism that enables the model to select the most salient object in the scene. The saliency increases with the color saturation and size of objects; moving objects are perceived as more salient than stationary objects.

Spatial transformations The spatial transformation system enables the model to flexibly change the reference position and reference orientation relative to which objects are represented. This enables the model to extract spatial relations between pairs of objects based on fixed relational templates (e.g., TO THE LEFT OF). Spatial transformations are expressed as steerable neural mappings.

Concepts Concepts of color (e.g., RED), motion direction (e.g., LEFTWARD), or relational concepts (e.g., TO THE LEFT OF) are represented by neural nodes. Their perceptual meaning is encoded in patterned synaptic connections between these discrete nodes and continuous feature spaces. Conceptual representations are understood as an interface to language.

Process organization The model controls which of its parts are active and inactive at any moment in time. Due to this organization of processes, the model is autonomous and only requires an initial input to complete its task.

This chapter devotes a separate section to each of these five parts and describes them in detail, including, in particular, a mathematical formalization of the entire model. Please refer to Table 3.1 for an overview and short description of all dynamic elements, dynamic neural fields and dynamic neural nodes, and to Table 3.2 for a listing of all recurring variable names used throughout this chapter. Appendix A.1 contains a list of parameter values used in the equations.

The model was implemented and parameterized using the software framework *cedar*. This chapter focuses instead on a mathematical level of description and does not address implementation details. Please refer to Section 2.2.8 for information on the numerical implementation of neural dynamics. A short paragraph with technical details regarding the model can be found in Appendix A.3.

name	variable	description
color/space perception field	$u_{\text{PCS}}(x, y, c, t)$	object representation over color and space
motion/space perception field	$u_{\text{PMS}}(x, y, \phi, t)$	object representation over motion direction and space
color/space attention field	$u_{\text{ACS}}(x, y, c, t)$	attentional focus on color and space
motion/space attention field	$u_{\text{AMS}}(x, y, \phi, t)$	attentional focus on motion direction and space
color attention field	$u_{\text{AC}}(c, t)$	guides attentional focus on color
color CoS field	$u_{\text{ACcs}}(c, t)$	checks attended object for color
motion attention field	$u_{\text{AM}}(\phi, t)$	guides attentional focus on motion direction
motion CoS field	$u_{\text{AMcs}}(\phi, t)$	checks attended object for motion direction
selective spatial attention field	$u_{\text{AS}}(x, y, t)$	selective attentional focus for space
multi-peak spatial attention field	$u_{\text{ASm}}(x, y, t)$	relays spatial positions to spatial transformations
reference field	$u_{\text{R}}(x, y, t)$	spatial position of the reference object
target field	$u_{\text{T}}(x, y, t)$	spatial position of the target object
target IOR field	$u_{\text{IR}}(x, y, t)$	inhibition-of-return (IOR) for the target object
target IOR CoS field	$u_{\text{IRcs}}(x, y, t)$	condition of satisfaction (CoS) for the target IOR
relational candidates field	$u_{\text{RC}}(x, y, t)$	relative positions between target and reference objects
relational response field	$u_{\text{RR}}(x, y, t)$	relative position of selected reference object
spatial relation CoS field	$u_{\text{Scs}}(x, y, t)$	relative positions, rotated to align with motion; checks for overlap with spatial templates
spatial relation CoD field	$u_{\text{Scd}}(x, y, t)$	analogous to spatial relation CoS field; checks for overlap with inverse spatial template
rotation field	$u_{\text{ROT}}(\phi, s, t)$	holds current reference angle used for rotation
rotation default-direction field	$u_{\text{ROTd}}(\phi, s, t)$	holds default reference angle for rotation
rotation selection field	$u_{\text{ROTs}}(\phi, s, t)$	selects rotation angle; either object motion direction or default
target color memory nodes	$\vec{u}_{\text{TCM}}(t)$	language interface to color of target object
target color production nodes	$\vec{u}_{\text{TCP}}(t)$	encodes perceptual meaning of target colors
target motion memory nodes	$\vec{u}_{\text{TMM}}(t)$	language interface to motion direction of target object
target motion production nodes	$\vec{u}_{\text{TMP}}(t)$	encodes perceptual meaning of target motion direction
reference color memory nodes	$\vec{u}_{\text{RCM}}(t)$	language interface to color of reference object
reference color production nodes	$\vec{u}_{\text{RCP}}(t)$	encodes perceptual meaning of reference colors
spatial relation memory nodes	$\vec{u}_{\text{SM}}(t)$	language interface to spatial relations
spatial relation production nodes	$\vec{u}_{\text{SP}}(t)$	encodes perceptual meaning of spatial relations
prior intention node	$u_{\text{P}}(t)$	when active, its process will be activated
intention node	$u_{\text{I}}(t)$	when active, its process is currently active
CoS node	$u_{\text{C}}(t)$	signals that its process is successfully completed
CoS memory node	$u_{\text{M}}(t)$	signals that its process has been completed in the past

Table 3.1: Overview of fields and nodes of the model, grouped by the part of the model they appear in. From top to bottom, these are: perception, attention, spatial transformation, concepts, and process organization. Note that the four nodes in the process organization section are present for every process.

variable	description
x	horizontal space of the camera image
y	vertical space of the camera image
t	time
c	color
ϕ	motion direction
r	scale
u_{A}	activation of a field with identifier 'A'
w_{A}	synaptic weight with identifier 'A'
$k_{\mathrm{B,A}}$	kernel from a source 'A' to a target 'B'
g	sigmoid function and output of fields/nodes
τ	time scale of dynamics
h	(negative) resting level of dynamics
s	external input into a field or node
ξ	Gaussian white noise

Table 3.2: Variables used throughout this chapter

3.1 Perception

The perceptual system (top right white box in the overview diagram, Figure 3.1 on page 40) takes input from a camera and builds a representation of all objects visible in the camera image. That representation resides in two three-dimensional dynamic neural fields. The *color/space perception field* is defined over the two spatial dimensions x and y of the camera image and over the color dimension c. This field always has a stable peak of activation whenever there is a colored object visible in the camera image. The *motion/space perception field* is defined over the same two spatial dimensions x and y of the camera image and over the motion direction dimension ϕ. A peak in this field thus represents the spatial position of an object and the direction in which it is moving. This field always has a stable peak of activation whenever an object is moving in the scene; it has no peak for stationary objects.

To create input to the perception fields, each video frame of the camera goes through several preprocessing steps. Since this model is not intended to make predictions about the human visual system, the preprocessing of the camera input is implemented algorithmically. It is a placeholder for a neurally plausible model of human visual processing (Lomp et al., 2017).

Lomp, O., Faubel, C., & Schöner, G. (2017). A neural-dynamic architecture for concurrent estimation of object pose and identity. *Frontiers in Neurorobotics*, *11*(April), 1–17

3.1.1 Color perception

Each frame of the camera input is preprocessed based on generic image processing algorithms that crop the image, scale it down, and convert it to the hue, saturation, value (HSV) color space. In-

put to the color/space perception field is then produced in a three-dimensional space (camera space x, y and color c) that scales with the color saturation of objects in the scene (Figure 3.2).[2] High values thus appear around objects with uniform, saturated colors, while low values are in areas with low saturated colors (e.g., black and white).The color/space perception field is parameterized such that colored objects always create a peak in the field and that the peak is roughly the size of the object.

Please note that this simplified preprocessing assumes that the scene only features unicolored objects in front of a white background. It works best with objects that appear circular on the camera image (e.g., balls), because the Gaussian lateral interaction kernel of the color/space perception field reinforces this shape.

The activation variable u_{PCS} of the color/space perception field evolves in time t based on the following differential equation

$$\begin{aligned}\tau \dot{u}_{\text{PCS}}(x,y,c,t) = &- u_{\text{PCS}}(x,y,c,t) + h + w_\xi \cdot \xi_{\text{PCS}}(x,y,c,t) \\ &+ [k_{\text{PCS,PCS}} * g(u_{\text{PCS}})](x,y,c,t) \\ &+ [k_{\text{PCS,C}} * s_{\text{C}}](x,y,c,t),\end{aligned} \tag{3.1}$$

which is based on the equation for a multi-dimensional dynamic neural field (Equation 2.6). The second line formalizes the lateral interaction within the field and the third line denotes the preprocessed input s_{C} convolved with a kernel $k_{\text{PCS,C}}$.[3]

3.1.2 Motion perception

For the motion/space perception field, the preprocessing consists of a neural dynamic implementation of the counter-change model of motion perception (Berger et al., 2012). An illustration of this part of the model is shown in Figure 3.3. The input s_{C} that is given to the color/space perception field is used as an input to this model as well, but the color dimension is contracted

$$s(x,y,t) = \max_c(s_{\text{C}}(x,y,c,t)). \tag{3.2}$$

For each position in the camera image, transient detectors[4] are defined. The type reacting to positive change follows the differential equations

$$\tau\, \dot{u}_{\text{p}}(x,y,t) = -u_{\text{p}}(x,y,t) + s(x,y,t) - v_{\text{p}}(x,y,t), \tag{3.3}$$

$$\tau_{\text{v}}\, \dot{v}_{\text{p}}(x,y,t) = -v_{\text{p}}(x,y,t) + s(x,y,t). \tag{3.4}$$

The type reacting to negative change is defined analogously

$$\tau\, \dot{u}_{\text{n}}(x,y,t) = -u_{\text{n}}(x,y,t) - s(x,y,t) - v_{\text{n}}(x,y,t), \tag{3.5}$$

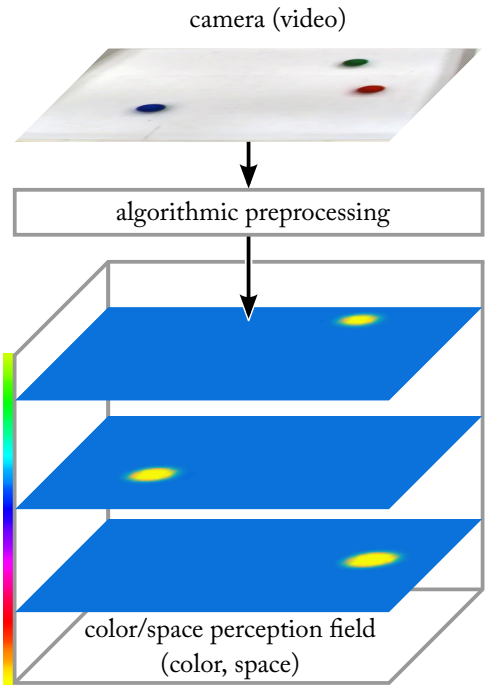


FIGURE 3.2: Color perception of the model.

[2] Please refer to Appendix A.2 for a more detailed description of the numerical implementation.

[3] The squared brackets denote a convolution operator, as formalized in Equation 2.17 on page 27.

Berger, M., Faubel, C., Norman, J., Hock, H., & Schöner, G. (2012). The counter-change model of motion perception: An account based on dynamic field theory. In A. E. P. Villa (Ed.), *ICANN 2012, Part I, LNCS 7552* (pp. 579–586). Berlin Heidelberg: Springer

[4] See Section 2.2.4.

$$\tau_{\mathrm{v}}\,\dot{v}_{\mathrm{n}}(x,y,t) = -v_{\mathrm{n}}(x,y,t) - s(x,y,t). \tag{3.6}$$

Please note that the input $s(x,y,t)$ acts inhibitorily on this second type of transient detector. The output of both types of transient detectors is given by the semi-linear rectifier function $f(u) = \max(0,u)$.

Motion is perceived when negative change at one position coincides with positive change at another position. Comparing the response of pairs of positive and negative detectors for a fixed distance r_0 between positions yields a response

$$\begin{aligned} s_{\mathrm{T}}(x,y,\phi,t) =& f(u_{\mathrm{p}}(x,y,t)) \\ & \cdot f(u_{\mathrm{n}}(x - r_0 \cdot \cos(\phi), y - r_0 \cdot \sin(\phi), t)) \end{aligned} \tag{3.7}$$

that depends on time t and spans the three-dimensional space defined by the two-dimensional space of the image, x and y, as well as the angle ϕ between the compared positions.To represent the perceived motion, this response becomes input to the motion/space perception field whose activation u_{PMS} follows the dynamics

$$\begin{aligned} \tau \dot{u}_{\mathrm{PMS}}(x,y,\phi,t) =& -u_{\mathrm{PMS}}(x,y,\phi,t) + h + w_\xi \cdot \xi_{\mathrm{PMS}}(x,y,\phi,t) \\ & + [k_{\mathrm{PMS,PMS}} * g(u_{\mathrm{PMS}})](x,y,\phi,t) \\ & + [k_{\mathrm{PMS,T}} * s_{\mathrm{T}}](x,y,\phi,t). \end{aligned} \tag{3.8}$$

The activation represents where along x and y a moving object is detected and in which direction ϕ that object is moving. The motion/space perception field is parameterized such that moving objects always create a peak in the field.

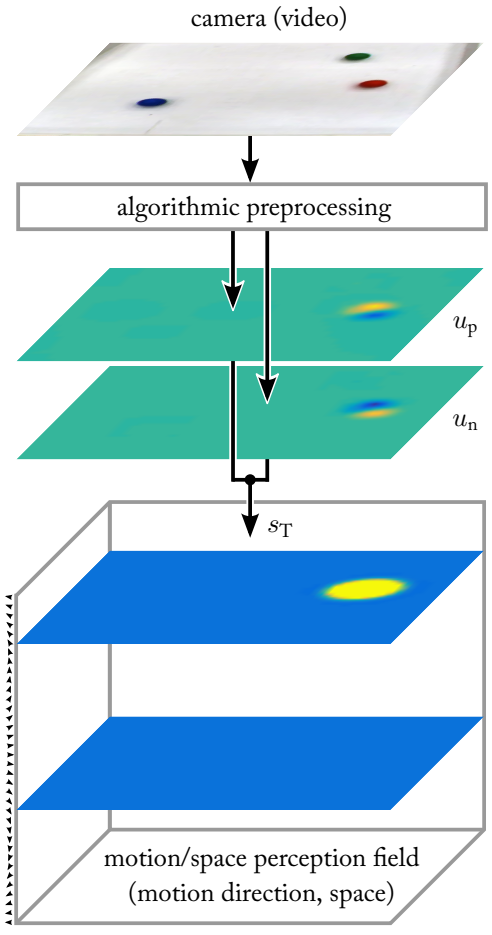


FIGURE 3.3: Motion perception of the model. The activation of the transient detectors, u_{p} and u_{n}, is shown in a different color code than the activation in the field to make their characteristic shape more visible.

3.2 Attention

This section introduces the attentional system of the model (central white box in Figure 3.1 on page 40). At the core of the attentional system are two three-dimensional dynamic neural fields that are defined over the same dimensions as the two perceptual fields: the *color/space attention field* is defined over the color dimension c and retinal space x and y; the *motion/space attention field* is defined over the motion direction ϕ and retinal space x and y. If a peak comes up in these attention fields, the object it represents is interpreted to be in attentional focus. Each field receives input from its corresponding perceptual field, reflecting the feature values and spatial position of all objects in the scene. However, this input is not strong enough to form peaks in the fields, it yields only subthreshold bumps of activation. Attentional processes are modeled by giving

additional input to the fields along single feature dimensions, highlighting a certain color, motion direction, or spatial position, and thereby pushing subthreshold bumps that overlap with the input through the detection instability.

Input to the attention fields comes from fields that are defined over single feature dimensions: the *color attention field* is defined over the color dimension c and gives input to the color/space attention field; the *motion attention field* is defined over the motion direction dimension ϕ and gives input to the motion/space attention field. These two fields implement part of a *feature attention* mechanism as they control which feature values are attended to. For instance, if there is a peak in the color attention field that represents green colors, the expansion coupling to the color/space attention field leads to a subthreshold input in that field in the form of a sheet (Figure 3.4). Along the color dimension, this sheet of activation is centered on the position that is encoding for the color red; throughout the other two (spatial) dimensions the sheet is homogeneous. The increased activation in the sheet pushes overlapping bumps of activation through the detection instability. This brings their respective objects, all of which are of red color, into attentional focus. Analogously, a peak in the motion attention field brings all objects into attentional focus that move into a certain direction.

Further input into the two three-dimensional attention fields is given by the *selective spatial attention field* that is defined over the spatial dimensions x and y. This field only allows for a single peak to form at any given time. If a peak forms, the expansion coupling to both the color/space attention field and the motion/space attention field leads to subthreshold input in those fields in the form of a cylinder. The cylinder is centered on the position of the peak along the two spatial dimensions and is homogeneous throughout the color and motion direction dimensions in the respective fields. The increased activation in the cylinder highlights objects that are located at the given spatial position; the selective spatial attention field thus implements a mechanism for *spatial attention*.

The color/space attention field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{ACS}}(x,y,c,t) = &- u_{\text{ACS}}(x,y,c,t) + h + w_\xi \cdot \xi_{\text{ACS}}(x,y,c,t) \\ &+ [k_{\text{ACS,ACS}} * g(u_{\text{ACS}})](x,y,c,t) \\ &+ [k_{\text{ACS,PCS}} * g(u_{\text{PCS}})](x,y,c,t) \\ &+ [k_{\text{ACS,AC}} * g(u_{\text{AC}})](c,t) \\ &+ [k_{\text{ACS,AS}} * g(u_{\text{AS}})](x,y,t) \\ &- w_{\text{ACS,SR}} \cdot g(u_{\text{SR}}(t)),\end{aligned} \tag{3.9}$$

where u_{ACS} is its own activation, u_{PCS} is the activation of the col-

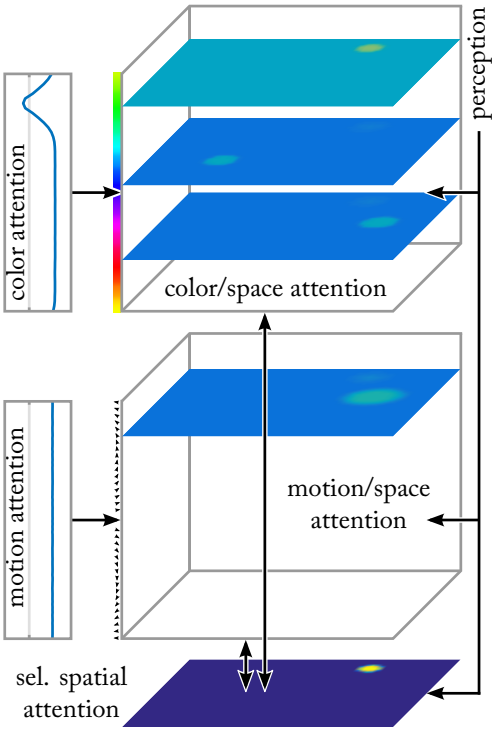


Figure 3.4: Attentional system of the model.

or/space perception field, u_{AC} is the activation of the color attention field, and u_{AS} is the activation of the selective spatial attention field. The inhibitory coupling formalized in the last line comes from a suppression node of a 'reset process', which will be explained in Section 3.5.

Analogously, the activation u_{AMS} of the motion/space attention field follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{AMS}}(x,y,\phi,t) = &- u_{\text{AMS}}(x,y,\phi,t) + h + w_\xi \cdot \xi_{\text{AMS}}(x,y,\phi,t) \\ &+ [k_{\text{AMS,AMS}} * g(u_{\text{AMS}})](x,y,\phi,t) \\ &+ [k_{\text{AMS,PMS}} * g(u_{\text{PMS}})](x,y,\phi,t) \\ &+ [k_{\text{AMS,AM}} * g(u_{\text{AM}})](\phi,t) \\ &+ [k_{\text{AMS,AS}} * g(u_{\text{AS}})](x,y,t) \\ &- w_{\text{AMS,SR}} \cdot g(u_{\text{SR}}(t)),\end{aligned} \tag{3.10}$$

where u_{PMS} is the activation of the motion/space perception field, u_{AM} is the activation of the motion attention field, and u_{AS} is the activation of the selective spatial attention field. Again, the inhibitory coupling formalized in the last line will be explained in Section 3.5.

3.2.1 Feature attention

The feature attention mechanism of the model is in part formed by the color attention field and the motion attention field. They give input to the three-dimensional attention fields and control which features are attended to. The color attention field with activation u_{AC} evolves in time based on the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{AC}}(c,t) = &- u_{\text{AC}}(c,t) + h + w_\xi \cdot \xi_{\text{AC}}(c,t) \\ &+ [k_{\text{AC,AC}} * g(u_{\text{AC}})](c,t) \\ &+ \sum_{i=1,\dots,N_{\text{C}}} W_{\text{C}i}(c) \cdot g(u_{\text{TCP}_i}(t)) \\ &+ \sum_{j=1,\dots,N_{\text{C}}} W_{\text{C}j}(c) \cdot g(u_{\text{RCP}_j}(t)),\end{aligned} \tag{3.11}$$

where the last two lines are the input from two arrays of neural nodes, $\vec{u}_{\text{TCP}}$ and $\vec{u}_{\text{RCP}}$, that represent discrete color concepts like RED or GREEN for two roles that objects may have in a scene. The perceptual meaning of the concepts is encoded in the connection weights $\vec{W}_{\text{C}}(c)$ between the nodes and the field. The activation of the nodes and their connection weights are expressed in vector form where $N_{\text{C}} = \dim(\vec{u}_{\text{TCP}}) = \dim(\vec{u}_{\text{RCP}}) = \dim(\vec{W}_{\text{C}})$. This coupling will be explained in detail in Section 3.4.

Analogously, the motion attention field with activation u_{AM} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{AM}}(\phi, t) = &- u_{\mathrm{AM}}(\phi, t) + h + w_{\xi} \cdot \xi_{\mathrm{AM}}(\phi, t) \\ &+ [k_{\mathrm{AM,AM}} * g(u_{\mathrm{AM}})](\phi, t) \\ &+ \sum_{i=1,\ldots,N_{\mathrm{M}}} W_{\mathrm{M}i}(\phi) \cdot g(u_{\mathrm{TMP}_i}(t)),\end{aligned} \tag{3.12}$$

where the last line formalizes input from a similar array of neural nodes, $\vec{u}_{\mathrm{TMP}}, = N_{\mathrm{M}}$, that represent discrete concepts of motion direction, for instance LEFTWARD or UPWARD. The meaning of the concepts is encoded in the connection weights $\vec{W}_{\mathrm{M}}(\phi)$. Please note that $\dim(\vec{W}_{\mathrm{M}}) = \dim(\vec{u}_{\mathrm{TMP}}) = N_{\mathrm{M}}$. This coupling will also be explained in Section 3.4.

[5] See Section 2.2.7.

For each of these attention fields, there is an additional condition of satisfaction (CoS)[5] field defined over the same dimension (color c and motion direction ϕ, respectively). The *color CoS field* receives subthreshold input both from the color attention field and the color/space attention field and forms a peak if the inputs match. Analogously, the *motion CoS field* matches input from the motion attention field and the motion/space attention field. In this view, the one-dimensional attention fields can be thought of as the *intention fields* of elementary behaviors that govern the attention of their respective feature dimension.[6]

[6] See Section 2.2.7.

The color CoS field with activation u_{ACcs} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{ACcs}}(c, t) = &- u_{\mathrm{ACcs}}(c, t) + h + w_{\xi} \cdot \xi_{\mathrm{ACcs}}(c, t) \\ &+ [k_{\mathrm{ACcs,ACcs}} * g(u_{\mathrm{ACcs}})](c, t) \\ &+ [k_{\mathrm{ACcs,AC}} * g(u_{\mathrm{AC}})](c, t) \\ &+ \max_{x,y}([k_{\mathrm{ACcs,ACS}} * g(u_{\mathrm{ACS}})](x, y, c, t)) \\ &- \max_{c'}(g(u_{\mathrm{AC}}(c', t))),\end{aligned} \tag{3.13}$$

where the third line is localized input from the color attention field, u_{AC}, reflecting the feature values that are in attentional focus. In order to form peaks, it has to overlap with localized input from the color/space attention field, u_{ACS} (line 4). The global inhibitory input (line 3) from the color attention field, u_{AC}, is required because the color CoS field should also form a peak based only on input from the color/space attention field in case no peak is present in the color attention field.

Analogously, the activation u_{AMcs} of the motion CoS field is

governed by the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{AMcs}}(\phi, t) = &- u_{\text{AMcs}}(\phi, t) + h + w_\xi \cdot \xi_{\text{AMcs}}(\phi, t) \\ &+ [k_{\text{AMcs,AMcs}} * g(u_{\text{AMcs}})](\phi, t) \\ &+ [k_{\text{AMcs,AM}} * g(u_{\text{AM}})](\phi, t) \\ &+ \max_{x,y}([k_{\text{AMcs,AMS}} * g(u_{\text{AMS}})](x, y, \phi, t)) \\ &- \max_{\phi'}(g(u_{\text{AM}}(\phi', t))).\end{aligned} \tag{3.14}$$

3.2.2 Spatial attention

The selective spatial attention field fulfills more than one function within the model. Its primary function is to bring attentional focus to a unique spatial position. It does so by providing input to the color/space attention field as well as to the motion/space attention field. Since the selective spatial attention field only allows a single peak at any point in time, the attentional focus can only be at a single spatial position. In case the selective spatial attention field receives input on multiple spatial positions, for instance from the perceptual fields, it makes a selection decision and forms a peak at the position with the strongest activation. Input also comes from spatial transformations reflecting the spatial position of an object that has been selected based on how well it fits with a spatial relation.[7] The selective spatial attention field thus relays the response from the spatial transformations back to the attentional system so that the features of the object can be extracted.

[7] This will be explained in detail in Section 3.3.

The second function of the selective spatial attention field is to integrate the feature representations that are distributed over multiple fields. The coupling between the selective spatial attention field and the two three-dimensional attention field is bidirectional along the shared spatial dimensions x and y. This essentially binds all feature values of each object together through the fields' shared spatial dimensions x and y. Whenever an object is brought into attentional focus by highlighting a certain feature, for instance its color, the peak will be projected into the selective spatial attention field, forming a peak there at the object's spatial position. This creates a subthreshold cylinder of activation in both three-dimensional attention fields, bringing up peaks that represent the other features of the object, for instance its motion direction. This integration of distributed feature representations is a vital aspect of feature integration theory (Treisman & Gelade, 1980).

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136

As a third function, the selective spatial attention field implements a saliency mechanism that enables the model to bring attentional focus to a spatial position in absence of other attentional cues. This is required, for instance, to generate a description of a scene,

where the only input comes from the camera and no further instructions are given. The saliency mechanism is modeled by couplings from the perception fields to the selective spatial attention field. Since these inputs are summed, moving objects are more salient than stationary objects. Furthermore, the larger the object and the more saturated its colors, the more salient it is. Since the strength of the input from the three-dimensional attention fields varies depending on how many features are in attentional focus, there are inhibitory connections from the color attention field and the motion attention field to the selective spatial attention field. That way, if a feature is brought into attentional focus by a peak in one of those fields, the input to the selective spatial attention field does not vary as strongly.

A last function of the spatial attention mechanism is to relay the spatial position of objects that are in attentional focus to other fields (downward in Figure 3.1). In some cases, multiple object positions have to be relayed; since the selective spatial attention field cannot have more than one peak, this requires a *multi-peak spatial attention field*. It is defined over the same spatial dimensions x and y and receives the same input as the selective spatial attention field but allows multiple peaks to form.

Contrary to the selective spatial attention field, the multi-peak spatial attention field does not give input to the three-dimensional attention fields. This is because input from the multi-peak spatial attention field would enable multiple spatial positions to be highlighted and, as a consequence, multiple objects with (possibly) different features being brought into attentional focus. This is a problem when the model extracts features from these objects as feature binding may be lost.

The selective spatial attention field is coupled to the multi-peak spatial attention field such that a peak in the selective spatial attention field will inhibit all other peaks in the multi-peak spatial attention field, leaving a single peak in both fields. The selective spatial attention field can be "activated" and "deactivated" by homogeneous input. In the deactivated state, all activation in the field remains below threshold. This enables the multi-peak spatial attention field to have multiple peaks and relay them to other fields. When the selective spatial attention field is activated, both spatial attention fields become selective and relay the position of a single object to the three-dimensional attention fields. One can think of the two fields as a functional unit, a spatial attention field where the selectivity can be activated and deactivated by homogeneous input.

The selective spatial attention field evolves based on the follow-

ing differential equation

$$\begin{aligned}
\tau_{\text{AS}}\dot{u}_{\text{AS}}(x,y,t) = &- u_{\text{AS}}(x,y,t) + h_{\text{AS}} + w_{\xi} \cdot \xi_{\text{AS}}(x,y,t) \\
&+ [k_{\text{AS,AS}} * g(u_{\text{AS}})](x,y,t) \\
&+ \max_{c}([k_{\text{AS,ACS}} * g(u_{\text{ACS}})](x,y,c,t)) \\
&+ \max_{\phi}([k_{\text{AS,AMS}} * g(u_{\text{AMS}})](x,y,\phi,t)) \\
&+ \max_{c}([k_{\text{AS,PCS}} * g(u_{\text{PCS}})](x,y,c,t)) \\
&+ \max_{\phi}([k_{\text{AS,PMS}} * g(u_{\text{PMS}})](x,y,\phi,t)) \\
&- \max_{c}(g(u_{\text{AC}}(c,t))) \\
&- \max_{\phi}(g(u_{\text{AM}}(\phi,t))) \\
&- [k_{\text{AS,IR}} * g(u_{\text{IR}})](x,y,t) \\
&+ C_{\text{SU}}(x,y,t) \\
&+ w_{\text{AS,GPI}}\, g(u_{\text{GPI}}(t)) \\
&+ w_{\text{AS,GAI}}\, g(u_{\text{GAI}}(t)),
\end{aligned} \tag{3.15}$$

where u_{AS} is its own activation variable, the third to sixth line are the inputs from the color/space attention field (u_{ACS}), the motion/space attention field (u_{AMS}), the color/space perception field (u_{PCS}), and the motion/space perception field (u_{PMS}), respectively. The seventh and eighth line are global inhibitory inputs from the color attention field (u_{AC}) and motion attention field (u_{AM}), respectively. The inputs normalize the input from the three-dimensional attention fields, depending on the number of features specified in the relational phrase. The ninth and tenth line are inputs from the 'target IOR field',[8] which inhibits the model from selecting the same object twice and the 'relational response field',[9] whose activation represents a match between the spatial position of objects and relational templates. Both fields will be explained in more detail later. The last two lines formalize input from the process organization system,[10] which will also be explained later in more detail. Here, both inputs determine whether or not the spatial attention mechanism is selective.

[8] See Section 3.3.2 and Equation 3.19.

[9] See Section 3.3.5 and Equation 3.35.

[10] The input comes from the intention nodes of the perceptual boost process (penultimate line) and the spatial attention process (last line).

The activation u_{ASm} of the multi-peak spatial attention field is

governed by the differential equation

$$\begin{aligned}\tau_{\text{ASm}}\dot{u}_{\text{ASm}}(x,y,t) = &- u_{\text{ASm}}(x,y,t) + h_{\text{ASm}} + w_\xi \cdot \xi_{\text{ASm}}(x,y,t)\\ &+ [k_{\text{ASm,ASm}} * g(u_{\text{ASm}})](x,y,t)\\ &+ \max_c([k_{\text{ASm,ACS}} * g(u_{\text{ACS}})](x,y,c,t))\\ &+ \max_\phi([k_{\text{ASm,AMS}} * g(u_{\text{AMS}})](x,y,\phi,t))\\ &+ \max_c([k_{\text{ASm,PCS}} * g(u_{\text{PCS}})](x,y,c,t))\\ &+ \max_\phi([k_{\text{ASm,PMS}} * g(u_{\text{PMS}})](x,y,\phi,t))\\ &- \max_c(g(u_{\text{AC}}(c,t)))\\ &- \max_\phi(g(u_{\text{AM}}(\phi,t)))\\ &- [k_{\text{ASm,IR}} * g(u_{\text{IR}})](x,y,t)\\ &+ [k_{\text{ASm,AS}} * g(u_{\text{AS}})](x,y,t)\\ &+ w_{\text{ASm,GPI}}\, g(u_{\text{GPI}}(t))\\ &- w_{\text{ASm,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{3.16}$$

which is mostly analogous to Equation 3.15. Line ten formalizes input from the selective spatial attention field (u_{AS}), exciting the spatial position of the selected object and inhibiting all other positions. The last two lines are input from the process organization system described in Section 3.5, where the penultimate line is input from the intention node of the 'perceptual boost process', which brings the field into a dynamic regime where it can form peaks and the last line is strong inhibitory input from the suppression node of the 'reset process'.

3.3 Spatial transformations

This section introduces the spatial transformation system of the model, which extracts spatial relations between a pair of objects (bottom white box in Figure 3.1 on page 40). Extracting the spatial relation between two objects requires that they are brought into the attentional foreground and that a representation of their spatial positions is built up. From these positions, one can extract the relative spatial position of one of the objects with respect to the other by shifting the representation to center it on the second object. Such a shift can be implemented by a steerable neural mapping. The model uses this idea twice in succession to generate an object representation that is both shifted, centered on a reference position, and rotated according to the motion direction of one of the objects.

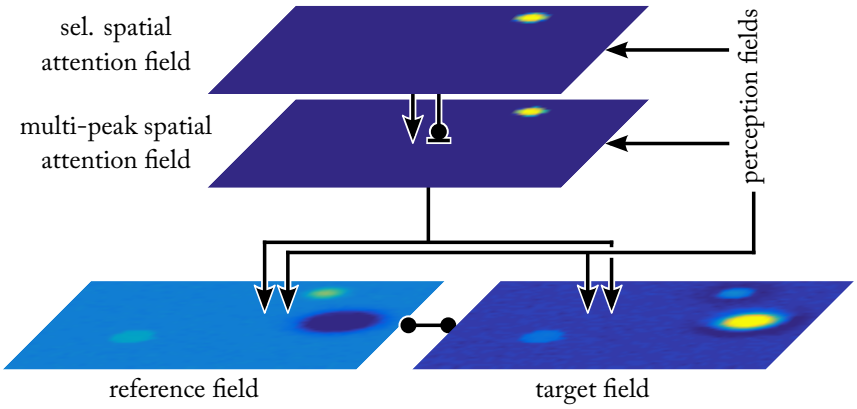


FIGURE 3.6: The target field and reference field and their connection to the spatial attention fields. In this example, the reference object is currently in the attentional foreground.

3.3.1 Target and reference

Consider the following example: given the visual input shown in Figure 3.5 and the phrase "the red object to the left of the green object", the model has to find both a red and a green object and determine whether the red object is to the left of the green object. In this example, the red object is the *target object*, the object the phrase is referring to. The green object is the *reference object*; it needs to be found in the scene only to facilitate finding the correct target object. The description "to the left of" refers to a spatial relation that specifies the relative position of the target object with respect to the reference object. In the model, this relative position is determined by a steerable neural mapping.[11] This requires that the position of the target object and the position of the reference object are represented in separate dynamic neural fields that give input to the steerable neural mapping. The spatial position of the target object is represented by a peak of activation in the *target field*, which is defined over the spatial dimensions x and y. The spatial position of the reference object (or multiple candidates for the reference object) are held by the *reference field*, defined over the same spatial dimensions. Both fields receive input from the multi-peak spatial attention field, reflecting the spatial positions of the objects that are currently in attentional focus (Figure 3.6). Attentional focus is sequentially directed toward the target object and the reference object while at the same time raising the resting level of the corresponding field (target field and reference field). When the resting level of the target field is raised, the selective spatial attention field is activated, enforcing a selection decision for a single target object. When the resting level of the reference field is raised, the selective spatial attention field is deactivated, allowing for multiple candidate positions to be represented in the reference field. Both the sequentiality

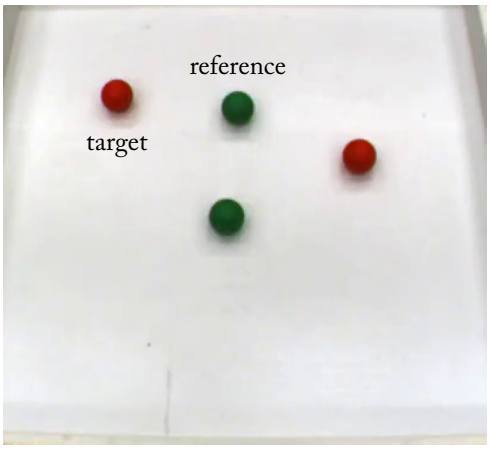


FIGURE 3.5: *Target object* and *reference object* given the phrase "the red object to the left of the green object".

[11] See Section 2.2.5.

and the activation and deactivation of the spatial attention field are controlled by the process organization system, which is explained in Section 3.5.

Both the target field and the reference field receive additional input from the color/space perception field and the motion/space perception field, enabling the fields to track the position of moving objects even after they have initially formed peaks. Finally, the target field and the reference field inhibit each other such that a given spatial position is only ever represented in one of the fields.

The activation u_{T} of the target field evolves in time based on the following differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{T}}(x,y,t) = &- u_{\mathrm{T}}(x,y,t) + h_{\mathrm{T}} + w_{\xi} \cdot \xi_{\mathrm{T}}(x,y,t) \\ &+ [k_{\mathrm{T,T}} * g(u_{\mathrm{T}})](x,y,t) \\ &+ \max_{c}([k_{\mathrm{T,PCS}} * g(u_{\mathrm{PCS}})](x,y,c,t)) \\ &+ \max_{\phi}([k_{\mathrm{T,PMS}} * g(u_{\mathrm{PMS}})](x,y,\phi,t)) \\ &+ [k_{\mathrm{T,ASm}} * g(u_{\mathrm{ASm}})](x,y,t) \\ &- [k_{\mathrm{T,R}} * g(u_{\mathrm{R}})](x,y,t) \\ &+ w_{\mathrm{T,TTI}}\, g(u_{\mathrm{TTI}}(t)) \\ &- w_{\mathrm{T,SR}}\, g(u_{\mathrm{SR}}(t)),\end{aligned} \tag{3.17}$$

[12] See Section 3.5 for details on the last two inputs.

where the third and fourth line are contraction couplings from the color/space perception field (u_{PCS}) and the motion/space perception field (u_{PMS}), respectively, the fifth line is input from the multi-peak spatial attention field (u_{ASm}), and the sixth line is inhibitory input from the reference field (u_{R}). The penultimate line is input from the process organization system, more specifically, the intention node of the 'target field process'; it brings the target field into a dynamic regime where it can form peaks. The last line of Equation 3.17 is strong inhibitory input from the suppression node of the 'reset process'.[12]

The activation u_{R} of the reference field is governed by the following differential equation, which is structured analogously to Equa-

tion 3.17

$$\begin{aligned}\tau\dot{u}_{\mathrm{R}}(x,y,t) = &- u_{\mathrm{R}}(x,y,t) + h + w_{\xi} \cdot \xi_{\mathrm{R}}(x,y,t) \\ &+ [k_{\mathrm{R,R}} * g(u_{\mathrm{R}})](x,y,t) \\ &+ \max_{c}([k_{\mathrm{R,PCS}} * g(u_{\mathrm{PCS}})](x,y,c,t)) \\ &+ \max_{\phi}([k_{\mathrm{R,PMS}} * g(u_{\mathrm{PMS}})](x,y,\phi,t)) \\ &+ [k_{\mathrm{R,ASm}} * g(u_{\mathrm{ASm}})](x,y,t) \\ &- [k_{\mathrm{R,T}} * g(u_{\mathrm{T}})](x,y,t) \\ &+ w_{\mathrm{R,RFI}}\, g(u_{\mathrm{RFI}}(t)) \\ &- w_{\mathrm{T,SR}}\, g(u_{\mathrm{SR}}(t)),\end{aligned} \tag{3.18}$$

Here, the penultimate line is input from the intention node of the 'reference field process', which will bring the reference field into a dynamic regime where it can form peaks.[13]

[13] See Section 3.5.

3.3.2 Target inhibition-of-return

In some cases it may be necessary to find the correct target object by trial-and-error. The scene shown in Figure 3.5 is such a case, for instance. Here, to find the correct target object, the model has to choose one of the red objects in the scene and check whether it conforms with the description (i.e., whether it is to the left of a green object). It would not help to start by selecting the reference object first, as the same problem would arise. In case a chosen candidate does not match the description, the process has to be repeated with another candidate. To solve this problem, the model uses an *inhibition-of-return (IOR)* mechanism, common in models of visual attention (Itti & Koch, 2001). To keep track of which objects have already been checked, the *target IOR field* holds a representation of the spatial positions of all objects that have been represented in the target field since the beginning of the task. The target IOR field is defined over the same spatial dimensions x and y as the target field and receives input from it. It also receives input from the color/space perception field and the motion/space perception field so that it can track moving objects. The lateral interaction of the target IOR field features local excitation that is strong enough to support peaks even after input from the target field ceases. To prevent the same object from being attended to more than once, it gives inhibitory input to both the selective spatial attention field and the multi-peak spatial attention field.[14]

Itti, L. & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203

[14] See Equations 3.15 and 3.16.

$$\begin{aligned}\tau \dot{u}_{\text{IR}}(x,y,t) = &- u_{\text{IR}}(x,y,t) + h_{\text{IR}} + w_\xi \cdot \xi_{\text{IR}}(x,y,t) \\ &+ [k_{\text{IR,IR}} * g(u_{\text{IR}})](x,y,t) \\ &+ [k_{\text{IR,T}} * g(u_{\text{T}})](x,y,t) \\ &+ \max_c([k_{\text{IR,PCS}} * g(u_{\text{PCS}})](x,y,c,t)) \\ &+ \max_\phi([k_{\text{IR,PMS}} * g(u_{\text{PMS}})](x,y,\phi,t)) \\ &+ w_{\text{IR,TII}}\, g(u_{\text{TII}}(t)),\end{aligned} \tag{3.19}$$

where the third line is input from the target field (u_{T}) and the fourth and fifth line are input from the color/space perception field (u_{PCS}) and motion/space perception field (u_{PMS}), respectively. The last line is input from the intention node of the 'target IOR process', which is part of the process organization system explained in Section 3.5. It brings the target IOR field into a dynamic regime where it can form peaks.

To determine whether the peak that is currently in the target field has already formed in the target IOR field, the *target IOR CoS field* takes subthreshold input from both of these fields (it is defined over the same spatial dimensions x and y). Only if the inputs overlap does the target IOR CoS field form a peak. Its activation, u_{IRcs}, is governed by the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{IRcs}}(x,y,t) = &- u_{\text{IRcs}}(x,y,t) + h + w_\xi \cdot \xi_{\text{IRcs}}(x,y,t) \\ &+ [k_{\text{IRcs,IRcs}} * g(u_{\text{IRcs}})](x,y,t) \\ &+ [k_{\text{IRcs,T}} * g(u_{\text{T}})](x,y,t) \\ &+ [k_{\text{IRcs,IR}} * g(u_{\text{IR}})](x,y,t) \\ &- w_{\text{IRcs,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{3.20}$$

where the third line formalizes input from the target field (u_{T}) and the fourth line is input from the target IOR field (u_{IR}). The last line is strong inhibitory input from the suppression node of the 'reset process'.[15]

[15] See Section 3.5.

3.3.3 Relative position

Based on the spatial position of the target object and the (possibly multiple) candidates for the reference object, a steerable neural mapping produces activation that represents the position of the target object relative to the position of all reference objects (Figure 3.7). The resulting activation is projected into the *relational candidates field*, which is defined over the spatial dimensions x and y. Its acti-

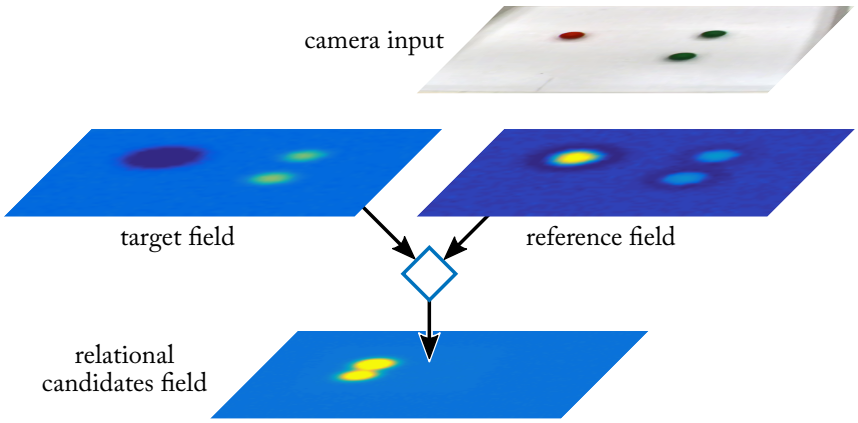


FIGURE 3.7: Example of a spatial transformation that yields the relative position of the (red) target object with respect to all (green) reference objects.

vation u_{RC} evolves based on the differential equation

$$\begin{aligned}\tau_{\mathrm{RC}}\dot{u}_{\mathrm{RC}}(x,y,t) = &- u_{\mathrm{RC}}(x,y,t) + h + w_\xi \cdot \xi_{\mathrm{RC}}(x,y,t) \\ &+ [k_{\mathrm{RC,RC}} * g(u_{\mathrm{RC}})](x,y,t) \\ &+ \iint dx'dy'\, A_{\mathrm{SD}}(x',y',t) \\ &\qquad B_{\mathrm{SD}}(x-x',y-y',t).\end{aligned} \tag{3.21}$$

The last term formalizes how the steerable neural mapping is implemented here as a convolution. As first input

$$A_{\mathrm{SD}}(x,y,t) = [k_{\mathrm{T}} * g(u_{\mathrm{T}})](x,y,t), \tag{3.22}$$

it takes the output of the target field (u_{T}), convolved with a kernel (k_{T}). As second input

$$B_{\mathrm{SD}}(x,y,t) = [k_{\mathrm{R}} * g(u_{\mathrm{R}})](x,y,t), \tag{3.23}$$

it takes the output of the reference field (u_{R}), convolved with a kernel (k_{R}).

Figure 3.7 shows an example of this transformation, where the activation of the target field holds a single peak that represents a possible target object and the reference field holds two peaks that are candidates for the reference object. The activation in the relational candidates field holds two peaks, each representing the relative position of the (possible) target object to one of the (possible) reference objects.

3.3.4 Rotation

Representing the relative position of the target object is sufficient to resolve spatial relations in static scenes. For instance, to determine

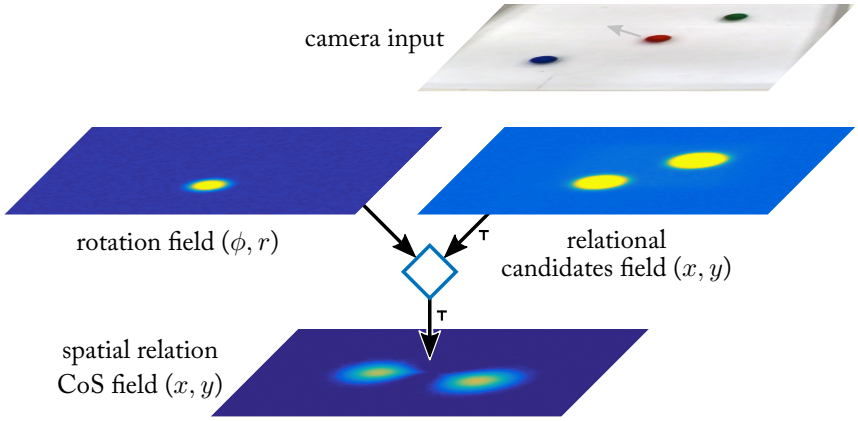


FIGURE 3.8: Example of a spatial transformation that aligns the relative positions of the target object with its motion direction. The relational candidates field contains a representation of relative positions of the target object (red) with respect to the other objects. The activation in the spatial relation CoS field represents the same positions in a rotated reference frame, where the motion direction of the target object is aligned with the y-axis of the plot. A small "T" next to an arrow denotes a conversion between Cartesian and polar coordinates.

whether the target object is to the left of any of the reference objects, one can overlay the activation in the relational candidates field with a pattern that expresses the spatial relation TO THE LEFT OF. However, it is not sufficient to resolve relations in dynamic scenes. To determine whether the target object is moving toward any of the reference objects, for instance, the pattern that expresses the spatial relation TOWARD depends on the motion direction of the target object. For this reason, the representation of the relative position of the target object in the relational candidates field is transformed further, essentially rotating it around the center, depending on the motion direction of the target object (Figure 3.8). This way, the spatial relations in dynamic scenes can be expressed by a fixed pattern. Rotation in Cartesian coordinates is implemented by a shift of the representation in polar coordinates. As before, the shift is achieved by a steerable neural mapping, which is implemented with a convolution here. Conversion between Cartesian coordinates (x, y) and polar coordinates (angular coordinate ϕ, radial coordinate r) is determined by

$$\phi = \tan^{-1}\left(\frac{y}{x}\right), \tag{3.24}$$

$$r = \sqrt{x^2 + y^2}, \tag{3.25}$$

$$x = r\cos(\phi), \tag{3.26}$$

$$y = r\sin(\phi). \tag{3.27}$$

Since the above equations establish a one-to-one mapping between Cartesian coordinates and polar coordinates, the conversion can be expressed neurally by fixed synaptic connections.

The steerable neural mapping implementing the rotation projects activation into the *spatial relation CoS field* and the *spatial relation CoD field*,[16] both of which are defined over Cartesian spatial

[16] CoD stands for "condition of dissatisfaction", the opposite of a condition of satisfaction (CoS).

dimensions x and y.

Both fields match the input against spatial templates that represent relational concepts such as TOWARD or TO THE LEFT OF (Figure 3.9). If the spatial relation CoS field forms a peak during such a match, the model converged on a combination of target and reference object that fits a given or existing relation. The field is selective and thus only allows for a single object to match any of the relational templates. If the spatial relation CoD field forms a peak, on the other hand, the current combination of target object and reference object candidates does not fit the relation. This triggers that the process of matching objects against spatial relations is repeated with a new target object. This mechanism is part of the process organization system and will be explained in Section 3.5.

The spatial relation CoS field with activation u_{Scs} follows the differential equation

$$\begin{aligned}\tau_{\text{Scs}}\dot{u}_{\text{Scs}}(x,y,t) = &- u_{\text{Scs}}(x,y,t) + h_{\text{Scs}} + w_\xi \cdot \xi_{\text{Scs}}(x,y,t) \\ &+ [k_{\text{Scs,Scs}} * g(u_{\text{Scs}})](x,y,t) \\ &+ \iint d\phi' dr'\, A_{\text{RD}}(\phi', r', t) \\ &\qquad B_{\text{RD}}(\phi - \phi', r - r', t) \\ &+ \sum_{i=1,\dots,N_{\text{R}}} W_{\text{R}i}(x,y) \cdot g(u_{\text{SP}_i}(t)) \\ &+ w_{\text{Scs,SRI}}\, g(u_{\text{SRI}}(t)) \\ &- w_{\text{Scs,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{3.28}$$

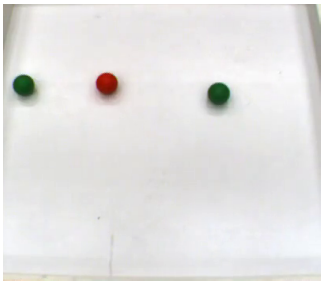
where the third and fourth line formalize the steerable neural mapping, which is implemented as a convolution between

$$A_{\text{RD}}(\phi, r, t) = [k_{\text{RC}} * g(u_{\text{RC}})](x, y, t), \tag{3.29}$$

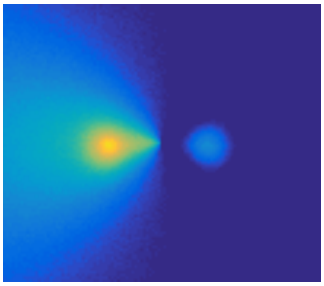
the output of the relational candidates field (u_{RC}), convolved with a Gaussian kernel (k_{RC}) and converted to polar coordinates and

$$B_{\text{RD}}(\phi, r, t) = [k_{\text{ROT}} * g(u_{\text{ROT}})](\phi, r, t), \tag{3.30}$$

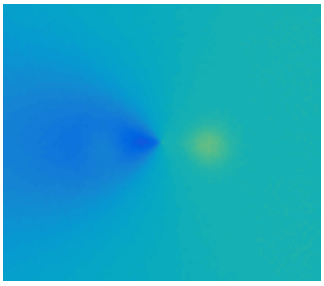
the output of the rotation field (u_{ROT}), convolved with a Gaussian kernel (k_{ROT}); this field will be explained later in this section. The fifth line of Equation 3.28 formalizes input from an array of neural nodes, $\vec{u}_{\text{SP}}$, that represent discrete relational concepts like TO THE LEFT OF or TOWARD. The meaning of the concepts is encoded in the connection weights $\vec{W}_{\text{R}}(x,y)$. Please note that $\dim(\vec{W}_{\text{R}}) = \dim(\vec{u}_{\text{SP}}) = N_{\text{R}}$. This will be explored in more detail in Section 3.4. The sixth line is input from the intention node of the 'spatial relational field process', which is part of the process organization system



(a) camera input



(b) spatial relation CoS field



(c) spatial relation CoD field

FIGURE 3.9: Activation of the (b) spatial relation CoS field and the (c) spatial relation CoD field. Both fields receive input representing the relative spatial position of two objects, visible as bumps of activation to the left and right of the center of the plots. Only the object on the left overlaps with the spatial template TO THE LEFT OF in the spatial relation CoS field. If this objects was not in the scene, the bump on the right would form a peak in the spatial relation CoD field.

(Section 3.5) and brings the spatial relation CoS field into a dynamic regime where it can form peaks. The last line is strong inhibitory input from the suppression node of the 'reset process'.[17]

[17] See Section 3.5.

The activation u_{Scd} of the spatial relation CoD field follows an analogous differential equation

$$\begin{aligned}\tau_{\text{Scd}}\dot{u}_{\text{Scd}}(x,y,t) = &- u_{\text{Scd}}(x,y,t) + h_{\text{Scd}} + w_\xi \cdot \xi_{\text{Scd}}(x,y,t) \\ &+ [k_{\text{Scd,Scd}} * g(u_{\text{Scd}})](x,y,t) \\ &+ \iint d\phi' dr'\, A_{\text{RD}}(\phi', r', t) \\ &\qquad B_{\text{RD}}(\phi - \phi', r - r', t) \\ &+ \sum_{i=1,\ldots,N_{\text{R}}} W_{\text{R}i}(x,y) \cdot g(u_{\text{SP}_i}(t)) \\ &+ w_{\text{Scd,SRI}}\, g(u_{\text{SRI}}(t)) \\ &- w_{\text{Scd,SR}}\, g(u_{\text{SR}}(t)) \\ &- w_{\text{Scd,Scs}} \max_{x',y'}([k_{\text{Scd,Scs}} * g(u_{\text{Scs}})](x', y', t)).\end{aligned} \tag{3.31}$$

The only differences are that the array of neural nodes projects their activation inhibitorily into the spatial relation CoD field and that the resting level and time constant are different. Furthermore, the field receives strong homogeneous inhibitory input from the spatial relation CoS field (u_{Scs}; last line). This prevents the spatial relation CoD field from forming a peak if the spatial relation CoS field signals a match between an object position and a spatial template.

The steerable neural mapping that implements the rotational transformation and produces the input for the spatial relation CoS field and the spatial relation CoD field takes as first input the output of the relational candidates field transformed into polar coordinates.[18] The second input comes from the *rotation field*, defined over polar coordinates ϕ and r.[19] Along the angular dimension ϕ, its activation represents either the current motion direction of the target object or a default direction in case the target object is stationary. The decision between these two alternatives is implemented in two fields that give input to the rotation field, both defined over polar coordinates. The *rotation default-direction field* receives fixed localized input along both the angular dimension ϕ and the radial dimension r and thus represents a default direction and scale for the transformation. The default direction is required to enable a transformation even if the target object is not moving; in this case the transformation rotates it by zero degrees. The *rotation selection field* receives the same fixed input along the radial dimension r, but along ϕ it receives activation from the motion CoS field holding the

[18] See Equations 3.28 and 3.29.

[19] See Equation 3.30.

current motion direction of the target object. The rotation selection field only forms a peak if there is also a peak in the motion CoS field, that is, if the target object is moving and its motion direction has been extracted. When the rotation selection field forms a peak, it homogeneously inhibits the rotation default-direction field. Thus, if the target object is moving, its motion direction is represented by the rotation field; otherwise, the default direction is represented. Along the radial dimension r, the fixed scale is used in any case.

The activation u_{ROTs} of the rotation selection field is governed by the differential equation

$$\begin{aligned}\tau_{\text{ROTs}}\dot{u}_{\text{ROTs}}(\phi, r, t) = &- u_{\text{ROTs}}(\phi, r, t) + h + w_\xi \cdot \xi_{\text{ROTs}}(\phi, r, t) \\ &+ [k_{\text{ROTs,ROTs}} * g(u_{\text{ROTs}})](\phi, r, t) \\ &+ [k_{\text{ROTs,AMcs}} * g(u_{\text{AMcs}})](\phi, t) \\ &+ s_{\text{ROTs,scale}}(r) \\ &+ w_{\text{ROTs,TMI}}\, g(u_{\text{TMI}}(t)),\end{aligned} \tag{3.32}$$

where the third line formalizes input from the motion CoS field (u_{AMcs}). The fourth line denotes the fixed input, $s_{\text{ROTs,scale}}$, for the scale, which is a Gaussian along the radial dimension r. The last line is excitatory input from the intention node of the 'target motion field process', which is part of the process organization system.[20] The input brings the rotation selection field into a dynamic regime where it can form peaks.

[20] See Section 3.5.

The activation u_{ROTd} of the rotation default-direction field follows the differential equation

$$\begin{aligned}\tau_{\text{ROTd}}\dot{u}_{\text{ROTd}}(\phi, r, t) = &- u_{\text{ROTd}}(\phi, r, t) + h + w_\xi \cdot \xi_{\text{ROTd}}(\phi, r, t) \\ &+ [k_{\text{ROTd,ROTd}} * g(u_{\text{ROTd}})](\phi, r, t) \\ &- \max_{\phi', r'}([k_{\text{ROTd,ROTs}} * g(u_{\text{ROTs}})](\phi', r', t)) \\ &+ s_{\text{ROTd,scale}}(r) \\ &+ s_{\text{ROTd,direction}}(\phi),\end{aligned} \tag{3.33}$$

where the third line is homogeneous inhibitory input from the rotation selection field (u_{ROTs}). The fourth and fifth line are fixed inputs for the scale and default direction that are Gaussians along their respective dimension.

The activation u_{ROT} of the rotation field thus follows the differ-

ential equation

$$\begin{aligned}\tau_{\text{ROT}}\dot{u}_{\text{ROT}}(\phi,r,t) = &- u_{\text{ROT}}(\phi,r,t) + h + w_\xi \cdot \xi_{\text{ROT}}(\phi,r,t) \\ &+ [k_{\text{ROT,ROT}} * g(u_{\text{ROT}})](\phi,r,t) \\ &+ [k_{\text{ROT,ROTs}} * g(u_{\text{ROTs}})](\phi,r,t) \\ &+ [k_{\text{ROT,ROTd}} * g(u_{\text{ROTd}})](\phi,r,t),\end{aligned} \tag{3.34}$$

where the third line formalizes the input from the rotation selection field (u_{ROTs}) and the fourth line is the input from the rotation default-direction field (u_{ROTd}).

3.3.5 Inverse transformations

A peak in the spatial relation CoS field signals that for the selected target object, a reference object exists that has a fitting spatial relation to it. In order to extract features from that reference object, the position of the peak in the spatial relation CoS field is projected back into the selective spatial attention field. This requires transformations that invert those described above. The *relational response field* receives input from the inverse rotation transformations. Its activation u_{RR} is governed by the differential equation

$$\begin{aligned}\tau\dot{u}_{\text{RR}}(x,y,t) = &- u_{\text{RR}}(x,y,t) + h + w_\xi \cdot \xi_{\text{RR}}(x,y,t) \\ &+ [k_{\text{RR,RR}} * g(u_{\text{RR}})](x,y,t) \\ &+ \iint d\phi' dr'\, A_{\text{RU}}(\phi',r',t) \\ &\qquad B_{\text{RU}}(\phi'-\phi, r'-r, t) \\ &+ w_{\text{RR,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{RR,DI}}\, g(u_{\text{DI}}(t)),\end{aligned} \tag{3.35}$$

where the convolution in the third and fourth line formalizes how the relative, rotated spatial position of the target object is transformed into its original orientation, where its first input,

$$A_{\text{RU}}(\phi,r,t) = [k_{\text{Scs}} * g(u_{\text{Scs}})](x,y,t), \tag{3.36}$$

consists of the output of the spatial relation CoS field (u_{Scs}), convolved with a kernel k_{Scs} and converted to polar coordinates. The other input

$$B_{\text{RU}}(\phi,r,t) = [k_{\text{ROT}} * g(u_{\text{ROT}})](\phi,r,t), \tag{3.37}$$

to the convolution comes from the rotation field (u_{ROT}), convolved with a kernel (k_{ROT}). The last two lines of Equation 3.35 formalize input from the process organization system, from the intention nodes of the 'ground relation process' and the 'describe process'.[21] Both inputs bring the relational response field into a dy-

namic regime that enables it to form a peak.

The output of the relational response field holds the relative position of the target object with respect to the reference object. This is transformed further into the absolute position of the reference object and feeds into the selective spatial attention field, bringing attentional focus to the spatial position of the reference object. The transformation is implemented as another convolution, where the input into the selective spatial attention field[22] is determined by

[22] See Equation 3.15.

$$C_{\text{SU}}(x, y, t) = \iint dx'dy' \; A_{\text{SU}}(x - x', y - y', t) \; B_{\text{SU}}(x', y', t), \tag{3.38}$$

which takes as first input

$$A_{\text{SU}}(x, y, t) = [k_{\text{RR}} * g(u_{\text{RR}})](x, y, t), \tag{3.39}$$

the output of the relational response field (u_{RR}) convolved with a kernel (k_{RR}) and as second input

$$B_{\text{SU}}(x, y, t) = [k_{\text{T}} * g(u_{\text{T}})](x, y, t), \tag{3.40}$$

the output of the target field (u_{T}), also convolved with a kernel (k_{T}).

Transforming the spatial position of the reference object into its original reference frame and feeding it into the selective spatial attention field enables the model to extract the features of the reference object and facilitates that the object can be described based on discrete concepts (e.g., a discrete color or motion direction). The part of the model that implements these discrete concepts is the focus of the next section.

3.4 Concepts

Conceptual representations express discrete concepts such as the color RED, the motion direction RIGHTWARD, or the spatial relation TO THE LEFT OF (bottom left box in the overview Figure 3.1 on page 40). They are understood here as an interface to language, as concepts may be mapped onto words and vice versa (Logan & Sadler, 1996). In the model, the perceptual meaning of a concept is represented by a discrete dynamic neural node[23] and its synaptic connections to a dynamic neural field that is defined over continuous feature dimensions (i.e., color, motion direction, and space). For instance, the color RED is represented by a neural node that is synaptically connected to every position in the color attention field. The connection weights have large values for regions in the field that

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13, pp. 493–529). Cambridge, MA, USA: MIT Press

[23] See Equation 2.7 on page 24.

code for red colors and low values elsewhere. When the node is activated, a peak comes up in the color attention field, ultimately bringing objects of red color into attentional focus. Conversely, when a red object is in attentional focus, the node representing RED is activated through connections from the color CoS field.

The model implements the color concepts RED, YELLOW, GREEN, and BLUE, the concepts of motion direction LEFTWARD, RIGHTWARD, UPWARD, and DOWNWARD, and the concepts of spatial relations TO THE LEFT OF, TO THE RIGHT OF, ABOVE, BELOW, TOWARD, and AWAY FROM. Each of these concepts is represented by a *pair* of interconnected neural nodes (Figure 3.10). The *memory node* acts as an interface to language. Since language is not part of the model, the node acts as an interface for the user, who can activate the node to specify which object the model is supposed to ground in the scene (e.g., RED and RIGHTWARD). Conversely, when the model is describing a scene, the user can read out the response from the activation of the memory nodes. Suprathreshold activation in the memory node does not have a direct impact on the rest of the model. It projects activation onto the *production node*, the second node in the pair, which acts as a gate to the rest of the model and can be activated and deactivated by the process organization system. The production node has patterned connections to the feature attention field and feature CoS attention field, encoding the meaning of the concept in terms of continuous feature dimensions close to the sensory-motor system. The production node also projects activation onto the memory node and is able to activate it as a feature description of an object that is in attentional focus.

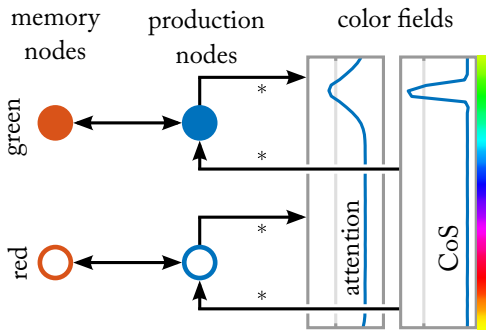


FIGURE 3.10: Diagram of the memory node and production node of the exemplary color concepts GREEN and RED. The arrows with stars denote synaptic connections that have a weight pattern. This pattern encodes the perceptual meaning of the concept. The patterns are implemented here as Gaussian functions with a mean at the color the concept represents.

In a relational phrase like "the red object to the left of the green object", color concepts can appear in two different roles. In the example, the concept RED describes the target object while the concept GREEN describes the reference object. The concepts are specific to these roles and must remain bound to them throughout their use in the model. The model thus has a dedicated pair of memory node and production node for every color concept in both the target role and the reference role. The concepts for motion direction are only represented for the target role because the model does not handle situations in which this feature is specified for the reference object. The concepts for spatial relations are not tied to any role since they describe relations rather than individual objects.

To summarize, the model has a memory node and a production node for the 'target color', the 'reference color', the 'target motion direction', and the 'spatial relation'. All nodes within each of these categories are governed by the same differential equation, which is explained next.

3.4.1 Color concepts

The activation of all *target color memory nodes* is written as a vector $\vec{u}_{\text{TCM}}$ that consists of the activation values of the individual nodes, each of which represents one of the following discrete color concepts: RED, YELLOW, GREEN, and BLUE. The activation of all nodes is governed by the differential equation

$$\begin{aligned}\tau\dot{\vec{u}}_{\text{TCM}}(t) = &- \vec{u}_{\text{TCM}}(t) + h + w_\xi \cdot \vec{\xi}(t) \\ &+ w_{\text{TCM,TCM}}\, g(\vec{u}_{\text{TCM}}(t)) \\ &- w_{\text{TCMgi}} \sum_{i=1,\ldots,N_{\text{C}}} g(u_{\text{TCM}_i}(t)) \\ &+ w_{\text{TCM,TCP}}\, g(\vec{u}_{\text{TCP}}(t)) \\ &+ \vec{s}_{\text{TCM,U}}(t) \\ &+ w_{\text{TCM,DI}}\, g(u_{\text{DI}}(t)) \\ &- w_{\text{TCM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{3.41}$$

which is based on the equation for a dynamic neural node (Equation 2.7). The second line formalizes the self-excitation, where $w_{\text{TCM,TCM}} > 0$ is a scalar weight. The third line is global inhibition between the individual nodes; every node receives inhibitory input from all other nodes, where $\dim(\vec{u}_{\text{TCM}}) = N_{\text{C}}$. The fourth line is excitatory input from the target color production nodes with activation $\vec{u}_{\text{TCP}}$. The fifth line denotes input given by the user, usually at the beginning of a task. If the task description includes the color of the target object, as in the phrase "the red object to the left of the red object", then the corresponding input is activated by the user. In this example, at a time t_0, the user would activate input for the node that represents the color RED for the target object: $\vec{s}_{\text{TCM,U}}(t_0) = (1, 0, 0, 0)^T$. The last two lines in Equation 3.41 are inputs from the process organization system[24]; the penultimate line is from the intention node of the 'describe process', which brings the target color memory nodes into a dynamic regime in which they can be activated from the target color production nodes. The last line is strong inhibitory input from a suppression node activated by the 'reset process'.

[24] See Section 3.5.

The activation $\vec{u}_{\text{TCP}}$ of all *target color production nodes* is expressed here analogously in vector form for the same color concepts as above.

It evolves in time based on the differential equation

$$\begin{aligned}\tau_{\text{TCP}}\dot{\vec{u}}_{\text{TCP}}(t) = &- \vec{u}_{\text{TCP}}(t) + h + w_{\xi} \cdot \vec{\xi}(t) \\ &+ w_{\text{TCP,TCP}}\, g(\vec{u}_{\text{TCP}}(t)) \\ &- w_{\text{TCPgi}} \sum_{i=1,\ldots,N_{\text{C}}} g(u_{\text{TCP}i}(t)) \\ &+ w_{\text{TCP,TCM}}\, g(\vec{u}_{\text{TCM}}(t)) \\ &+ w_{\text{TCP,ACcs}} \max_{c}(\vec{W}_{\text{C}}(c) \cdot g(u_{\text{ACcs}}(c,t))) \\ &+ w_{\text{TCP,TI}}\, g(u_{\text{TI}}(t)),\end{aligned} \tag{3.42}$$

where the second line is the self-excitation and the third line is the inhibition between all nodes with $\dim(\vec{u}_{\text{TCP}}) = N_{\text{C}}$. The fourth line is the excitatory input from the target color memory nodes (u_{TCM}). The fifth line formalizes the input from the color CoS field with activation u_{ACcs}. Each of the target color production nodes receives input from the field, where the synaptic strength at each position along the color dimension c is weighted with a function

$$W_{\text{C}i}(c) = \exp\left(-\frac{(c-\mu_{ci})^2}{2\sigma_c^2}\right), i = 1, \ldots, N_{\text{C}} \tag{3.43}$$

that is specific to each concept and encodes its perceptual meaning. All functions $\vec{W}_{\text{C}}, \dim(\vec{W}_{\text{C}}) = N_{\text{C}}$, are Gaussians defined over the color dimension c and differ only in their mean value μ_{ci}. Please note that these are the same weighting functions that determine the input strengths from the target color production nodes to the color attention field.[25] In the fifth line of Equation 3.42, the weighted output of the color CoS field is contracted and given as input to each of the nodes. The last line of the equation formalizes input from the intention node of the 'target process', which is part of the process organization system.[26] The input brings the target color production nodes into a dynamic regime in which they can be activated by input from the target color memory nodes or the color CoS field.

The differential equations that govern the activation of the reference color memory nodes and reference color production nodes are analogous to those for the target color memory nodes and target color production nodes, respectively. The activation of the reference color memory nodes is denoted in the same vector notation as

[25] See Equation 3.11 in Section 3.2.1.

[26] See Section 3.5.

above for the same color concepts and follows the equation

$$\begin{aligned}\tau\dot{\vec{u}}_{\text{RCM}}(t) = &- \vec{u}_{\text{RCM}}(t) + h + w_\xi \cdot \vec{\xi}(t) \\ &+ w_{\text{RCM,RCM}}\, g(\vec{u}_{\text{RCM}}(t)) \\ &- w_{\text{RCMgi}} \sum_{i=1,\ldots,N_{\text{C}}} g(u_{\text{RCM}i}(t)) \\ &+ w_{\text{RCM,RCP}}\, g(\vec{u}_{\text{RCP}}(t)) \\ &+ \vec{s}_{\text{RCM,U}}(t) \\ &+ w_{\text{RCM,DI}}\, g(u_{\text{DI}}(t)) \\ &- w_{\text{RCM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{3.44}$$

and the activation of the reference color production nodes evolves in time according to the equation

$$\begin{aligned}\tau_{\text{RCP}}\dot{\vec{u}}_{\text{RCP}}(t) = &- \vec{u}_{\text{RCP}}(t) + h + w_\xi \cdot \vec{\xi}(t) \\ &+ w_{\text{RCP,RCP}}\, g(\vec{u}_{\text{RCP}}(t)) \\ &- w_{\text{RCPgi}} \sum_{i=1,\ldots,N_{\text{C}}} g(u_{\text{RCP}i}(t)) \\ &+ w_{\text{RCP,RCM}}\, g(\vec{u}_{\text{RCM}}(t)) \\ &+ w_{\text{RCP,ACcs}} \max_c (\vec{W}_{\text{C}}(c) \cdot g(u_{\text{ACcs}}(c,t))) \\ &+ w_{\text{RCP,RI}}\, g(u_{\text{RI}}(t)).\end{aligned} \tag{3.45}$$

[27] See Equation 3.11 on page 47.

Please note that $\dim(\vec{u}_{\text{RCM}}) = \dim(\vec{u}_{\text{RCP}}) = N_{\text{C}}$. Furthermore, note that the reference color production nodes also give input to the color attention field that is weighted by $\vec{W}_{\text{C}}(c)$.[27]

3.4.2 Motion direction concepts

The nodes that represent the concepts of motion direction are structured analogously and follow similar equations. The activation of all *target motion memory nodes* is written as a vector $\vec{u}_{\text{TMM}}$ that consists of the activation values of the individual nodes, each of which represents one of the following discrete concepts of motion direction: LEFTWARD, RIGHTWARD, UPWARD, and DOWNWARD. The activation

of all nodes is governed by the differential equation

$$\begin{aligned}\tau\dot{\vec{u}}_{\text{TMM}}(t) = &- \vec{u}_{\text{TMM}}(t) + h + w_\xi \cdot \vec{\xi}(t) \\ &+ w_{\text{TMM,TMM}}\, g(\vec{u}_{\text{TMM}}(t)) \\ &- w_{\text{TMMgi}} \sum_{i=1,\ldots,N_{\text{M}}} g(u_{\text{TMM}i}(t)) \\ &+ w_{\text{TMM,TMP}}\, g(\vec{u}_{\text{TMP}}(t)) \\ &+ \vec{s}_{\text{TMM,U}}(t) \\ &+ w_{\text{TMM,DI}}\, g(u_{\text{DI}}(t)) \\ &- w_{\text{TMM,SR}}\, g(u_{\text{SR}}(t))\end{aligned} \tag{3.46}$$

and the target motion production nodes with the activation vector $\vec{u}_{\text{TMP}}$ follow the differential equation

$$\begin{aligned}\tau_{\text{TMP}}\dot{\vec{u}}_{\text{TMP}}(t) = &- \vec{u}_{\text{TMP}}(t) + h + w_\xi \cdot \vec{\xi}(t) \\ &+ w_{\text{TMP,TMP}}\, g(\vec{u}_{\text{TMP}}(t)) \\ &- w_{\text{TMPgi}} \sum_{i=1,\ldots,N_{\text{M}}} g(u_{\text{TMP}i}(t)) \\ &+ w_{\text{TMP,TMM}}\, g(\vec{u}_{\text{TMM}}(t)) \\ &+ w_{\text{TMP,AMcs}} \max_{\phi}(\vec{W}_{\text{M}}(\phi) \cdot g(u_{\text{AMcs}}(\phi, t))) \\ &+ w_{\text{TMP,TI}}\, g(u_{\text{TI}}(t)),\end{aligned} \tag{3.47}$$

where the fifth line formalizes input from the motion CoS field with activation u_{AMcs}. Here, the concepts of motion direction are encoded in the connection weights

$$W_{\text{M}i}(\phi) = \exp\left(-\frac{(\phi - \mu_{\phi_i})^2}{2\sigma_\phi^2}\right), i = 1, \ldots, N_{\text{M}} \tag{3.48}$$

defined over the motion direction ϕ. Please note that $\dim(\vec{W}_{\text{M}}) = \dim(\vec{u}_{\text{TMM}}) = \dim(\vec{u}_{\text{TMP}}) = N_{\text{M}}$. Analogously to the concepts of colors, all weighting functions are Gaussians that are defined over ϕ and differ only in their mean values μ_{ϕ_i}. Please note that these are the same weighting functions that determine the input strengths from the target motion production nodes to the motion attention field.[28]

[28] See Equation 3.12 on page 48.

3.4.3 Spatial relation concepts

The nodes that represent concepts of spatial relations are structured analogously and follow similar equations to the nodes described

above. The activation of all *spatial relation memory nodes* is written as a vector $\vec{u}_{\text{SM}}$ which consists of the activation values of the individual nodes, each of which represents one of the following discrete concepts of spatial relations: TO THE LEFT OF, TO THE RIGHT OF, ABOVE, BELOW, TOWARD, and AWAY FROM. The activation of all nodes is governed by the differential equation

$$\begin{aligned}\tau\dot{\vec{u}}_{\text{SM}}(t) = &- \vec{u}_{\text{SM}}(t) + h_{\text{SM}} + w_\xi \cdot \vec{\xi}(t) \\ &+ w_{\text{SM,SM}}\, g(\vec{u}_{\text{SM}}(t)) \\ &- w_{\text{SMgi}} \sum_{i=1,\ldots,N_{\text{R}}} g(u_{\text{SM}_i}(t)) \\ &+ w_{\text{SM,SP}}\, g(\vec{u}_{\text{SP}}(t)) \\ &+ \vec{s}_{\text{SM,U}}(t) \\ &+ w_{\text{SM,DI}}\, g(u_{\text{DI}}(t)) \\ &- w_{\text{SM,SR}}\, g(u_{\text{SR}}(t)) \\ &+ w_{\text{SM,SNI}}\, g(u_{\text{SNI}}(t)),\end{aligned} \tag{3.49}$$

where the third line expresses global inhibition between the individual nodes; every node receives inhibitory input from all other nodes, where $\dim(\vec{u}_{\text{SM}}) = N_{\text{R}}$. The last line is excitatory input from the intention node of the 'spatial memory node process', which is part of the process organization system,[29] and brings the spatial relation memory nodes into a dynamic regime in which they can be activated by input from the spatial relation production nodes.

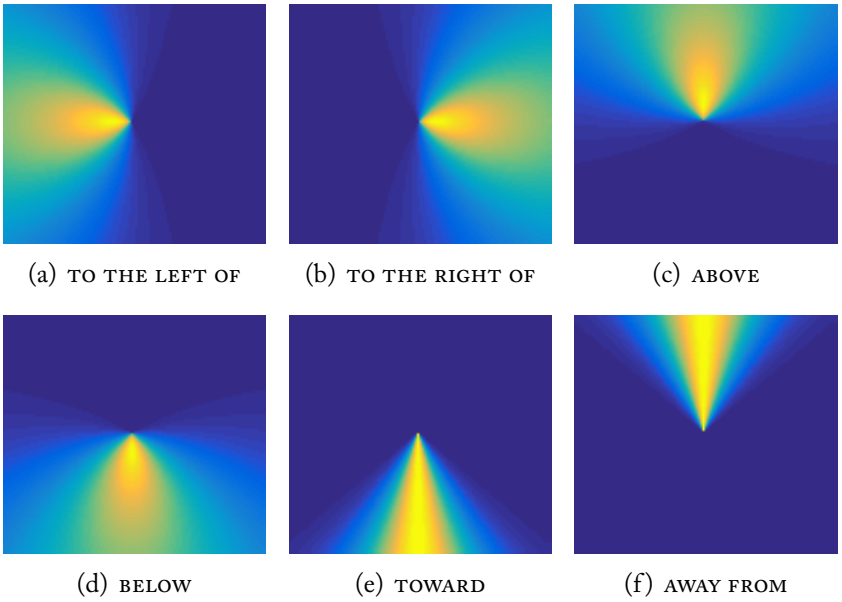
[29] See Section 3.5.

The *spatial relation production nodes* with activation $\vec{u}_{\text{SP}}$ follow a differential equation analogous to Equation 3.42

$$\begin{aligned}\tau_{\text{SP}}\dot{\vec{u}}_{\text{SP}}(t) = &- \vec{u}_{\text{SP}}(t) + h + w_\xi \cdot \vec{\xi}(t) \\ &+ w_{\text{SP,SP}}\, g(\vec{u}_{\text{SP}}(t)) \\ &+ w_{\text{SP,SM}}\, g(\vec{u}_{\text{SM}}(t)) \\ &+ w_{\text{SP,Scs}} \max_{x,y}(\vec{W}_{\text{R}}(x,y) \cdot g(u_{\text{Scs}}(x,y,t))) \\ &+ \vec{c}_{\text{SP,DI}}\, g(u_{\text{DI}}(t)) \\ &+ \vec{c}_{\text{SP,MD}}\, g(u_{\text{MD}}(t)) \\ &+ w_{\text{SP,SI}}\, g(u_{\text{SI}}(t)),\end{aligned} \tag{3.50}$$

where the fourth line formalizes the input from the spatial relation CoS field with activation u_{Scs}, the output of which is multiplied with the weighting functions

$$W_{\text{R}i}(x,y) = a \cdot \exp\left(\frac{-(\phi - \mu'_{\phi_i})^2}{2\sigma_\phi^2} + \frac{-(r-\mu_r)^2}{2\sigma_r^2}\right), \quad i = 1,\ldots,N_{\text{R}} \tag{3.51}$$



(a) TO THE LEFT OF (b) TO THE RIGHT OF (c) ABOVE

(d) BELOW (e) TOWARD (f) AWAY FROM

FIGURE 3.11: Relational concepts $\vec{W}_{\mathrm{R}}(x, y)$ as encoded in the synaptic connection strengths between spatial relation production nodes and the spatial relation CoS field.

that encode the meaning of each of the spatial relation concepts, where $\dim(\vec{W}_{\mathrm{R}}) = N_{\mathrm{R}}$. All functions consist of a Gaussian function in polar coordinates ϕ and r that is converted into Cartesian coordinates based on Equation 3.26–3.27. See Figure 3.11 for a plot of the weighting functions. The individual concepts differ mostly in the parameter μ'_{ϕ_i}, which determines the direction in which the template is oriented. For the weighting functions of the concepts of spatial motion relations TOWARD and AWAY FROM, the Gaussian function over the radial dimension r is very broad such that it has no relevant effect; the weighting functions are thus shaped like a beam (Figures 3.11(e), 3.11(f)). The weighting functions for the other concepts TO THE LEFT OF, TO THE RIGHT OF, ABOVE, and BELOW are inspired by behavioral data (Logan & Sadler, 1996) and are based on functions used in earlier models (Lipinski et al., 2012). Please note that these are the same weighting functions that determine the input strengths from the spatial relation production nodes to the spatial relation CoS field.[30]

The fifth line in Equation 3.50 formalizes input from the intention node of the 'describe process', which is part of the process organization system.[31] The input is weighted with the vector $\vec{w}_{\mathrm{SP,DI}} = (b, b, b, b, -a + b, -a + b)^T$, where $a, b > 0$, such that the spatial relation production nodes that represent concepts of spatial relations (the first four of the vector) receive higher input than the nodes that represent concepts of movement relations (the last two). The sixth line in Equation 3.50 formalizes input from a *motion detection node* that is activated whenever an object in the scene is moving. The output of that node is weighted with the vector $\vec{w}_{\mathrm{SP,MD}} =$

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13, pp. 493–529). Cambridge, MA, USA: MIT Press

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511

[30] See Equation 3.28 on page 59.

[31] See Section 3.5.

$(-a, -a, -a, -a, a, a)^T$ such that the spatial relation production nodes that represent spatial relational concepts (the first four of the vector: TO THE LEFT OF, TO THE RIGHT OF, ABOVE, BELOW) are inhibited and the nodes that represent motion spatial relational concepts (the last two: TOWARD, AWAY FROM) are excited whenever there is motion in the scene. The activation u_{MD} of the motion detection node evolves in time based on the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{MD}}(t) = &- u_{\text{MD}}(t) + h + w_\xi \cdot \xi_{\text{MD}}(t) \\ &+ w_{\text{MD,MD}}\, g(u_{\text{MD}}(t)) \\ &+ w_{\text{MD,PMS}} \max_{\phi,x,y}([k_{\text{MD,PMS}} * g(u_{\text{PMS}})](\phi, x, y, t)) \\ &+ w_{\text{MD,DI}}\, g(u_{\text{DI}}(t)),\end{aligned} \tag{3.52}$$

where the third line formalizes input from the motion/space perception field with activation u_{PMS} that has a peak whenever something is moving in the scene. The last line formalizes input from the intention node of the 'describe process', which is part of the process organization system.[32] This input brings the motion detection node into a dynamic regime in which it can be activated.

[32] See Section 3.5.

3.5 Process organization

One of the central problems in building a large-scale neural dynamic model as the one presented here, is the organization of its processes. For the model to function properly, the correct fields and nodes have to be brought into the right dynamic regimes at critical moments in time. While some processes may evolve in parallel, others may only be active in a sequence, or even only a certain sequential order. Importantly, organizing processes in this way should be done based on the same principles of neural dynamics that the rest of the model adheres to.

The following section explains the principles on which the process organization system is built. The entire system is then explained on the basis of these principles.

[33] See Section 2.2.7.

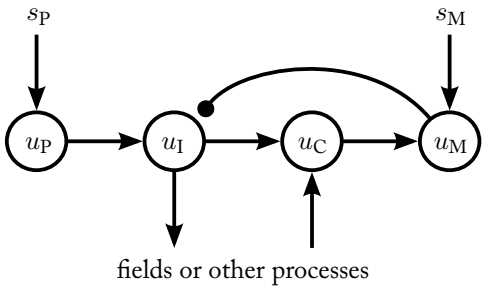


FIGURE 3.12: Diagram of a process in dynamic field theory. Nodes are represented by circles, where the names are the activation variables (see text). Excitatory connections are represented by regular arrows; inhibitory connections are denoted by lines ending in filled circles. All nodes have self-excitation, which is not shown here.

3.5.1 Processes

Each process that is controlled within the model is represented by a structure of four dynamic neural nodes (Figure 3.12). This structure is based on the elementary behaviors (EBs) of behavioral organization[33] but lacks the intention field and CoS field. They are thereby on a more abstract level with regards to the sensorimotor system.

The *prior intention node* represents whether the process will have to be activated at some point as part of another process. Its activation u_{P} evolves in time based on the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{P}}(t) = &- u_{\mathrm{P}}(t) + h + w_{\xi} \cdot \xi_{\mathrm{P}}(t) \\ &+ w_{\mathrm{P,P}}\, g(u_{\mathrm{P}}(t)) \\ &+ s_{\mathrm{P}}(t),\end{aligned} \tag{3.53}$$

where the second line is self-excitation and the third line is external input that comes from other processes to activate the node.

The *intention node* represents whether the process is currently having an impact on the model and is waiting to be finished. Its activation u_{I} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{I}}(t) = &- u_{\mathrm{I}}(t) + h_{\mathrm{I}} + w_{\xi} \cdot \xi_{\mathrm{I}}(t) \\ &+ w_{\mathrm{I,I}}\, g(u_{\mathrm{I}}(t)) \\ &+ w_{\mathrm{I,P}}\, g(u_{\mathrm{P}}(t)) \\ &- w_{\mathrm{I,M}}\, g(u_{\mathrm{M}}(t)),\end{aligned} \tag{3.54}$$

where the third line is excitatory input from the prior intention node (u_{P}), and the fourth line is inhibitory input from the CoS memory node (u_{M}), which deactivates the intention node. Some processes have to remain active even though their CoS is reached. In that case the connection weight $w_{\mathrm{I,M}}$ is set to zero.

The *condition of satisfaction (CoS) node* is activated as soon as the process is successfully finished. Its activation u_{C} is governed by

$$\begin{aligned}\tau \dot{u}_{\mathrm{C}}(t) = &- u_{\mathrm{C}}(t) + h_{\mathrm{C}} + w_{\xi} \cdot \xi_{\mathrm{C}}(t) \\ &+ w_{\mathrm{C,C}}\, g(u_{\mathrm{C}}(t)) \\ &+ w_{\mathrm{C,I}}\, g(u_{\mathrm{I}}(t)) \\ &+ s_{\mathrm{C}}(t),\end{aligned} \tag{3.55}$$

where the third line is excitatory input from the intention node (u_{I}). The fourth line is input from a field or from other processes; this will be explained in more detail in the next section.

The *CoS memory node* represents whether the process has already successfully finished as part of another process. Its activation u_{M} follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{M}}(t) = &- u_{\mathrm{M}}(t) + h_{\mathrm{M}} + w_{\xi} \cdot \xi_{\mathrm{M}}(t) \\ &+ w_{\mathrm{M,M}}\, g(u_{\mathrm{M}}(t)) \\ &+ w_{\mathrm{M,C}}\, g(u_{\mathrm{C}}(t)) \\ &+ s_{\mathrm{M}}(t),\end{aligned} \tag{3.56}$$

where the second line is strong self-excitation that keeps the node active even if the CoS node, which gives it input (line three), is

deactivated. The fourth line is input from other processes; it is the same input that activates the prior intention node such that $s_{\mathrm{M}}(t) = s_{\mathrm{P}}(t)$.

3.5.2 Sequences of processes

Like all dynamics that are part of this model, processes are continuously updated and directly react to external input. However, in some cases it may be required that processes only impact the model in a certain context, that two processes may not be active at the same time, or that processes may only become active in a certain sequential order. Such constraints on the sequential activation can be implemented in DFT using the precondition nodes and suppression nodes from behavioral organization.[34] As a reminder: a *precondition node* ensures that a process B is only activated once a process A is finished; a *suppression node* ensures that a process C is not active at the same time as a process D.

[34] See Section 2.2.7.

The suppression node can be used between two processes without a change from its definition in Section 2.2.7. The precondition node is also defined the same way except that it receives its inhibition from the CoS memory node of the process that is activated first (process A in the example above) rather than from its CoS node.

3.5.3 Heterarchy of processes

Processes can be structured both as a hierarchy and as a heterarchy. Processes on the lowest hierarchical level directly interact with the fields of the model. Their intention nodes give input to activate certain fields while their CoS nodes receive input in return that signals whether the processes have finished. These low-level processes can be reused in different contexts by more abstract processes that are defined on higher hierarchical levels. The structure I am proposing in this model allows for processes on higher hierarchical levels to activate an arbitrary number of lower-level processes. Furthermore, it allows for processes on lower levels to be activated by multiple higher-level processes, thereby creating a heterarchy. Additionally, constraints can be placed on the sequential order in which an arbitrary pair of processes is activated; more on this later.

The couplings that structure processes into a hierarchy are the same for all processes in the model. For a process A on a higher hierarchical level and several processes B_j on a lower hierarchical level (Figure 3.13), the coupling structure is as follows. Process A recruits the lower level processes B_j by projecting activation from its intention node (u_{AI}) to each of their prior intention node and

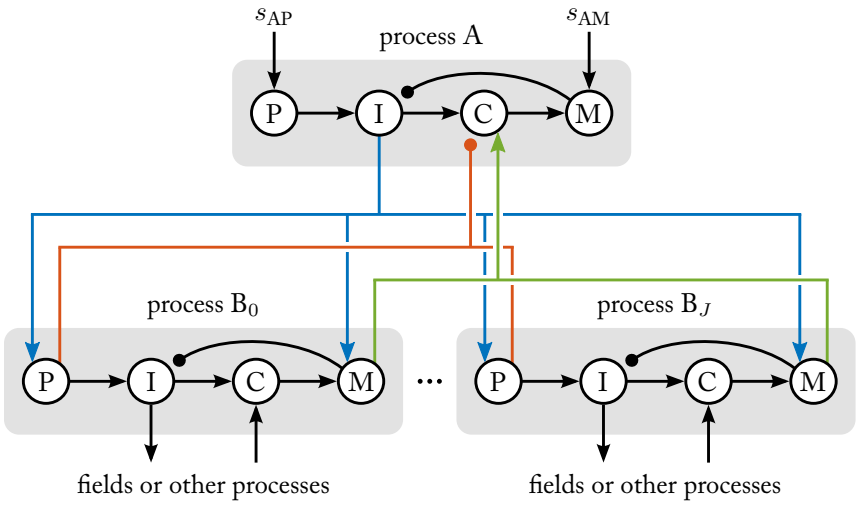


FIGURE 3.13: Hierarchy of processes with process A on the higher level recruiting processes B_j on the lower level. The letters in the nodes correspond to the indices of activation variables: "P" for prior intention nodes, "I" for intention nodes, "C" for CoS nodes, and "M" for CoS memory nodes. Blue connections are formalized in Equation 3.57 while red and green connections are explained in Equation 3.58.

CoS memory node such that

$$s_{\mathrm{B}_j\mathrm{P}}(t) = s_{\mathrm{B}_j\mathrm{M}}(t) = w_{\mathrm{AI}}\, g(u_{\mathrm{AI}}(t)), \tag{3.57}$$

where $s_{\mathrm{B}_j\mathrm{P}}(t)$ is the input to the prior intention nodes of the processes B_j, $s_{\mathrm{B}_j\mathrm{M}}(t)$ is the input to their CoS memory nodes (blue lines in Figure 3.13). The input is strong enough to activate the prior intention nodes but not the CoS memory nodes. The activated prior intention nodes, in turn, inhibit the CoS node of process A (orange lines in Figure 3.13), requiring additional input from the CoS memory nodes of all processes B_j (green lines). This means that the CoS of process A depends on the CoS of all processes B_j; it is only finished once all behaviors on the lower level are finished. The input into the CoS node of process A is thus determined by

$$s_{\mathrm{AC}}(t) = \sum_j g(u_{\mathrm{B}_j\mathrm{M}}(t)) - g(u_{\mathrm{B}_j\mathrm{P}}(t)), \tag{3.58}$$

where $u_{\mathrm{B}_j\mathrm{M}}$ is the activation of the CoS memory nodes and $u_{\mathrm{B}_j\mathrm{P}}$ the activation of the prior intention nodes of processes B_j. That way, the subthreshold activation of the CoS node of process A is lowered for every process that is recruited and the corresponding CoS input from that process is required to raise the activation.

For processes on the lowest level of the hierarchy the input into the CoS node comes directly from fields within the model. This requires that the resting level of the CoS node is lowered such that the activation of the node is only raised above threshold with input from the relevant field in the model.

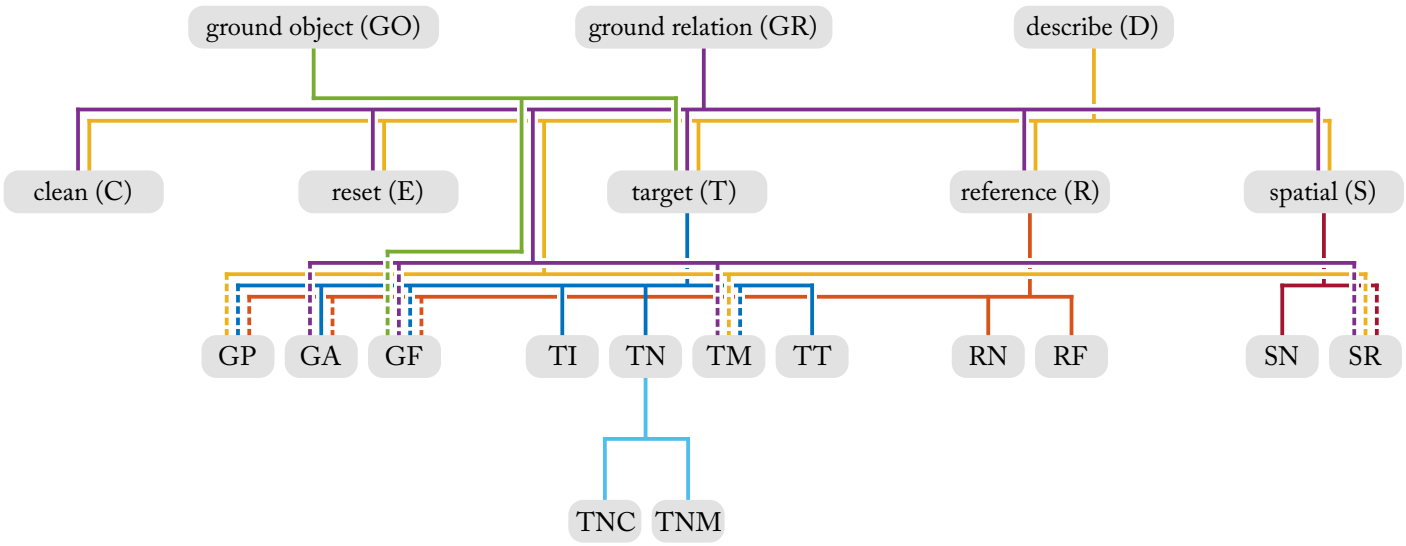


FIGURE 3.14: Heterarchy of all processes

A heterarchy could for instance be created if another process A_0 recruited a subset or all of the processes Bj. This is used in the model and will be mentioned in the following section that describes all processes that the model organizes.

3.5.4 Description of all processes

This section describes the process organization system of the model. It is hinted at in the top left box of the overview Figure 3.1 on page 40 but is too complex to illustrate in full. The model incorporates 21 processes, each of which is represented by the four dynamical nodes explained above (i.e., prior intention node, intention node, CoS node, and CoS memory node). With a few exceptions, all processes have the same parameters. The processes are structured in a heterarchy roughly subdivided into four hierarchical levels (Figure 3.14). The coupling that brings about this structure adheres to the principles described above. The following describes each process on a functional level. Please refer to Appendix B for a complete mathematical description of each process.

First hierarchical level

The highest hierarchical level has three processes that control different tasks of the model. These processes are activated by user input and for a given task, only one of the processes is ever activated at the same time.

Ground object process (GO) This process lets the model search the scene for a single object that has features specified by the user, for instance an object that is red and moving upward. The process thus implements a grounding task for a single object. It activates the target process (T) on the next lower hierarchical level and enables it to activate the feature process (GF) one further level below (green lines in Figure 3.14).

Ground relation process (GR) When this process is activated, the model will search for a target object that is described both by given features as well as a spatial relation to a reference object, whose features are also given. The process thus implements a grounding task for a pair of objects. The 'ground relation process' activates processes on lower hierarchical levels that control the search for the target object, the reference object, and the spatial relation (violet lines in Figure 3.14). Furthermore, processes on even lower levels receive activation, because they should only be active when the 'ground relation process' is active in combination with a process on the mid-level. Finally, the 'ground relation process' also activates nodes that determine the sequentiality in which processes on the next lower hierarchical level are activated; this will be explained later.

Describe process (D) This process generates a description of the scene it is currently presented with; it thus implements a description task. If there is only a single object in the scene, the model extracts the features of that object (e.g., "red"). If there are multiple objects in the scene, the model extracts the features and spatial relation of two of those objects (e.g., "red to the left of green"). Like the 'ground relation process', it activates all processes on the next lower hierarchical level as well as processes on lower hierarchical levels that should only be active as part of the 'describe process' (yellow lines in Figure 3.14). It also activates nodes that determine the sequential order in which processes on lower levels are activated. Some of the processes it activates give input to nodes (e.g., all memory nodes) that would be activated by user input in other tasks. This enables the model to activate the nodes as responses based only on peaks of activation in the attentional system.

Second hierarchical level

The second hierarchical level holds three processes that ground the three elements of a spatial phrase: the target object, the reference object, and the spatial relation. Two additional processes are required to support the organization of these grounding processes.

Target process (T) When this process is activated, the model grounds the target object. In case of a grounding task, where the relevant feature values of the object are given by the user, the model directs attentional focus to these features in order to find matching objects. In case of a description task, where no feature values of the object are given, the model directs attentional focus to the most salient object. The 'target process' activates several processes on the next lower hierarchical level (dark blue lines in Figure 3.14) that control aspects of the grounding of the target object.

Reference process (R) When this process is activated, the model grounds the reference object. This happens analogously to the grounding of the target object: in grounding tasks, the model directs attentional focus to specified features in order to find matching objects; in description tasks, the model directs attentional focus to the most salient object. The 'reference process' activates several processes on the next lower hierarchical level (orange lines in Figure 3.14) that control aspects of the grounding of the reference object. Since some of these processes are also activated by the 'target process' (T) (see Figure 3.14), the structure of processes is a heterarchy, not a hierarchy.

Spatial relation process (S) This process makes the model ground a spatial relation between two objects in the scene. In grounding tasks, the model activates a single concept of a spatial relation that is specified by the user (e.g., TO THE LEFT OF) and matches it against the relative positions between the selected candidates for the target and reference object. In description tasks, the model activates all available concepts of spatial relations and selects the one that fits best for the chosen target and reference object.

Clean process (C) When activated, the process ensures that any peaks have dissipated that may have formed in the color attention field and motion attention field while grounding the target object. This is required before the model can transition to grounding the reference object to avoid that remaining peaks affect the subsequent grounding process.

Reset process (E) This process inhibits large parts of the model in order to restart the process of grounding the target object and reference object. This is required if the selected objects do not match the specified spatial relation even though their other features may be as specified. The inhibition from the reset process initiates that a new combination of target object and reference object are grounded.

The peaks in the target IOR field, which are not affected by the inhibition, prevent the model from selecting a target object that has already been attended to.

Third hierarchical level

The third hierarchical level holds processes that control aspects of grounding the target object, the reference object, and the spatial relation. They are thus activated by the 'target process', the 'reference process', or the 'spatial relation process'. The first letter of their abbreviation reflects the process by which they are activated: target process: "T", reference process: "R", spatial relation process: "S". Some processes are activated both by the target process and the reference process. Their abbreviation begins with the letter "G" (for "generic").

Perceptual boost process (GP) This process enables the model to select an object based on salience when no features are specified. It does so by projecting homogeneous input into the selective spatial attention field and the multi-peak spatial attention field. The process can be activated both by the 'target process' and the 'reference process' as part of the high-level 'describe process'.

Spatial attention process (GA) This process makes the spatial attention system selective by projecting homogeneous input into the selective spatial attention field. The process can be activated both by the 'target process' and the 'reference process' as part of the high-level 'ground relation process'.

Feature process (GF) Checks whether features (e.g., color or motion direction) that were specified by the user have been found in the scene. More specifically, if only a single feature is specified, the process checks whether this has been found; if features in multiple feature dimensions are specified, it checks whether all of them have been found. The intention node of this process does not have connections to the rest of the model, only its CoS signal is used. The process is activated both by the 'target process' and the 'reference process' as part of the high-level 'ground object process' and 'ground relation process'.

Target IOR process (TI) This process enables the target IOR field to form peaks by projecting homogeneous activation into it. It also checks whether the object that is currently represented in the target field is also represented in the target IOR field. This process is activated by the 'target process' only.

Target memory node process (TN) Checks whether any of the target color memory nodes and target motion memory nodes are active. This is required when the model describes a scene to ensure a response has been given in the nodes. It does so by activating two processes on the lowest hierarchical level (light blue lines in Figure 3.14) that check this for the target color memory nodes and the target motion memory nodes, respectively.

Target motion field process (TM) Checks whether the motion perception system has extracted a motion direction from the perceptual input. This process is activated by the 'target process' but can only be activated as part of the higher-level 'ground relation process' and 'describe process'.

Target field process (TT) This process enables the target field to form peaks by projecting homogeneous activation into it. This process is activated by the 'target process' alone.

Reference memory node process (RN) This process is analogous to the 'target memory node process' as it checks whether any of the reference color memory nodes are active. It is activated only by the 'reference process'.

Reference field process (RF) This process is analogous to the 'target field process' as it enables the reference field to form peaks by projecting homogeneous activation into it. This process is activated by the 'reference process' alone.

Spatial memory node process (SN) This process checks whether any of the spatial relation memory nodes are active; it is thus analogous to the 'target memory node process' and 'reference memory node process'. When the 'describe process' is active, it additionally brings the spatial relation memory nodes into a dynamic regime where they can be activated from the spatial relation production nodes. The 'spatial memory node process' is activated only by the 'spatial relation process'.

Spatial relational field process (SR) This process enables the spatial relation CoS field and the spatial relation CoD field to form peaks by projecting homogeneous activation into them. It is thus analogous to the 'target field process' and the 'reference field process'. This process is activated by the 'spatial relation process' in conjunction with the 'ground relation process' and the 'describe process'.

Fourth hierarchical level

The fourth and lowest hierarchical level consists of only two processes, both of which are activated by the target memory node process.

Target memory node color process (TNC) This process checks whether any of the target color memory nodes are active.

Target memory node motion process (TNM) This process checks whether any of the target motion memory nodes are active.

3.5.5 Sequentiality

The model requires that some processes are activated in a certain sequential order. This section explains which processes have such constraints and how those are enforced. Please refer to Appendix B.22 for a complete mathematical description.

Grounding the target object and reference object

The target object and the reference object have to be grounded sequentially, they cannot be grounded in parallel. This is because the attentional system of the model does not encode the role (i.e., target or reference) of the object that is currently grounded. In the memory nodes and production nodes, the role is explicitly bound to the feature associated with that role. For instance, an active target color memory node may encode that the target object has the color red. The attentional system brings objects with red colors into the attentional foreground but does not encode whether the object is a target object or reference object. This binding is only kept up by ensuring that the spatial position of the found object is represented by a peak in the target field. This reflects behavioral data that shows that humans attend to target and reference sequentially when resolving spatial relations (Franconeri et al., 2012).

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227

The model provides that the target object is successfully grounded before the grounding of the reference object begins. This sequential order is arbitrary but was chosen here because the target object is commonly named first in relational phrases such as "The red object to the left of the green object". Moreover, in relational phrases that describe movement verbs like "The red object that is moving toward the green object", the moving object is defined here as the target object. Since moving objects are more salient than stationary objects, it is easier to ground a moving target object before grounding a stationary reference object than the other way around.

After the target object has been grounded, the target process is deactivated. However, it may take some time before activation decays below threshold in some fields of the model, most notably the color CoS field and the motion CoS field. If the reference process is activated before this happens, these remaining peaks interfere with the grounding of the reference object. This means the reference process can only be activated once the peaks in these fields have decayed—this is what the clean process checks for.

The model thus enforces that the target process is activated first, the clean process is activated after it, and the reference process is activated last. This is implemented using two precondition nodes: the first ensures that the clean process is activated after the target process, the second ensures that the reference process is activated after the clean process.

Producing a response after the relation has been evaluated

In description tasks, the spatial relation production nodes are all activated from the beginning to bias the selection of reference objects in the spatial relation CoS field. However, this means that the spatial relation memory nodes may only be brought into a dynamical regime where they can be activated when the actual spatial relation between the target object and the reference object has been evaluated. If they were brought into that regime before that, a random spatial relation memory node would activate based only on the noise within the system.

Thus, the model features a precondition node that only allows the spatial relation memory nodes to become active when there is a peak in the spatial relation CoS field. (The precondition node activates the 'spatial memory node process' only once the 'spatial relational field process' is finished.)

Similarly, the selection of a reference object from multiple candidates should only happen once an object has been selected in the spatial relation CoS field. Otherwise the wrong reference object could be selected. Since the selection happens in the selective spatial attention field, an additional precondition node ensures that the field can only form a peak once there is a peak in the spatial relation CoS field. (The precondition node activates the 'spatial attention process' once the 'spatial relational field process' is finished.) However, this precondition node can only be activated when the 'reference process' is active in conjunction with the 'ground relation process' or the 'describe process'.

Reset process

The 'reset process' inhibits large parts of the model when selected target and reference objects do not match the specified spatial relation. The inhibition is implemented with a suppression node that is activated by the intention node of the 'reset process'.

An additional precondition node ensures that the 'reset process' is only activated if there is a peak in the spatial relation CoD field, signaling that the target and reference object do not match the relation.

3.5.6 Organizational structures in the fields

The process organization is not solely based on the node structure but is fused with many fields within the core of the model that support it. There are two structures in the model that are particularly relevant to the process organization.

First, the fields that govern the feature attention of the model have a structure that is similar to the fields of EBs in behavioral organization.[35] In this view, the color attention field and motion attention field correspond to intention fields, because they determine what is in attentional focus and thus express the intention of the attentional behavior. The color CoS field and the motion CoS field, on the other hand, correspond to the CoS field. They check whether the peaks in their respective feature attention field overlaps with activation from the three-dimensional attention fields.

[35] See Section 2.2.7.

Second, the fields that evaluate the match between the selected target and reference object and the spatial relation have a similar structure. The spatial relation CoS field corresponds to a CoS field of an EB because it checks the overlap of an intended spatial relation with the actual relative position between the target and reference object. The spatial relation CoD field is a new concept of a condition of dissatisfaction (CoD) that is not present in the original EB of behavioral organization. Nevertheless, it is required whenever the mismatch between the intention and the current state of the behavior must be explicitly detected. There is no field to represent the intention in this case because the spatial relation is projected directly into the other fields.

Results 4

This section shows the performance of the model in different tasks and different visual scenes. Overall, the performance of the model is evaluated by simulating it in the software framework *cedar* in 104 different tests. 89 of these tests cover grounding tasks, where the model is presented with a phrase, such as "the red object to the left of the green object" and the model has to find the corresponding objects in the scene. The remaining 15 tests are description tasks, where the model has to produce such a phrase for a given visual scene. For all tests, the model remains the same; no parameters or connections are changed. The tests differ only in the visual input and the initial task input that the model receives. The task input corresponds to the user saying a phrase such as the one above. It is given at the beginning of the test by the user who activates corresponding dynamic neural nodes.

The visual input is unique in most tests, although some scenes are used for multiple tests. All visual scenes come from a video data set of 82 videos, all of which show colored balls on a white background. The data set was created specifically to test this model; it is designed to systematically probe how the model behaves under different conditions. The visual scenes vary, for instance, in the number of overall objects in the scene, the number of objects that match a description, or the number of moving objects or distractors. Please refer to Appendix C for a detailed description of the video data set and snapshots of all videos.

In all tests, the performance of the model is evaluated qualitatively, that is, it is determined whether the model behaves as expected. This is determined manually by observing the processes that unfold in the model while it is behaving. The expected behavior depends on the experiment but in most cases it includes that the

correct object is brought into attentional focus, that stable representations arise in the correct fields, and that conditions of satisfaction (CoS) of all involved processes are met.

The overall result of the tests is the model works as expected. For all grounding tasks, it grounds the given phrase in the scene where that is possible. Analogously, for all description tasks, it generates a phrase for the given scene, where possible.

Of the 104 tests that were conducted, this chapter describes and explains 14 in detail. For each of these tests, the visual input and the activation of the most relevant parts of the model are shown as they evolve over time.

4.1 Grounding tasks

In *grounding tasks*, a given phrase such as "the red object to the left of the green object" is matched to a scene. The task involves that all elements in the phrase, such as objects with specific features or relations between objects, are searched for and found in the scene. The phrase is regarded as grounded when the counterparts of all elements in the phrase have been found in the scene and perceptual representations of them have been formed.

In relation to a given visual scene, a phrase can lead to one of three cases. In the ideal case, the phrase refers to a target object that is uniquely identifiable in the visual scene. In case multiple objects match the description, one of them has to be selected as the target object. In the worst case, the description does not match any object in the scene. The process of grounding differs depending on which of these cases occurs.

Moreover, it depends on how specific the phrase is in describing the target object. To refer to a single object, it may specify only a single feature (e.g., "the red object"), or a conjunction of features ("the red object that is moving upward"). However, it may also involve multiple objects and their spatial relation, such as in the phrases "the red object that is to the left of the green object" or "the red object that is moving toward the green object".

The following section shows the results of systematically testing the model with all combinations of the variants stated above. In all cases, the phrase is supplied by a user who activates memory nodes that correspond to the concepts the phrase consists of. For instance, for the phrase "the red object that is moving toward the green object" the user would activate the target color memory node for the color RED, the reference color memory node for the color GREEN, and the spatial relation memory node for the movement relation TOWARD. The user then initiates the process of grounding the

phrase by activating one of the processes on the highest hierarchical level.[1] If the phrase refers only to a single object, the user activates the 'ground object process'. If the phrase specifies multiple objects and their relation, as in the example above, the user activates the 'ground relation process'. Once activated, the model autonomously grounds the given phrase, without further intervention by the user.

[1] See Section 3.5.

The following sections show tests of grounding tasks that differ both in the given phrase and the visual scene. Each test is referred to with an identifier (G1,…,G89). To aid understanding, the target object is always of red color, even in tests where the color is irrelevant.

4.1.1 Single features

The tests shown in this section establish that the attentional system and saliency system of the model work in a variety of settings. In all these tests, the task input given to the model only specifies a single feature of a single object.

Color

For a first set of tests, the task input corresponds to the phrase "the red object" and thus only specifies its color. In these tests, the user activates the target color memory node for the color RED and subsequently activates the 'ground object process'. Table 4.1 lists all of these tests (G1,…,G18). It summarizes information about the visual scene that was presented to the model in each test. In the columns from left to right, it shows the identifier of the test, the total number of objects in the scene, the number of red objects (targets) in the scene, how many of these red objects are moving, how many objects that are not red (distractors) are moving, whether or not a target can be found (marked with '+' or '−', respectively), as well as a language description of the visual scene. The test marked in blue is described in detail in the next section.

The result of the tests show that the model only brings objects into the attentional foreground that are specified by the phrase (in this case, red objects). In all visual scenes that contain a red object (those marked with '+'), the model is able to bring it into attentional foreground and thereby ground the phrase. This works irrespective of the number of distractor objects in the scene and whether they move or are stationary (e.g., G9). In case there are multiple potential targets in the visual scene (G10,…,G18), the model selects one of the objects while ignoring any distractor objects. In making this selection, the model prefers moving red objects over stationary red objects because they are more salient[2]. If multiple red objects are all

[2] See Section 3.2.2.

ID	# objects	# potential targets	# moving potential targets	# moving distractors	+/−	description
G1	0	0	0	0	−	no object in the scene
G2	4	0	0	0	−	no targets in the scene
G3	4	0	0	1	−	one moving distractor
G4	4	0	0	2	−	multiple moving distractors
G5	4	1	0	0	+	one target in the scene
G6	4	1	1	0	+	moving target
G7	4	1	0	1	+	target; moving distractor
G8	4	1	1	1	+	moving target/distractor
G9	4	1	0	2	+	multiple moving distractors
G10	4	2	0	0	+	two stationary targets
G11	4	2	0	1	+	stationary targets; moving distractor
G12	4	2	1	0	+	moving target
G13	4	2	0	2	+	only distractors are moving
G14	4	2	1	1	+	moving and stationary targets/distractors
G15	4	2	1	2	+	moving/stationary targets; moving distractors
G16	4	2	2	1	+	moving targets; moving/stationary distractors
G17	4	2	2	2	+	all objects moving
G18	4	2	2	0	+	only distractors are moving

Table 4.1: Tests of grounding tasks in which the phrase is "the red object". The test marked in blue is described in the next section. The sixth column from the left denotes whether the the phrase can be grounded in the scene (marked with "+") or not (marked with '−'). See text for details on how to interpret this table.

either moving or stationary, the model often selects the ones that are closer to the bottom of the visual scene. This is due to perspective distortion[3] in the video, which leads to objects on the bottom appearing larger, and thus more salient, than those at the top.

For all visual scenes that do not contain red objects (G1,…,G4), the model does not bring any object into the attentional foreground. Please note that in these cases, it keeps on searching; it does not detect that there is no red object to be found in the scene. This is the expected behavior since the model does not have a mechanism to detect this case.

[3] See Appendix C.

ID	# objects	# potential targets	# moving distractors	+/−	description
G19	0	0	0	−	no object in the scene
G20	4	0	0	−	no moving object in the scene
G21	4	0	1	−	motion, but not in target direction
G22	4	1	0	+	motion in target direction
G23	4	2	0	+	multiple moving in target direction
G24	4	1	1	+	motion in multiple directions

Table 4.2: Tests of grounding tasks in which the phrase is "the object moving to the right". The tests marked in blue are described in detail in the text.

Motion direction

For a second set of tests, the task input corresponds to the phrase "the object moving to the right" and thus only specifies its motion direction. In these tests, the user activates the target motion memory node for the motion direction RIGHTWARD and subsequently activates the 'ground object process'. Table 4.2 lists all of these tests (G19,…,G24). It is structured analogously to Table 4.1, but here, target objects are those moving rightward and distractor objects are stationary or moving into a different direction. The two tests marked in blue are described in detail in the following sections.

These tests show the same results as those conducted by specifying the color of the target object: the model only brings objects into the attentional foreground that are specified by the phrase (in this case, objects moving rightward). It does not bring distractor objects into attentional foreground (G20, G21, G24) and if multiple potential target objects are present in the scene, it selects one of them (G23).

Example with a unique target

The following describes the processes that unfold in the model when it grounds a target that is uniquely identifiable by a color that is specified (test G5, marked blue in Table 4.1). In this example, the model grounds the phrase "the red object" in a visual scene that contains exactly one red object and three objects of other colors; all objects in the scene are stationary.[4] Figure 4.1 shows how the activation in the model evolves over the course of the test. The panel in the top row shows the activation level of the intention node and the CoS node of the 'target process' over continuous time, where the activation

[4]The video used in this test is video 4.03. Please refer to Appendix C, in particular Figure C.41 for more information.

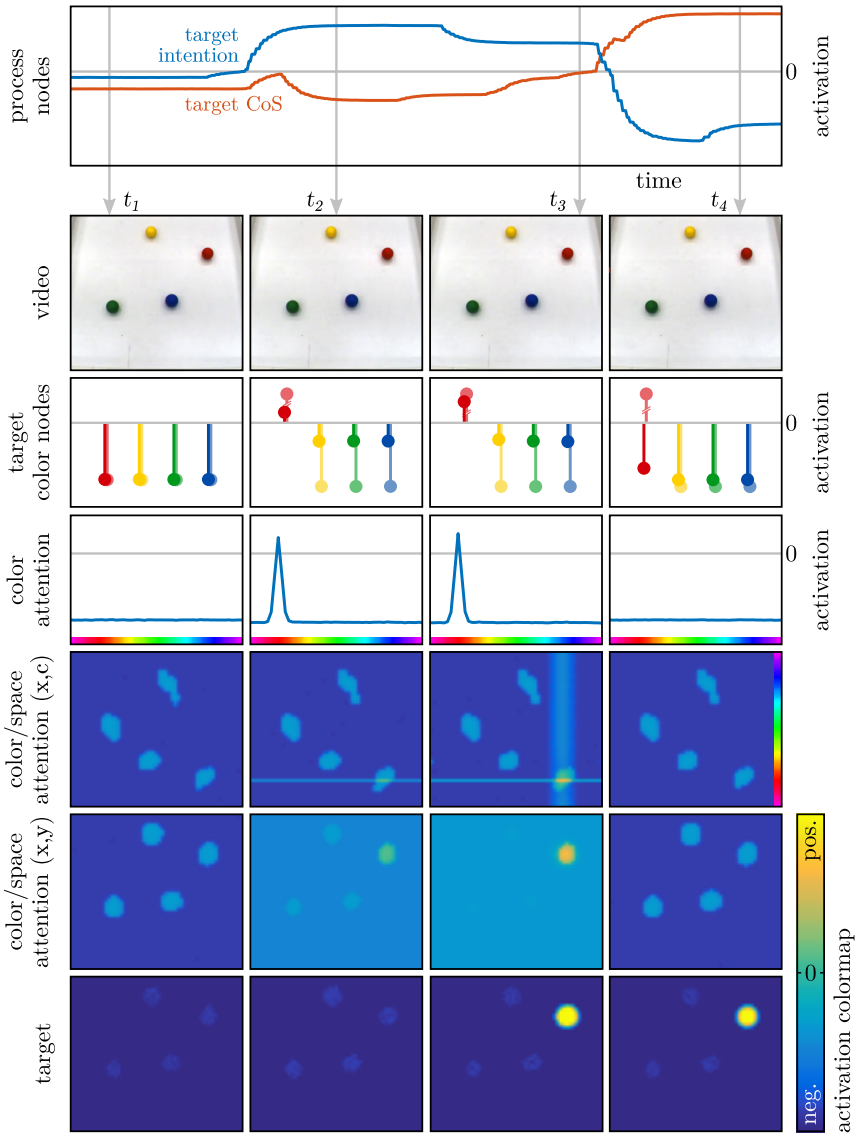


FIGURE 4.1: Grounding the phrase "the red object" in a scene with a unique target. The first row shows the time course of two nodes of the target process over continuous time. Rows 2–7 show snapshots at four points $t_1, \ldots, t_4$ in time (four columns). In the third row, the colors of the bars match the colors that the nodes represent. Solid bars show activation of the target color production nodes, transparent bars show activation of the target color memory nodes. The activation shown in the bottom three rows is color-coded using the colormap on the bottom right. See text for more detail.

threshold is marked with a (horizontal) gray line at zero. All panels below illustrate the state of the model at four points $t_1, \ldots, t_4$ in time (four columns). The second row from the top shows the video input. Since none of the objects in the scene is moving, the input does not change noticeably over time. The third row from the top shows the activation of the target color production nodes (illustrated in solid colors) and the target color memory nodes (illustrated in transparent colors).[5] The fourth row from the top shows the activation of the color attention field over its color dimension c. The fifth and sixth row from the top show the activation of the three-dimensional color/space attention field. Activation levels are denoted in a color-code defined by the colormap shown in the bottom right of Figure 4.1. The activation is shown twice, projected onto the horizontal space x and the color dimension c (fifth row)

[5]The activation of the target color memory node for the color RED at $t_2, \ldots, t_4$ is too large to fit into the panel at the chosen scale. This is denoted by two diagonal lines breaking up the bar that shows the activation level.

and onto the two spatial dimensions x, y (sixth row). The projections are computed by taking the maximum along the dimension not shown. The last row shows the color-coded activation of the target field over the spatial dimensions x, y.

At the beginning of the test, the model is already being simulated and is receiving visual input from the video. However, the activation in all fields and nodes of the model is below threshold. The only exception is the color/space perception field, which has four stable peaks of activation, each representing the spatial position and color of one of the objects in the scene. Figure 4.1 shows that at time t_1 all fields and nodes are below threshold. The plot of the color/space attention field shows four bumps of activation that are all below threshold. These are due to subthreshold input from the color/space perception field.

The user then introduces task input by manually giving excitatory input to the target color memory node for the color RED, thereby activating it. This input corresponds to the user saying the phrase "the red object". He then gives excitatory input to the prior intention node of the 'ground object process', which initiates the grounding of the phrase. This activates the target process on the next lower hierarchical level, which in turn activates multiple processes on lower hierarchical levels that each control an aspect of the grounding process of the target object.[6] Figure 4.1 shows that at time t_2 the 'target process' is active as the activation of its intention node (blue line in the top panel) is above threshold. This node directly gives homogeneous input to all target color production nodes. The input activates the target color production node for the color RED, because that node also gets subthreshold input from the active target color memory node. At time t_2, both nodes are active (third row from the top, second column from the left). The active target color production node projects activation into the color attention field and creates a peak there (fourth row, second column). The peak forms at the position coding for red colors because of the patterned synaptic connections between the target color production node for the color RED and the color attention field. The color attention field projects subthreshold input into the three-dimensional color/space attention field. Along the color dimension c, this input is localized and centered around the position coding for red colors. This shows up as a horizontal line of activation in the activation plot of the color/space attention field (fifth row, second column). Along the two spatial dimensions x, y, the input is homogeneous (sixth row, second column). The input overlaps with the subthreshold bump that codes for the spatial position and color of the red object in the scene. Due to that overlap, the activation in the color/space attention field at the spatial position of the red object rises above thresh-

[6] See Figure 3.14.

old. The color/space attention field projects its activation onto the multi-peak spatial attention field and the selective spatial attention field (both not shown in Figure 4.1), which are defined over the spatial dimensions x, y. Both fields are in a dynamic regime that enables them to form peaks. This is because the 'target process' has activated the 'spatial attention process', which gives homogeneous excitatory input to the selective spatial attention field.

At time t_3, a peak has formed in both the multi-peak spatial attention field and the selective spatial attention field. The latter projects activation back into the color/space attention field. Along the spatial dimensions x, y, this input is localized and centered onto the spatial position of the red object. Along the color dimension c, the input is homogeneous. This shows up as a thick vertical line of activation in the activation plot of the color/space attention field (fifth row, third column in Figure 4.1). The selective spatial attention field projects into the target field and forms a peak there (last row, third column). At this point in time, the phrase "the red object" is grounded: the red object in the scene has been identified and a stable representation of its features (spatial position and color) has been formed. The CoS of all active processes activated by the target process are met: there are peaks in the color attention field (CoS of the feature process), in the selective spatial attention field (CoS of the spatial attention process), in the target IOR field (CoS of the target IOR process), and in the target field (CoS of the target field process); furthermore, one of the target color memory nodes is active (CoS of the target memory node process). Since the CoS memory node of all these processes is active, the CoS node of the target process reaches the threshold (red line in the top panel in Figure 4.1 at t_3), activates the CoS memory node of the process, which in turn inhibits its intention node. This deactivates all processes on the lower levels and returns them to their initial state. The active CoS memory node of the target process activates the CoS node (and subsequently the CoS memory node) of the 'ground object process'. This deactivates the target process.

At time t_4, the activation level of most nodes and fields of the model has returned below threshold. However, the elements that constitute the grounding of the phrase "the red object" remain above threshold: the target color memory node for the color RED is active and the target field has a peak at the spatial position of the red object.

Example with multiple potential targets

The following example covers a case in which there are multiple objects in the scene that fit the description (test G23 in Table 4.2).

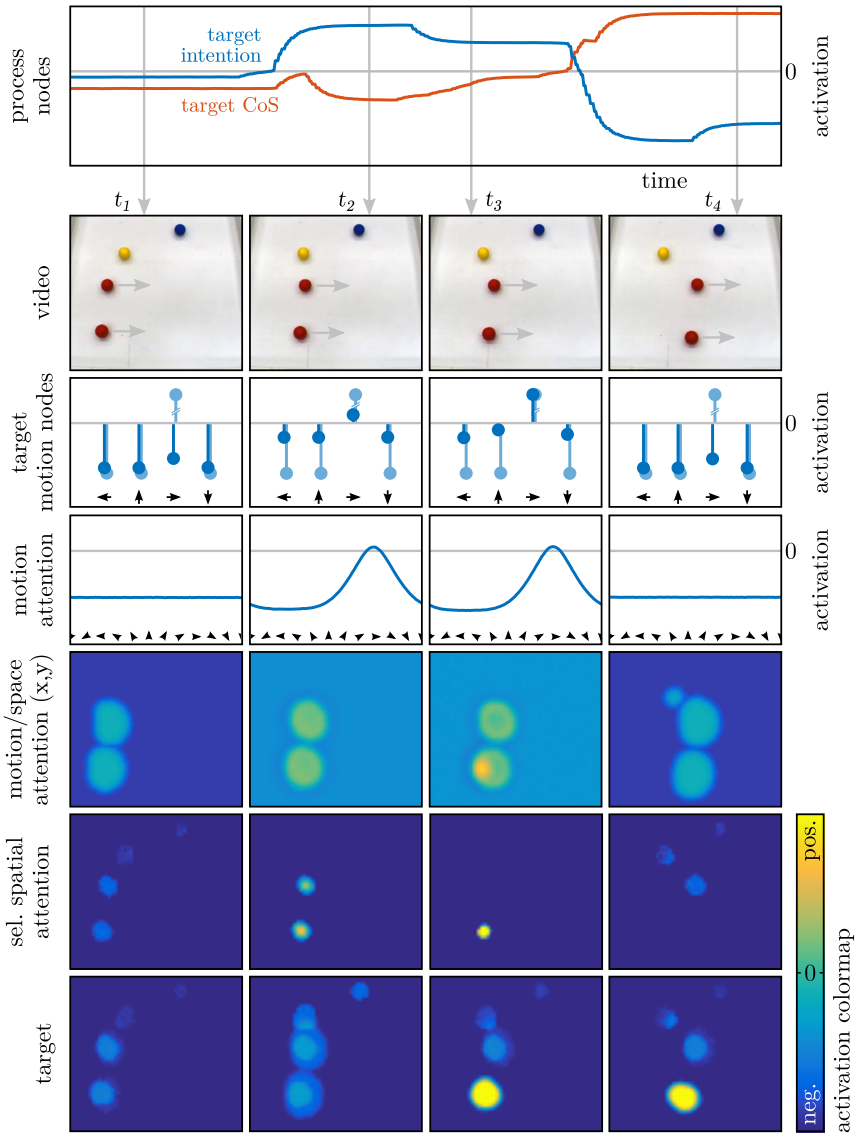


FIGURE 4.2: Grounding the phrase "the object moving rightward" in a scene with two matching objects. The figure can be interpreted analogously to Figure 4.1. In the third row, the concepts of motion direction that the nodes represent are indicated by the arrows underneath the activation bars. Analogous to before, solid bars show the activation of the target motion production nodes and transparent bars show the activation of the target motion memory nodes. See text for more detail.

Additionally, the example shows that the model can also deal with the feature dimension of motion direction. Thus, in this test, the task input corresponds to the phrase "the object moving rightward". The visual scene consists of two objects that are moving to the right as well as two stationary objects.[7]

Figure 4.2 shows how the activation in the model evolves over the course of the test. It is structured analogously to Figure 4.1 with a few notable differences. In the third row, the activation of the target motion memory nodes (solid colors) and target motion production nodes (transparent colors) is shown. The concepts of motion direction that the nodes represent are indicated by the arrows underneath the activation markers (from left to right: LEFTWARD, UPWARD, RIGHTWARD, and DOWNWARD movement). In the fourth row, the activation of the motion attention field is shown.

[7]The video used in this test is video 4.15c. Please refer to Appendix C, in particular Figure C.71 for more information.

In the fifth row, the activation of the motion/space attention field is shown, color-coded according to the color map in the bottom right of Figure 4.2. Here, the three-dimensional activation of the field is projected onto the spatial dimensions x, y by taking the maximum along the motion direction dimension ϕ. In the sixth line, the activation of the selective spatial attention field is shown, also color-coded.

At the beginning of the test, the model is already being simulated and fed with video input. The user supplies the task input by giving excitatory input to the target motion memory node that represents the movement direction RIGHTWARD; the node is activated by that input. At time t_1 the node is already active (third row, first column in Figure 4.2), while the target motion production node has not yet been activated. Apart from this active node, the model is thus in its initial state with the activation of most fields below threshold. The exception are the three-dimensional color/space perception field and the motion/space perception field (both not shown in Figure 4.2). The former features four peaks, representing the spatial position and color of the four objects in the scene. The latter features two peaks, representing the spatial position and motion direction of the two moving objects. This activation is projected onto the motion/space attention field, which remains below threshold but has two subtreshold bumps of activation at the spatial positions of the two moving objects (fifth row, first column). Similar input is also projected onto the selective spatial attention field (sixth row, first column) and target field (last row, first column), where it creates subthreshold bumps as well. The selective spatial attention field also receives subthreshold input from the color/space perception field, localized and centered on all objects, including the stationary ones. The input is barely visible in Figure 4.2 (sixth row, first column).

The user then gives excitatory input to the prior intention node of the 'ground object process', activating that node. This initiates the process of grounding the phrase "the object moving rightward". This process is analogous to the one explained in the previous example but uses those fields and nodes in the model that code for motion direction, instead of those coding for color. The 'ground object process' activates the target process on the lower level, which in turn activates processes on lower levels. The target process gives homogeneous input to the target motion production nodes, activating the node that represents the concept RIGHTWARD (third row, second column in Figure 4.2) because it also receives input from the corresponding target motion memory node. This node projects its activation onto the motion attention field through patterned synaptic connections, creating a peak at the position coding for the motion direc-

tion RIGHTWARD (fourth row, second column). The (suprathreshold) activation in this field is projected onto the three-dimensional motion/space attention field. Along the motion direction dimension ϕ, this input is localized and centered on the position coding for rightward motion. Along the spatial dimensions x, y, the input is homogeneous. It overlaps with the two subthreshold bumps that are localized at the positions of the two moving objects and brings these bumps above threshold (fifth row, second column). The activation in the motion/space attention field is projected onto the multi-peak spatial attention field and the selective spatial attention field. As in the previous example, the selective spatial attention field is in a dynamic regime where it can form a peak. At time t_2 it is in the process of making a selection decision: two bumps of activation are visible, with the lower one being slightly stronger (sixth row, second column). The multi-peak spatial attention field reacts slower to change than the selective spatial attention field because it is parameterized with a slower time scale.[8] At time t_2, the multi-peak spatial attention field has not yet formed a peak (not shown in Figure 4.2) and is thus not giving strong localized input to the target field, which remains below threshold (last row, second column).

[8] This means that the parameter τ_{ASm} in the differential equation of the multi-peak spatial attention field is larger than τ_{AS} of the selective spatial attention field.

At time t_3, the selective spatial attention field has made a selection decision by forming a single peak at the spatial position of the lower moving object and inhibiting the activation in the rest of the field (sixth row, third column). The field projects back into the motion/space attention field, highlighting the selected object there (fifth row, third column). Since the selective spatial attention field has a peak, it makes the multi-peak spatial attention field selective as well by giving it local excitatory input as well as global inhibitory input. The single remaining peak is projected into the target field, representing the selected target object (last row, third column). At this point in time, the model has grounded the phrase "the object moving rightward" by finding matching objects in the scene, making a selection decision between two candidates, and forming a representation of the spatial position of the selected object in the target field. As in the previous example, the CoS of all active processes are met. Figure 4.2 shows the activation of the CoS node of the target process in the top row (red line), which crosses the threshold between t_3 and t_4.

At time t_4, most of the nodes and fields return below threshold because the processes that initially activated them have deactivated. What remains are an active target motion memory node for the motion direction RIGHTWARD (third row, fourth column), and peaks in the target field (last row, fourth column) as well as the target IOR field (not shown). Please note that the peak in the target field tracks the moving spatial position of the selected target object. This is due

to excitatory input it always receives from the color/space perception field and the motion/space perception field. The same is true for the target IOR field.

Example with no target

The following example covers a case in which no object in the scene fits the given description (test G21 in Table 4.2). As in the last example, the task input corresponds to the phrase "the object moving rightward". However, the visual scene consists of three static objects as well as an object that is moving leftward, that is, into a different direction than searched for.[9]

[9]The video used in this test is video 4.1a. Please refer to Appendix C, in particular Figure C.37 for more information.

Figure 4.3 shows how the activation in the model evolves over the course of the test. It is analogous to Figure 4.2 with the difference that plots for only three points $t_1, \ldots, t_3$ in time are shown (three columns) and that the activation of the three-dimensional motion/space attention field is shown here projected onto the horizontal spatial dimension x and the motion direction dimension ϕ (by taking the maximum along the vertical spatial dimension y).

The processes that happen in the model are also largely analogous to the last example. At time t_1, activation in the model is below threshold except for the perceptual fields (not shown) as well as the target motion memory node for the motion direction RIGHTWARD (third row, first column in Figure 4.3) that the user activated. At time t_2, the corresponding target motion production node is active as well (third row, second column) because the user has initiated the grounding process by activating the 'ground object process'. The motion attention field has a peak at the position coding for the motion direction RIGHTWARD (fourth row, second column). The activation in this field projects onto the three-dimensional motion/space attention field. Along the spatial dimensions x, y, the input is homogeneous; along the motion direction dimensions ϕ, it is localized at the position coding for the motion direction RIGHTWARD. In Figure 4.3 this is shown as a horizontal bar of activation (fifth row, second column). However, this bar does not overlap with the subtreshold bump of activation that is in the field as well. This is because the subtreshold bump codes for the spatial position and motion direction of the green object moving *leftward* in the scene. Since it is moving leftward, the bump is localized in the field at positions coding for the motion direction LEFTWARD. As no bump overlaps with the input from the motion attention field, the motion/space attention field does not form a peak. Therefore, the model does not bring any object into attentional focus. For the rest of the example, the activation in the model does not undergo any significant change. The peaks in the color/space perception field and the motion/space per-

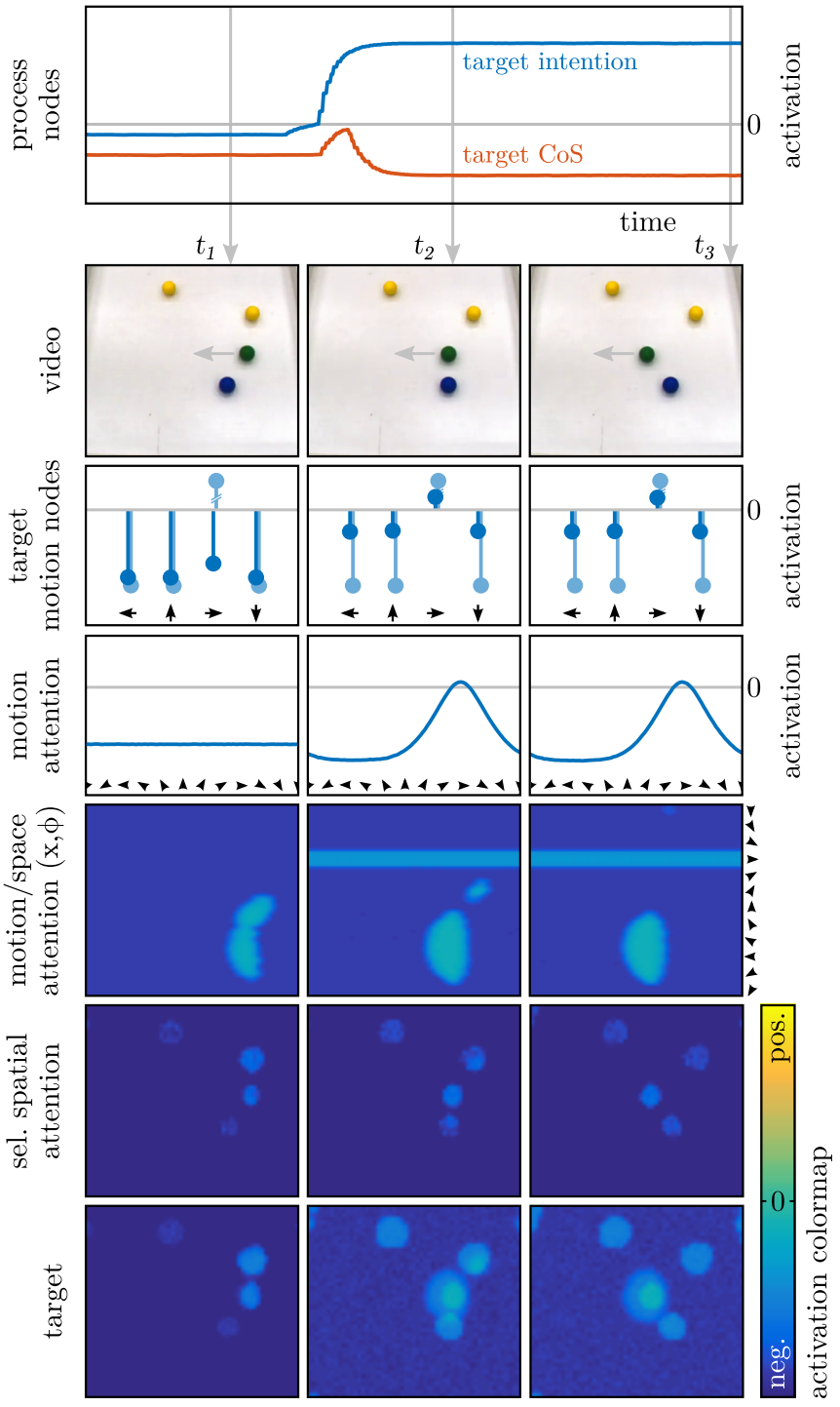


FIGURE 4.3: Grounding the phrase "the object moving rightward" in a scene where no such object exists. The figure can be interpreted analogously to Figure 4.2. See text for more detail.

ception field as well as the subtreshold bumps in other fields coding for spatial position track the spatial position of the moving object (visible in the plots between t_3 and t_4), but it does not create a peak anywhere.

Since no object can be found in the scene that matches the description, the CoS of most active processes cannot be met. This is

shown, for instance, by the activation of the CoS node of the target process (red line in the top panel), which stays below threshold. Nevertheless, the result of the example is positive because the behavior of the model is as expected. To be able to detect and react to the non-existence of an object would require an additional mechanism, which is not part of the model.

4.1.2 Feature conjunctions

The following tests establish that the attentional system and saliency system of the model work when the object that is to be searched for is specified by a conjunction of multiple features. In all tests, the task input given to the model specifies two features, color and motion direction, of a single object. The task input corresponds to the phrase "the red object moving rightward". In all tests, the user activates the target color memory node for the color RED, the target motion memory node for the motion direction RIGHTWARD, and subsequently the prior intention node of the 'ground object process'. Table 4.3 lists all of these tests (G25,…,G57). It is similar to Tables 4.1 and 4.2 and summarizes information about the visual scene that was presented to the model in each test. In the columns from left to right, it shows the identifier of the test, the total number of objects in the scene, the number of red objects in the scene, the number of moving objects, the number of objects moving rightward, the number of red objects moving rightward (potential target objects), whether or not a target can be found (marked with '+' or '−', respectively), as well as a language description of the visual scene. The tests marked in blue are described in detail in the following sections.

As before, the result of all tests show that the model only brings objects into the attentional foreground that are specified by the phrase (in this case, red objects that move rightward). In all visual scenes that contain a red object moving rightward (those marked with '+'), the model is able to bring it into attentional foreground and thereby ground the phrase. This works irrespective of the number of distractor objects in the scene and whether they move or are stationary (e.g., G53). In case there are multiple potential targets in the visual scene (G53,G54,G57), the model selects one of the objects while ignoring any distractor objects. As before, in making the selection the model often prefers objects that are closer to the bottom of the visual scene (due to perspective distortion in the video).

The tests also systematically check whether objects that do not match the description or that match it only partially are brought into attentional foreground. This could, for instance, be a stationary red object (G31), a red object moving leftward (G35), or a green ob-

ID	# objects	# color matches	# moving	# motion matches	# potential targets	+/−	description
G25	0	0	0	0	0	−	no object in the scene
G26	4	0	0	0	0	−	incorrect colors, no motion
G27	4	0	1	1	0	−	incorrect color, correct direction
G28	4	0	1	0	0	−	incorrect color, incorrect direction
G29	4	0	2	0	0	−	incorrect colors, incorrect directions
G30	4	0	2	1	0	−	incorrect colors, correct/incorrect directions
G31	4	1	0	0	0	−	correct color, no motion
G32	4	1	1	1	0	−	incorrect color, correct direction
G33	4	1	1	0	0	−	incorrect color, incorrect direction
G34	4	1	1	1	1	+	correct color, correct direction
G35	4	1	1	0	0	−	correct color, incorrect direction
G36	4	1	2	0	0	−	incorrect colors, incorrect directions
G37	4	1	2	1	0	−	incorrect colors, correct/incorrect directions
G38	4	1	2	1	1	+	correct color, correct direction; correct color, incorrect direction
G39	4	1	2	2	1	+	correct/incorrect color, correct direction
G40	4	1	2	0	0	−	correct/incorrect color, incorrect direction
G41	4	1	2	1	0	−	correct color, incorrect direction; incorrect color, correct direction
G42	4	2	0	0	0	−	multiple correct colors, no motion
G43	4	2	1	1	0	−	incorrect color, correct direction
G44	4	2	1	0	0	−	incorrect color, incorrect direction
G45	4	2	1	1	1	+	correct color, correct direction
G46	4	2	1	0	0	−	correct color, incorrect direction
G47	4	2	2	2	0	−	incorrect colors, correct directions
G48	4	2	2	0	0	−	incorrect colors, incorrect directions
G49	4	2	2	0	0	−	correct/incorrect colors, incorrect directions
G50	4	2	2	1	0	−	correct color, incorrect direction; incorrect color, correct direction
G51	4	2	2	1	1	+	correct color, correct direction; incorrect color, incorrect direction
G52	4	2	2	2	1	+	correct/incorrect color, correct direction
G53	4	2	3	2	2	+	correct colors, correct direction; incorrect color, incorrect direction
G54	4	2	3	3	2	+	correct/incorrect colors, correct direction
G55	4	2	2	0	0	−	only correct colors moving, incorrect directions
G56	4	2	2	1	1	+	only correct colors moving, correct/incorrect directions
G57	4	2	2	2	2	+	only correct colors moving, correct direction

Table 4.3: Tests of grounding tasks in which the phrase is "the red object moving rightward". The tests marked in blue are later described in detail. See text for details on how to interpret this table.

ject moving rightward (G27). Particularly interesting is the case in which multiple objects match the description partially, for instance a scene in which a red object is moving leftward and a green object is moving rightward (G41). This case shows that all matching features must also belong to the same object—in the model they are bound over the spatial position of the object. The result of the tests show that for all visual scenes that do not contain red objects moving rightward (those marked with '–'), the model does not bring any object into the attentional foreground. As explained for previous tests, in these cases the model keeps on searching; it does not detect that there is no target object to be found in the scene. This is the expected behavior since the model does not have a mechanism to detect this case.

Example with a unique target

The following describes the processes that unfold in the model when it grounds an object that is uniquely identifiable by a conjunction of features (test G56 in Table 4.3). In this example, the model grounds the phrase "the red object moving rightward" in a visual scene that contains exactly one such object, another red object that is moving leftward, and two stationary objects (blue and yellow).[10] Figure 4.4 shows how the activation in the model evolves over the course of the test. It is structured analogously to previous figures (Figures 4.1, 4.2, and 4.3) but shows the activation of the nodes and fields that code for color as well as those that code for motion direction. Both the activation of the color/space attention field (fifth row in Figure 4.4) and the motion/space attention field (sixth row) is shown projected onto the horizontal spatial dimension x and respective feature dimension of the field (color c and motion direction ϕ, respectively) by taking the maximum along the vertical spatial dimension y.

[10]The video used in this test is video 4.15b. Please refer to Appendix C, in particular Figure C.70 for more information.

The processes that unfold in this test are very similar to those in the first example (test G5), where a uniquely identifiable object is present in the scene. The difference is that a combination of features is used here in the phrase that describes the target object. Thus, at the beginning of the test, the user activates both the target color memory node for the color RED as well as the target motion memory node for the motion direction RIGHTWARD. At time t_1, the nodes are active (transparent bars in the third and fourth row, first column, of Figure 4.4). When the user initiates the process of grounding by activating the prior intention node of the 'ground object process', the corresponding target color production nodes and target motion production node also get activated (solid bars, third and fourth row, second column).

The active target color production node projects its activation

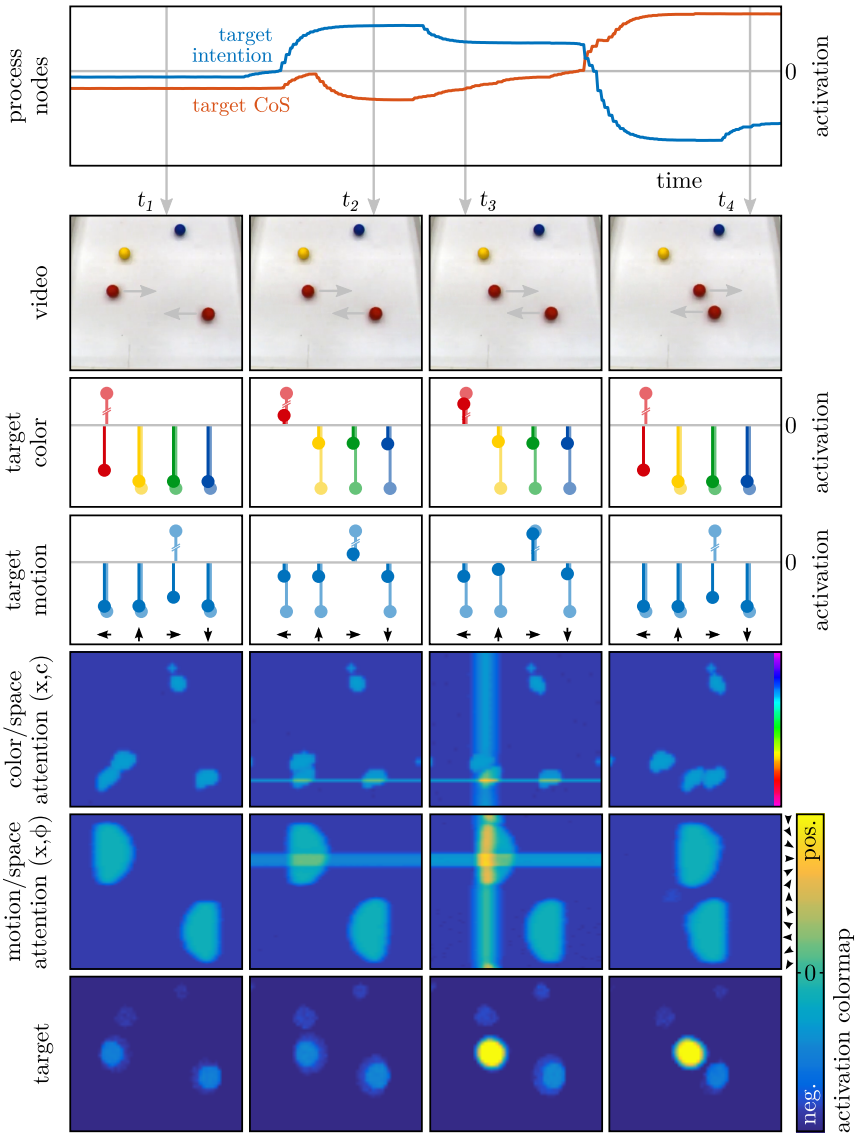


FIGURE 4.4: Grounding the phrase "the red object moving rightward" in a scene with a unique target. See text for more detail.

via patterned synaptic connections into the color attention field (not shown in Figure 4.4), creating a peak at the position coding for red colors. The color attention field projects into the color/space attention field. This input is localized along the color dimension c and centered at the position coding for red colors. Along the spatial dimensions x and y, it is homogeneous. The input is shown as a fine horizontal line in the plot in the fifth row and second column of Figure 4.4. It overlaps with the subthreshold bumps that code for the spatial position and color of *both* of the red objects in the scene.

Analogously, the target motion production node projects activation into the motion attention field (not shown in Figure 4.4), creating a peak at the position coding for the motion direction RIGHTWARD. The motion attention field projects into the motion/space attention field, where the input is localized along the dimension of

motion direction ϕ; along the spatial dimensions x, y it is homogeneous. The input is shown as a thick horizontal line in the plot in the sixth row and second column of Figure 4.4. It overlaps with the subthreshold bump that codes for the spatial position and motion direction of the object moving rightward, not with the bump that codes for the other red object, because that one is moving leftward.

The selective spatial attention field and multi-peak spatial attention field (both not shown in Figure 4.4) are parameterized such that they only form a peak if they receive input from those attention fields whose features are specified in the task input, in this case from the color/space attention field and the motion/space attention field. Along the spatial dimensions, the fields receive localized input at the positions of the two red objects. However, only the spatial position of the red object that is moving rightward receives localized input from both the color/space attention field *and* the motion/space attention field and can form a peak in the spatial attention fields. At time t_3, they have each formed a peak. The multi-peak spatial attention field is projecting its activation into the target field (last row, third column). The selective spatial attention field is projecting its activation back into the color/space attention field and motion/space attention field. This input is shown in Figure 4.4 by the vertical bar of activation (fifth and sixth row, third column).

At this point in time, the phrase "the red object moving rightward" has been grounded by the model. All processes that were active during the grounding processes are deactivated because their CoS is met. At time t_4, most of the activation in the model has returned below threshold. The target color memory node for the color RED and the target motion memory node for the motion direction RIGHTWARD remain active (third and fourth row, fourth column); so do the peak in the target field (last row, fourth column) and a peak in the target IOR field (not shown).

Example with multiple potential targets

The following example covers a case in which the description of the target object consists of a conjunction of features and there are multiple objects in the scene fitting the description (test G53 in Table 4.3). In this example, the model grounds the phrase "the red object moving rightward" in a visual scene that contains two such objects, as well as a yellow object that is moving leftward, and a stationary blue object.[11] Figure 4.5 shows how the activation in the model evolves over the course of the test. It is structured analogously to Figure 4.4.

The processes that unfold in this test are very similar to those in the last example (test G56), where a uniquely identifiable object is

[11]The video used in this test is video 4.14e. Please refer to Appendix C, in particular Figure C.67 for more information.

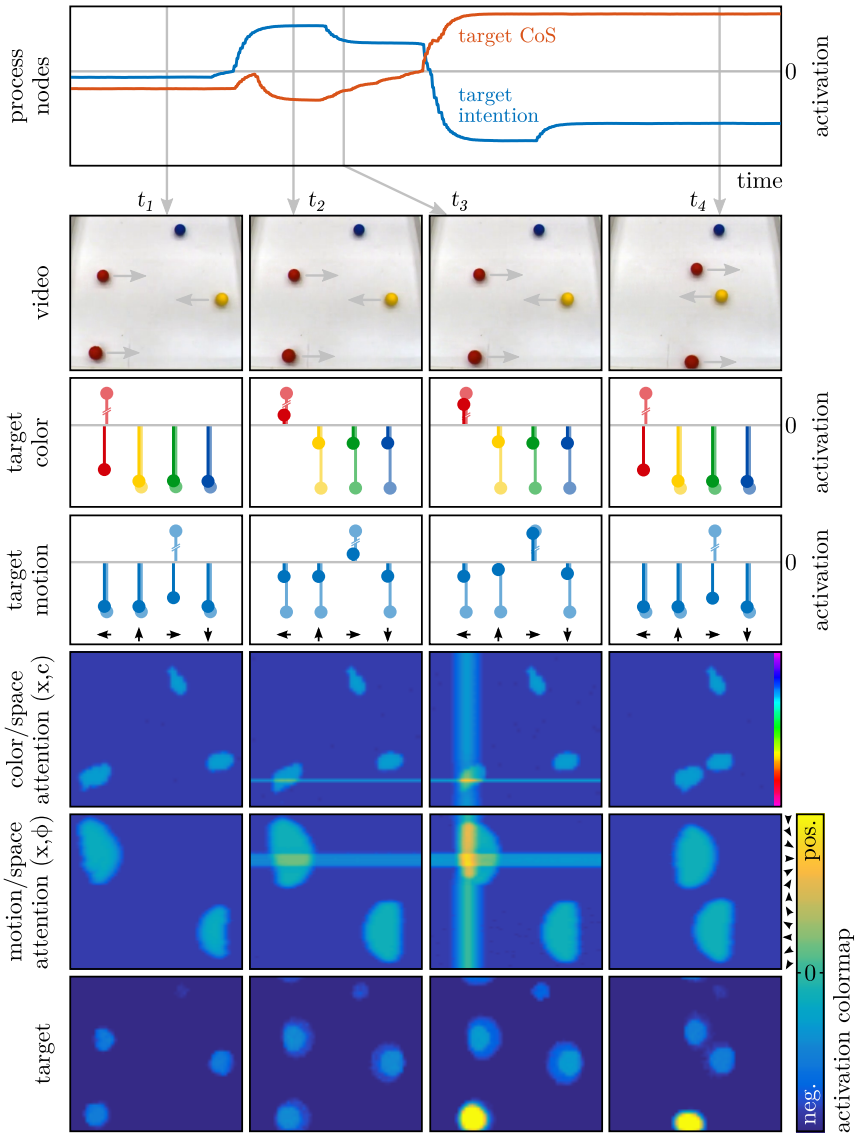


FIGURE 4.5: Grounding the phrase "the red object moving rightward" in a scene with two possible target objects. See text for more detail.

present in the scene. This is because the selective spatial attention field is always in a dynamic regime where it can form peaks when the target object is grounded. This means that the spatial attention mechanism will only ever form a single peak. Because of this, the processes in the model do not differ substantially if the scene contains one or multiple objects that fit the description.

At the beginning of the test, the user activates both the target color memory node for the color RED as well as the target motion memory node for the motion direction RIGHTWARD. The user initiates the grounding process by activating the prior intention node of the 'ground object process', the corresponding target color production nodes and target motion production node also get activated (third and fourth row, second column). All processes evolve analogously to the previous example.

At time t_3, the selective spatial attention field has made a selection decision and formed a peak that is centered on the lower of the two red objects. This peak is projected onto the target field (last row, third column of Figure 4.5) and into the color/space attention field and motion/space attention field (vertical lines of activation in the fifth and sixth row, third column).

At time t_4, most activation in the model has returned below threshold, except for activation in the target color memory node for the color RED (third row, fourth column), in the target motion memory node for the motion direction RIGHTWARD (fourth row, fourth column), the peak in the target field that is centered on and tracking the position of the lower red object in the scene (last row, fourth column), and the target IOR field (not shown). The object described in the phrase "the red object moving rightward" has been grounded in the scene.

Example with no target (only motion direction matches)

This example covers a case in which no object in the scene fits the given description but one of the objects matches the description partially (test G27 in Table 4.3). As in the last example, the task input corresponds to the phrase "the red object moving rightward". The visual scene consists of three static objects as well as an object that is moving rightward; however, that object is green rather than red.[12]

[12]The video used in this test is video 4.1b. Please refer to Appendix C, in particular Figure C.38 for more information.

Figure 4.6 shows how the activation in the model evolves over the course of the test. It is analogous to Figure 4.4.

At the beginning of the test, the user activates both the target color memory node for the color RED as well as the target motion memory node for the motion direction RIGHTWARD. The user initiates the grounding process by activating the prior intention node of the 'ground object process', the corresponding target color production nodes and target motion production node also get activated (third and fourth row, second column). All processes evolve analogously to the previous example. However, at time t_2, the input into the color/space attention field (horizontal line of activation in the fifth row, second column) does not overlap with any subtreshold bump representing an object. This is because the input is localized and centered on the position coding for red colors but there are no red objects in the scene. While the corresponding input into the motion/space attention field (horizontal line of activation in the sixth row, second column) does overlap with a subtreshold bump (the one representing the green object moving rightward) and forms a peak in that field, this is not sufficient to form a peak in the selective spatial attention field or multi-peak spatial attention field (both not

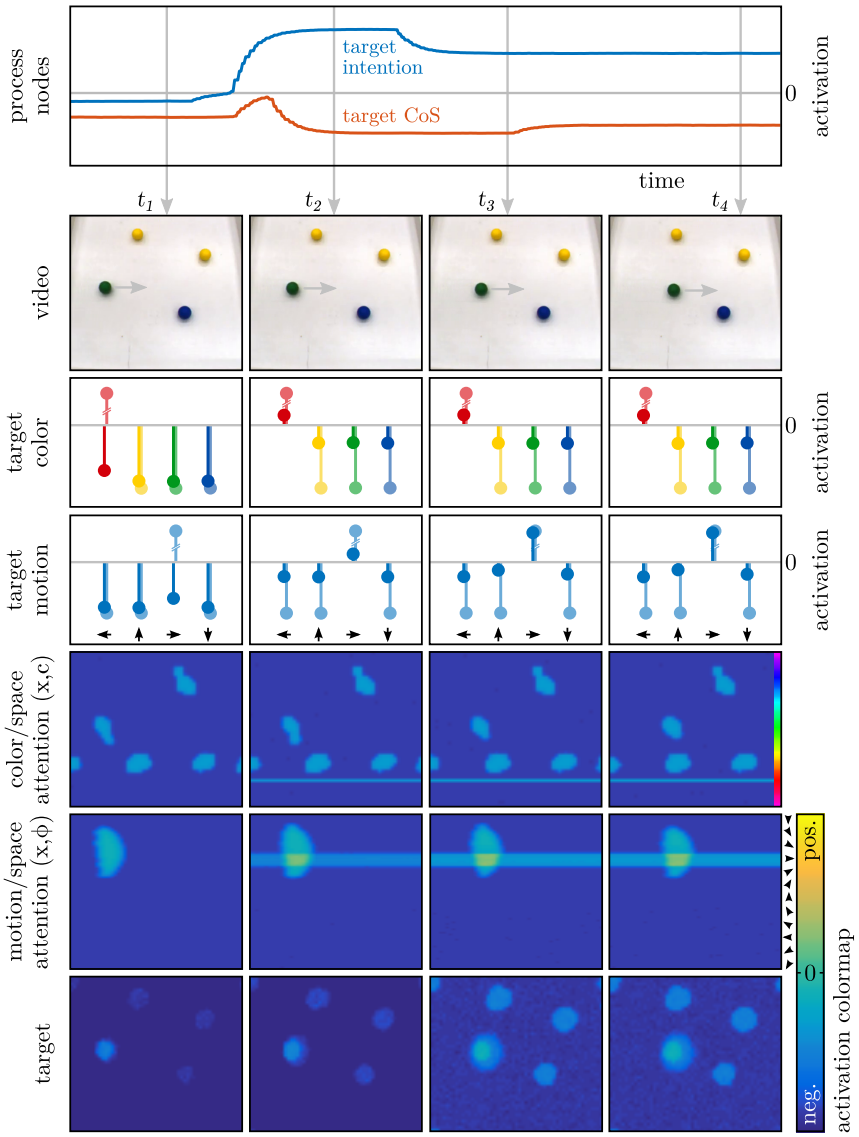


FIGURE 4.6: Grounding the phrase "the red object moving rightward" in a scene where no such object exists but one object matches the description partially. See text for more detail.

shown). Whenever a feature is specified in a phrase, that feature is required to be found in order for the spatial attention fields to form a peak.[13]

Since the object moving rightward is green instead of red, the model does not bring it into attentional foreground and does not ground it as the target object. Thus, until the end of the test, the subtreshold bumps in the fields track the moving position of the green object, but do not form a peak. As before, the model does not detect that there is no object in the scene that fits the description and instead keeps searching indefinitely. Since such a detection would require an additional mechanism, the behavior of the model is expected.

[13] This is implemented by inhibitory connections from the color attention field and motion attention field to the selective spatial attention field and multi-peak spatial attention field. See Section 3.2.2.

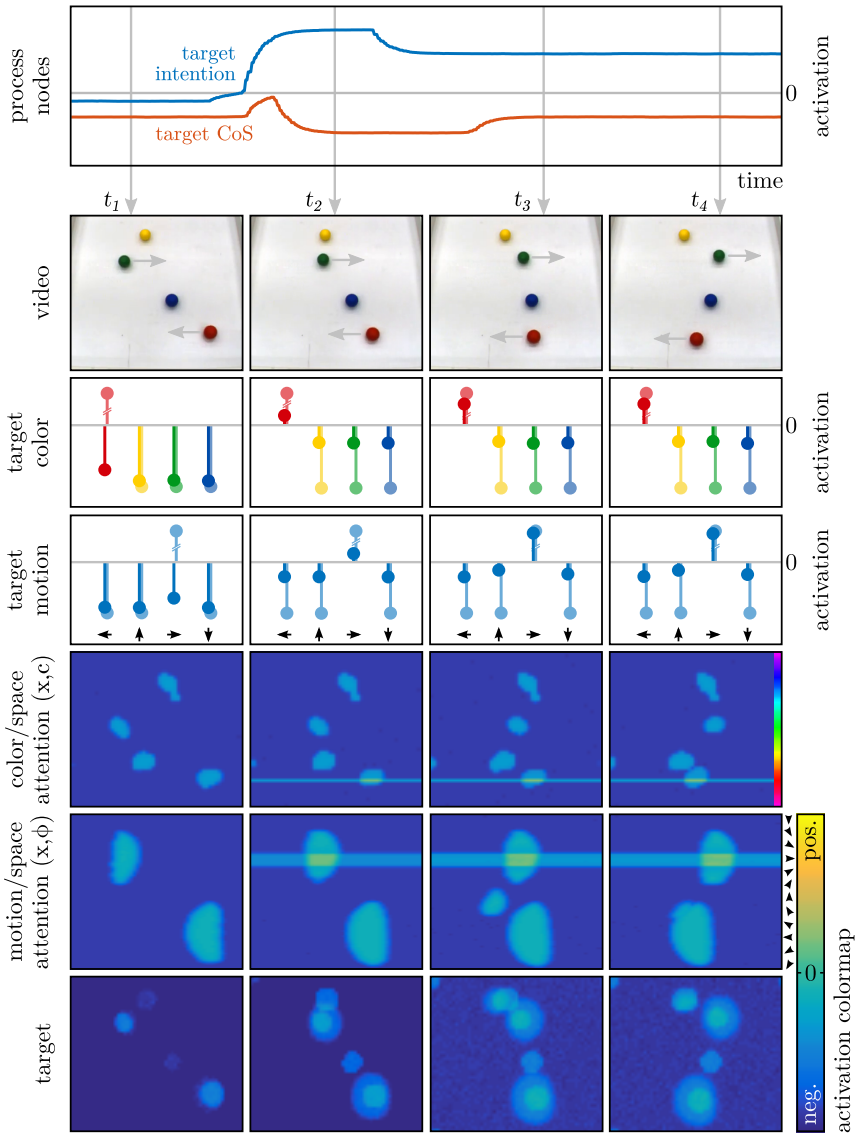


FIGURE 4.7: Grounding the phrase "the red object moving rightward" in a scene where no such object exists but two objects match the description partially. See text for more detail.

Example with no target (matching features not bound)

The following example covers a case in which no object in the scene fits the given description but two of the objects match the description partially (test G41 in Table 4.3). As in the last example, the task input corresponds to the phrase "the red object moving rightward". However, the visual scene contains only a red object moving leftward and a green object moving rightward; two stationary objects (blue and yellow) are also present.[14]

Figure 4.7 shows how the activation in the model evolves over the course of the test. It is analogous to Figure 4.4.

The processes that unfold in the model are very similar to the last example. The difference is that at time t_2 both the color/space attention field (fifth row, second column in Figure 4.7) and the mo-

[14]The video used in this test is video 4.6d. Please refer to Appendix C, in particular Figure C.49 for more information.

tion/space attention field (sixth row, second column) form a peak because input that is shown here as horizontal lines of activation overlaps with subthreshold bumps that represent objects. However, these peaks are at different spatial positions because they are associated with different objects. The input for the color/space attention field overlaps with the subthreshold bump that is associated with the red object moving leftward, while the input for the motion/space attention field overlaps with the subthreshold bump associated with the green object moving rightward. Since the peaks are at different spatial positions, their input into the selective spatial attention field and multi-peak spatial attention field does not overlap and thus does not produce peaks there. The model does not bring any of the objects in the scene into attentional foreground for the rest of the test.

This example shows that it is not sufficient that all features specified in the phrase are found in the scene. It is also required that these features belong to the same object and are bound by the same spatial position. If not, the objects in question are not grounded.

4.1.3 Relations between objects

The tests described next establish that the model can ground objects based on the spatial relations and movement relations they have with respect to other objects in the scene. In all tests, the task input given to the model specifies two objects using a single feature (here, color) for each of them, as well as a relation between the two objects. Unlike in the previous examples, there are always objects in the scene that match the color descriptions. The tests thus vary, for instance, whether the objects adhere to the specified relation, whether they move, and whether there is more than one pair of objects that may match the relation.

For a first set of tests, the task input corresponds to the phrase "the red object to the left of the green object" and both the target object (red) and the reference object (green) are uniquely identifiable in the scene. The relation that is specified between the objects is thus a spatial relation that is defined by their relative position. In all tests, the user activates the target color memory node for the color RED, the reference color memory node for the color GREEN, and the spatial relation memory node for the spatial relation TO THE LEFT OF. The user subsequently activates the prior intention node of the 'ground relation process'. Table 4.4 lists all of these tests (G58,…,G65). It is similar to Table 4.3 and summarizes information about the visual scene that was presented to the model in each test. In the columns from left to right, it shows the identifier of the test, the total number of objects in the scene, the number of red

ID	# objects	# potential targets	# potential references	# targets moving	# references moving	# relation fits	+/−	description
G58	2	1	1	0	0	1	+	no motion; fit
G59	2	1	1	0	0	0	−	no motion; no fit
G60	2	1	1	0	1	1	+	ref. moving; fit
G61	2	1	1	0	1	0	−	ref. moving; no fit
G62	2	1	1	1	0	1	+	tar. moving; fit
G63	2	1	1	1	0	0	−	tar. moving; no fit
G64	2	1	1	1	1	1	+	both moving; fit
G65	2	1	1	1	1	0	−	both moving, no fit

Table 4.4: Tests of grounding tasks in which the phrase is "the red object to the left of the green object". In all tests, the target object and reference object are uniquely identifiable in the scene. See text for detail.

objects in the scene (potential targets), the number of green objects (potential references), the number of moving red objects, the number of moving green objects, the number of object pairs for which the description fits, whether or not the relation between the target object and reference object matches the specified relation (marked with '+' or '−', respectively), as well as a language description of the visual scene.

The results of all tests show that the model only brings pairs of objects into the attentional foreground that are specified by the phrase (in this case, red objects to the left of green objects). In all visual scenes that contain a red object to the left of a green object (those marked with '+'), the model is able to sequentially bring both objects into the attentional foreground and form a representation in the target field and reference field, respectively. In all tests in which their relative spatial position matches the specified spatial relation, the model grounds the phrase. This works irrespective of whether or not the target object or reference object are moving (e.g., G60,G62,G64).[15]

[15] Even though the objects move in the scene, their relative position never changes qualitatively. For instance, if the red object is to the left of the green object at the beginning of the video, it will never cross over to its right.

The tests also systematically check whether pairs of objects that do not match the specified relation are brought into the attentional foreground. The result of the tests show that for all those cases (marked with '−'), the model does not ground the phrase. While the objects are brought into the attentional foreground, the model detects that their spatial relation does not match. It rejects a mismatching target object, but holds it in memory and searches for other red object in the scene. Since, in these tests, there is only ever a single red object in the scene, the model keeps searching indefi-

ID	# objects	# potential targets	# potential references	# tar. relation fits	# ref. relation fits	+/−	description
G66	3	1	2	0	0	−	multiple references, no fit
G67	3	1	2	1	1	+	multiple references, one fit
G68	3	1	2	1	2	+	multiple references, two fits
G69	3	2	1	0	0	−	multiple targets, no fit
G70	3	2	1	1	1	+	multiple targets, one fit
G71	3	2	1	2	1	+	multiple targets, two fits
G72	4	2	2	0	0	−	multiple tar/ref, no fit
G73	4	2	2	1	1	+	multiple tar/ref, one fit
G74	4	2	2	2	2	+	multiple tar/ref, two fits

Table 4.5: Tests of grounding tasks in which the phrase is "the red object to the left of the green object". In these tests, the visual scenes feature multiple objects that fit the color-description of the target and reference object. The tests differ in how many combinations of objects fit the specified spatial relation. See text for detail.

nitely. As explained for previous tests, the model does not detect that there is no other red object to be found. This is the expected behavior since the model does not have a mechanism to detect this case.

In a second set of tests, the task input remains the same[16] but the target object and reference object are no longer uniquely identifiable by their color. The scenes feature multiple objects that fit the color-description of the target and reference object, and vary in the number of combinations of these objects that fit the specified relation. Table 4.5 lists all of these tests (G66,…,G74). It is similar to Table 4.4 and summarizes information about the visual scene that was presented to the model in each test. In the columns from left to right, it shows the identifier of the test, the total number of objects in the scene, the number of red objects in the scene, the number of green objects, the number of red objects that are to the left of a green object, the number of green objects that are to the right of a red object, whether or not a target object can be found in the scene (marked with '+' or '−', respectively), as well as a language description of the visual scene.

[16]The task input corresponds to the phrase "the red object to the left of the green object".

As before, the result of all tests shows that the model only brings pairs of objects into the attentional foreground that are specified by the phrase (see, e.g., G67, G70, G73). If there are multiple objects that match the color-description of either the target object, the reference object, or both, the model makes selection decisions to ground exactly one pair of objects that fits the specified relation (e.g., G68, G71, G74). The target object is grounded first; its selection

ID	# objects	# target matches	# reference matches	# target moving	# reference moving	relation fits	+/−	description
G75	2	1	1	0	0	0	−	no motion
G76	2	1	1	0	1	0	−	reference moving
G77	2	1	1	1	0	1	+	tar. moving; fit
G78	2	1	1	1	0	0	−	tar. moving; no fit
G79	2	1	1	1	1	1	+	both moving; fit
G80	2	1	1	1	1	0	−	both moving; no fit

Table 4.6: Tests of grounding tasks in which the phrase is "the red object moving toward the green object". In these tests, the visual scenes feature objects that can be uniquely identified by their colors. See text for detail.

is based only on the saliency of the object. The reference object is grounded second; its selection is based both on its saliency and on its match with the specified relation.

The tests also systematically check whether pairs of objects that do not match the specified relation are brought into the attentional foreground. The result of the tests show that for all those cases (marked with '−'), the model does not ground the phrase (G66, G69, G72). Objects are brought into the attentional foreground, but the model detects that their spatial relation does not match. It rejects mismatching target object, but holds them in memory and searches for other red objects in the scene until one of them matches the description. If none matches, the model keeps searching for other red objects indefinitely. As explained for previous tests, this is the expected behavior.

Two additional sets of tests replicate the results shown above for task input that corresponds to the phrase "the red object moving toward the green object". As before, the first set of tests features visual scenes in which the target object and reference object can be uniquely identified by their color. Table 4.6 lists all of these tests (G75,…,G80). In the columns from left to right, it shows the identifier of the test, the total number of objects in the scene, the number of red objects in the scene (potential targets), the number of green objects (potential references), the number of moving red objects, the number of moving green objects, the number of pairs of objects for which the description fits, whether or not a target object can be found in the scene (marked with '+' or '−', respectively), as well as a language description of the visual scene.

A second set of tests features multiple objects that fit the color-description of the target and reference object, and varies the number

ID	# objects	# target matches	# reference matches	# tar. relation fits	# ref. relation fits	+/−	description
G81	3	1	2	0	0	−	multiple references, no fit
G82	3	1	2	1	1	+	multiple references, one fit
G83	3	1	2	1	2	+	multiple references, two fits
G84	3	2	1	0	0	−	multiple targets, no fit
G85	3	2	1	1	1	+	multiple targets, one fit
G86	3	2	1	2	1	+	multiple targets, two fits
G87	4	2	2	0	0	−	multiple tar/ref, no fit
G88	4	2	2	1	1	+	multiple tar/ref, one fit
G89	4	2	2	2	2	+	multiple tar/ref, two fits

Table 4.7: Tests of grounding tasks in which the phrase is "the red object moving toward the green object". In these tests, the visual scenes feature multiple objects that fit the description of the target and reference object. The tests differ in how many combinations of objects fit the specified spatial relation. See text for detail.

of combinations of objects that fit the specified relation. Table 4.7 lists all of these tests (G81,…,G89). In the columns from left to right, it shows the identifier of the test, the total number of objects, the number of red objects (potential targets), the number of green objects (potential references), the number of red objects that are to the left of green objects, the number of green objects to the right of red objects, whether or not a target object can be found in the scene (marked with '+' or '−', respectively), as well as a language description of the visual scene.

The results of both sets of tests show that the model only brings pairs of objects into the attentional foreground that are specified by the phrase (in this case, red objects moving toward green objects). In all visual scenes that show a red object moving toward a green object (those marked with '+'), the model sequentially brings both objects into the attentional foreground, forms representations of the objects in the target field and reference field, identifies that their relation matches the specified one, and thereby grounds the phrase. This works irrespective of whether or not the reference object is moving (G79). If there are multiple objects that match the color-description of either the target object, the reference object, or both, the model makes selection decisions to ground exactly one pair of objects that fits the specified relation (e.g., G83, G86, G89). The target object is grounded first; its selection is based only on its saliency. The reference object is grounded second; its selection is based both on its saliency and on its match with the specified relation.

Both sets of tests also systematically check whether pairs of objects that do not match the specified relation are brought into the attentional foreground. The tests show that for all those cases (marked with '–'), the model does not ground the phrase. If there are multiple red objects and the model initially selects one that is not moving toward a green object, it detects this, rejects the red object, but holds it in memory and searches for other red objects in the scene until one of them matches the description. If none matches, the model keeps searching for other red objects indefinitely. As explained for previous tests, this is the expected behavior.

Example with a unique target

The following describes the processes that unfold in the model when it grounds an object that is uniquely identifiable by the phrase "the red object to the left of the green object" (test G70 in Table 4.5). In this example, the model grounds the phrase in a visual scene that contains two red objects, one of them to the left and the other to the right of a green object. All objects in the scene are stationary.[17] Figure 4.8 shows how the activation in the model evolves over the course of the test. It is structured analogously to previous figures but shows more plots since in this example both the target object and the reference object as well as their spatial relation are grounded. The top panel shows the activation of the intention node and CoS node of the target process, the reference process, and the spatial relation process over the continuous time course of the test. The third row shows the activation level of the target color memory nodes (transparent colors; left side of the panel for each point in time), the target color production nodes (solid colors; left side of each panel), the reference color memory nodes (transparent colors; right side of each panel), and the reference color production nodes (solid colors; right side of each panel). As before, the colors of the bars match the colors that the nodes represent. The fourth row shows the activation level of the spatial relation memory nodes (transparent colors) and the spatial relation production nodes (solid colors) for the relations (from left to right) TO THE LEFT OF, TO THE RIGHT OF, ABOVE, BELOW, TOWARD, and AWAY FROM. For each node, the synaptic connection pattern to the relational candidates field is shown below in a color-code.[18] The last three rows show the activation of the target field, the reference field, and the spatial relation CoS field (from top to bottom), color-coded using the color map in the bottom right of Figure 4.8.

At the beginning of the test, the model is already being simulated and is receiving visual input from the video. However, the activation in all fields and nodes of the model is below the threshold.

[17] The video used in this test is video 3.05b. Please refer to Appendix C, in particular Figure C.28 for more information.

[18] This is meant as a symbol to denote which relational concept each bar represents.

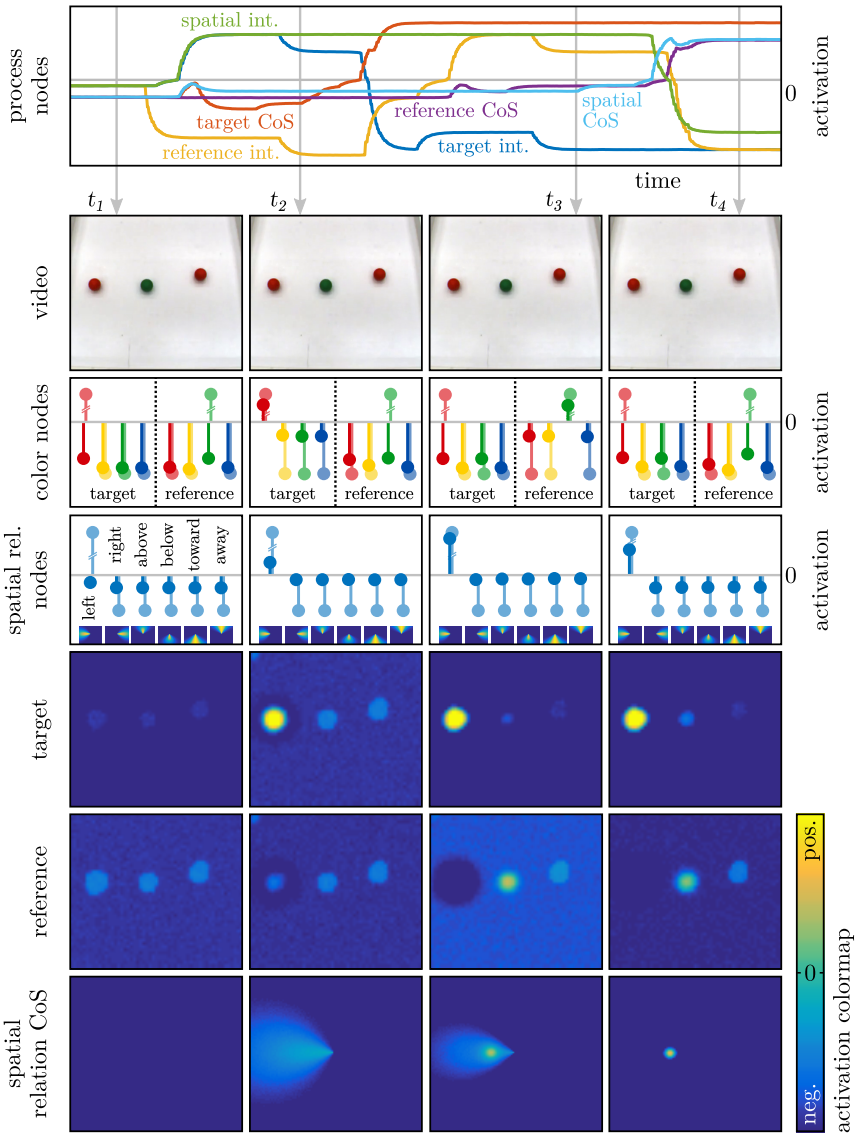


FIGURE 4.8: Grounding the phrase "the red object to the left of the green object" in a scene with a unique target. See text for more detail.

The only exception is the color/space perception field (not shown), which has three stable peaks of activation, each representing the spatial position and color of one of the objects in the scene. As task input, the user activates the target color memory node for the color RED, the reference color memory node for the color GREEN, and the spatial relation memory node for the spatial relation TO THE LEFT OF. This corresponds to the phrase "the red object to the left of the green object". Figure 4.8 shows that at time t_1 (first column) all fields and nodes are below threshold but the three memory nodes are active. Some localized subthreshold bumps are (barely) visible in the plots of the target field (fifth row) and reference field (sixth row). These are due to input from the color/space perception field. Shortly after t_1, the user initiates the grounding process by activating the prior intention node of the 'ground relation process'. Once

the intention node of this process is active, it gives input to the following processes on the next lower hierarchical level: the target process, the reference process, the spatial relation process, the clean process, and the reset process. In addition, it activates multiple precondition nodes that enforce these processes to become active in a sequential order: the target process and the spatial relation process become active first (see top panel of Figure 4.8), the clean process becomes active once the target process is finished (not shown in Figure 4.8), and the reference process becomes active once the clean process is finished (yellow line in the top panel, between t_2 and t_3). The reset process (not shown) is not activated in this example.

At time t_2 (second column), the target process and spatial relation process are active. These give homogeneous input to the target color production nodes and spatial relation production nodes, respectively, and activate the nodes that also receive input from the previously activated memory nodes (third and fourth row in Figure 4.8). The active target color production node for the color RED brings the two red objects in the scene into the attentional foreground. The model makes a selection decision based on the saliency of the objects and forms a representation of the spatial position of the left red object in the target field (fifth row). The active spatial relation production node for the spatial relation TO THE LEFT OF projects into the spatial relation CoS field via its patterned synaptic connections, leading to a localized, subthreshold pattern of activation on the left side of the field (last row). Once the target object has been grounded and the target process is finished, the clean process is activated (not shown), which ensures that the color attention field no longer has a peak. Once this process is finished, the reference process becomes active.

At time t_3 (third column), the reference process is active and gives homogeneous input to all reference color production nodes, activating the one representing the color GREEN (third row). This brings the green object in the scene into the attentional foreground and forms a representation of its spatial location in the reference field (sixth row). With peaks both in the target field and the reference field, the spatial relation CoS field receives input that reflects the relative position of the target object with respect to the reference object. In Figure 4.8 (last row), this is visible as a small yellow circle to the left of the center of the field. This input overlaps with the one representing the spatial relation TO THE LEFT OF, and thus forms a peak in the spatial relation CoS field. This peak shows that the selected red object is in fact to the left of the selected green object. At this moment, the phrase "the red object to the left of the green object" has been successfully grounded.

At time t_4 (fourth column), the CoS of all active processes have

been met and the activation in large parts of the model has returned to below threshold. The peaks in the target field, target IOR field (not shown), reference field, and spatial relation CoS field, as well as the activation in the memory nodes remain because of their high self-excitation. They represent the end-result of the grounding process.

Example with hypothesis testing

The following example describes the processes that unfold in the model when it grounds the phrase "the red object to the left of the green object" in a scene that contains multiple red and green objects, all of which are stationary. However, only one of the red objects is to the left of a green object (test G73 in Table 4.5).[19] Since there are multiple possible combinations of red and green objects in the scene, the model is required to select objects and check whether they match the specified relation. Please note that this form of *hypothesis testing* is not necessarily required in the visual scene of the previous example since one could begin by grounding the unique (green) reference object and infer the correct target object with the help of the spatial relation. However, in the visual scene of the current example, there are multiple candidates for both the target object and the reference object and testing hypotheses cannot be avoided. Figure 4.9 shows how the activation in the model evolves over the course of the test. It is structured analogously to Figure 4.8, with three additions. First, the top panel includes the activation of the intention node of the reset process. Second, there is an additional panel in the sixth row that shows activation snapshots of the target IOR field at four points $t_1, \ldots, t_4$ in time. Third, an analogous additional panel in the last row shows activation snapshots of the spatial relation CoD field. In both additional panels, the activation is shown using the same color-code as for the other fields (color map in the bottom right of Figure 4.9).

[19] The video used in this test is video 4.18b. Please refer to Appendix C, in particular Figure C.76 for more information.

At the beginning of the test, the model is already being simulated and is receiving visual input from the video. As task input, the user activates the target color memory node for the color RED, the reference color memory node for the color GREEN, and the spatial relation memory node for the spatial relation TO THE LEFT OF. This corresponds to the phrase "the red object to the left of the green object". The user then initiates the grounding process by activating the prior intention node of the 'ground relation process'. As in the previous example, the target process (blue line in the top panel of Figure 4.9) and spatial relation process (green line) activate first. At time t_1 (first column), the target process has brought one of the red objects in the scene into the attentional foreground and forms a rep-

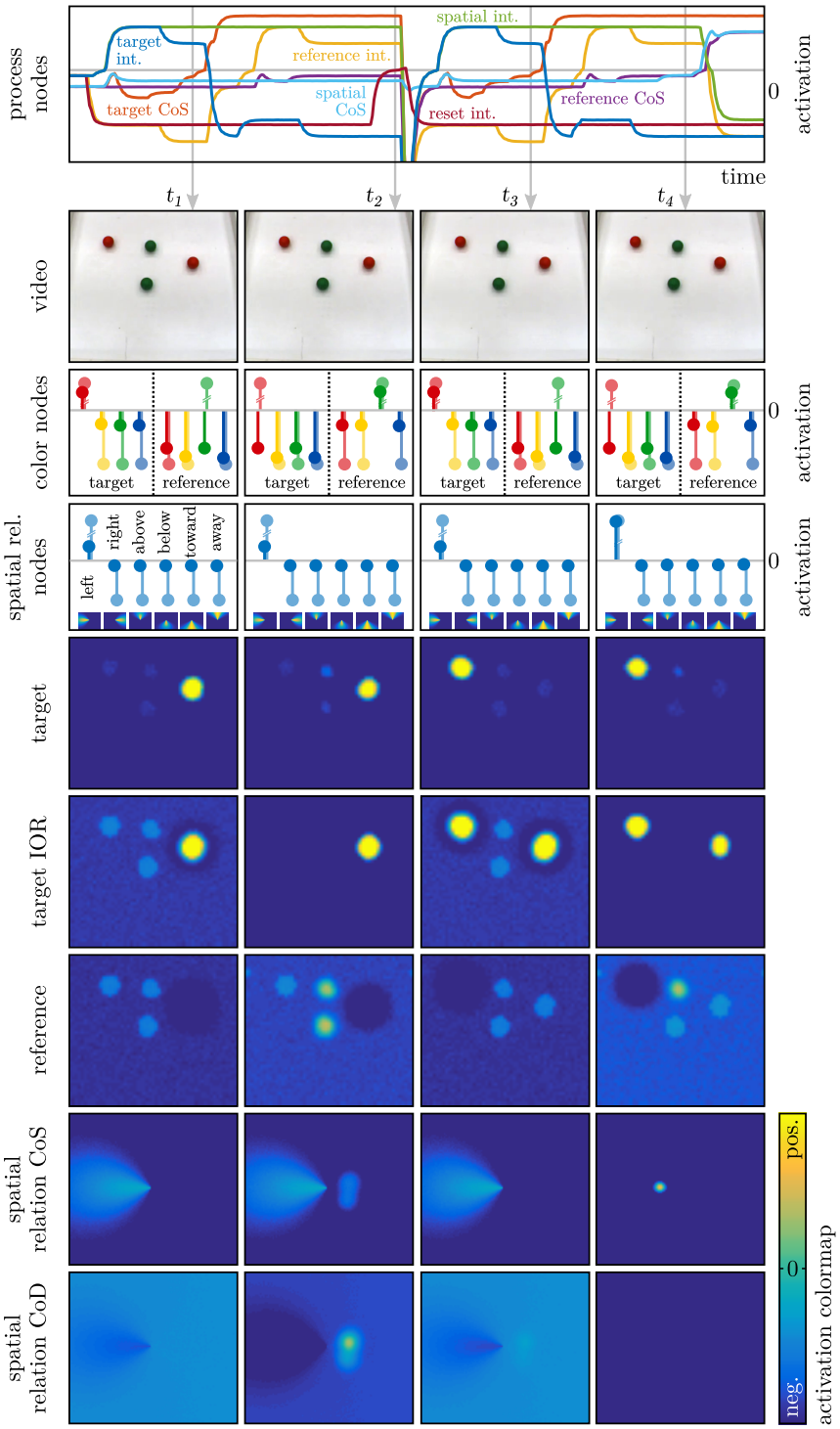


FIGURE 4.9: Grounding the phrase "the red object to the left of the green object" in a scene that requires hypothesis testing. The target object is uniquely identifiable by the description, but there are multiple objects that match the color description of the target object as well as the reference object. See text for more detail.

resentation of its spatial location in the target field (fifth row). Since the selection between the two red objects is based on their saliency only, the model selects the object on the right. The activation in the target field is projected onto the target IOR field (sixth row), which has strong self-excitation and forms a self-sustained peak. The spa-

tial relation process activates the spatial relation production node for the relation TO THE LEFT OF, which projects into both the spatial relation CoS field (eight row) and spatial relation CoD field (last row) via patterned synaptic connections. Please note that the input to the spatial relation CoS field is excitatory while the input to the spatial relation CoD field is inhibitory. The spatial relation CoS field thus matches all inputs that the spatial relation CoD field does not and vice versa.

At time t_2 (second column), the reference process is active (top panel, yellow line) and has brought the two green objects into the attentional foreground. They form two peaks in the reference field (seventh row). The spatial relation CoS field (eighth row) and spatial relation CoD field (last row) receive input from the spatial transformations that reflects the spatial position of the selected red object (represented in the target field) with respect to all green objects (represented in the reference field). Since the selected red object is to the right of both green objects, the subthreshold bumps that are input to the spatial relation CoS field (eight row) do not overlap with the input from the spatial relation production node and cannot form a peak. However, the input does form a peak in the spatial relation CoD field (last row). As soon as there is a peak in the spatial relation CoD field (shortly after t_2), the reset process is activated. This activates a suppression node that inhibits all other processes and large parts of the model.[20] In Figure 4.9 this is visible in the top panel, where the activation of all intention nodes and CoS nodes drops when the intention node of the reset process (dark red line) becomes active. Afterward, the grounding process begins anew, starting with the target process and the spatial relation process becoming active.

[20] See Section 3.5.

At time t_3 (third column), the target process has brought red objects in the scene in the foreground once more. However, since there is still a peak in the target IOR field at the position of the previously selected red object (the one on the right) and the target IOR field inhibits this spatial location in the spatial attention fields, the model selects a different red object. The target field (fifth row) thus forms a peak at the spatial location of the red object on the left. As before, the position of this object is represented by a self-sustained peak in the target IOR field (sixth row). At time t_3, the activation of the other three fields shown in Figure 4.9 (rows 7–9) have returned to a state below threshold, similar to that at time t_1 (first column). Once the reference process is active, it brings both green objects into the attentional foreground again.

As before, the spatial transformations project two subthreshold bumps of activation into the spatial relation CoS field and spatial relation CoD field, but this time the input matches the spatial rela-

tion TO THE LEFT OF in the spatial relation CoS field. At time t_4 (fourth column), the spatial relation CoS field has made a selection decision between the two potential reference objects: one of the two subthreshold bumps in the spatial relation CoS field fits the spatial relation better and forms a peak (eight row). It is the bump that reflects the relative position of the selected red object (represented in the target field) with respect to the upper green object in the scene. The other bump input is inhibited by strong global inhibitory interaction within the spatial relation CoS field. The peak that remains in the field is transformed back and projected into the spatial attention system, bringing the selected green object into the attentional foreground while suppressing the other green object. This is visible in the activation of the reference field (seventh row), which only has a single peak.

At this moment, the model has grounded the phrase "the red object to the left of the green object". The peaks in the target field, the reference field, and the spatial relation CoS field represent the red object, the green object, and their relation that the phrase refers to.

Example with multiple potential targets

The following example describes the processes that unfold in the model when it grounds the phrase "the red object moving toward the green object". The scene contains two red objects, each of which is moving toward a different, stationary green object.[21] There are thus multiple objects in the scene that fit the description of the phrase (test G89 in Table 4.7). Figure 4.10 shows how the activation in the model evolves over the course of the test. It is structured analogously to Figure 4.8 with the addition of a panel (seventh row) that shows the activation of the relational candidates field at four points $t_1, \ldots, t_4$ in time. The activation of the relational candidates field is visualized using the same color map (bottom right of Figure 4.10) as all other color-coded plots.

[21]The video used in this test is video 4.19c. Please refer to Appendix C, in particular Figure C.80 for more information.

At the beginning of the test, the model is already being simulated and is receiving visual input from the video. As task input, the user activates the target color memory node for the color RED, the reference color memory node for the color GREEN, and the spatial relation memory node for the spatial relation TOWARD. This corresponds to the phrase "the red object moving toward the green object". The user then initiates the grounding process by activating the prior intention node of the 'ground relation process'. As in the previous example, the target process (blue line in top panel of Figure 4.10) and spatial relation process (green line) activate first. At time t_2 (second column), the target process has brought one of the

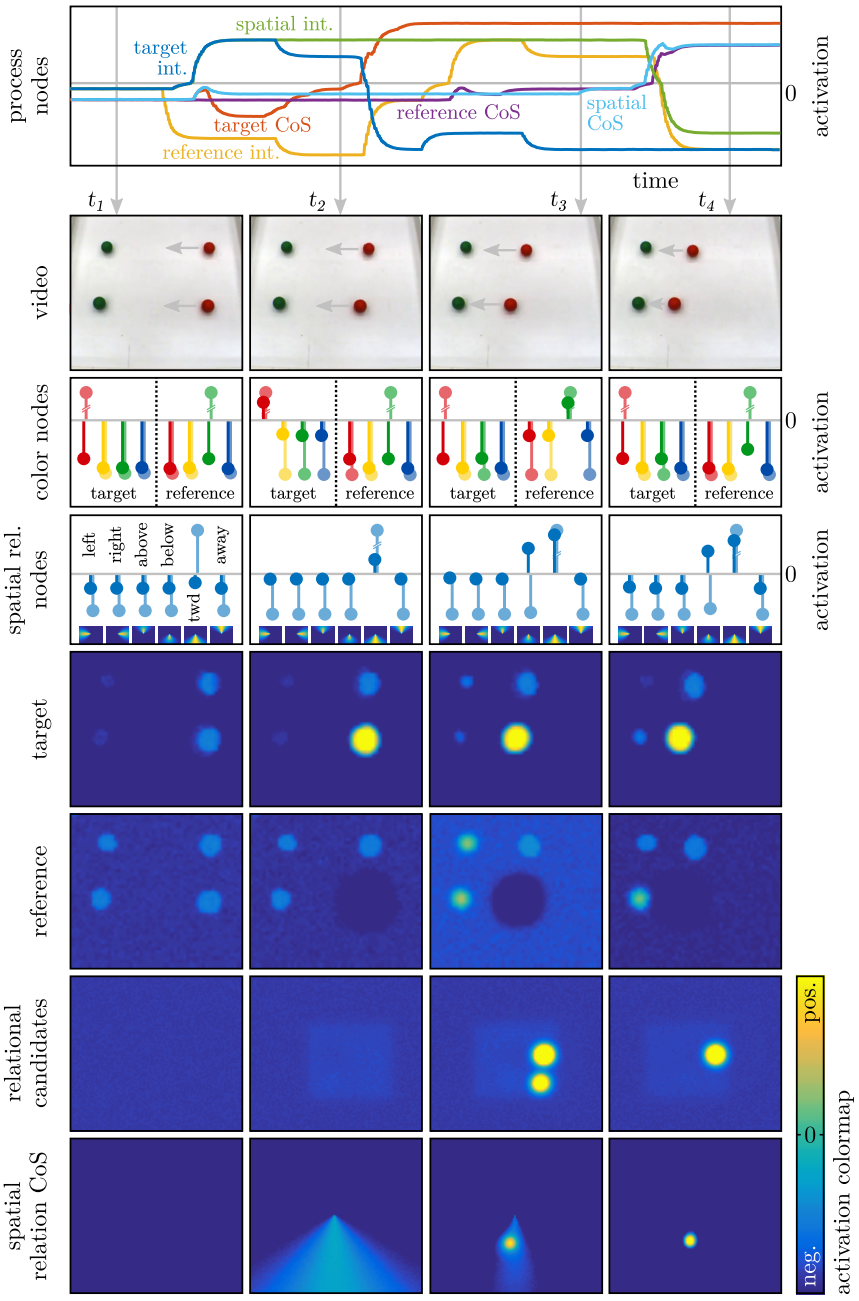


FIGURE 4.10: Grounding the phrase "the red object moving toward the green object" in a scene that features multiple objects fitting that description. See text for more detail.

red objects in the scene into the attentional foreground and forms a representation of its spatial location in the target field (fifth row). Since the selection between the two red objects is based on their saliency only, the model selects the lower object. While forming a representation of its spatial position, the model also extracts the motion direction of the selected red object (not shown in Figure 4.10). The spatial relation process activates the spatial relation production node for the spatial relation TOWARD (fourth row; abbreviated "twd" in the first column), which projects into the spatial relation CoS field (last row) via patterned synaptic connections.

At time t_3 (third column), the reference process is active (yellow line in top panel) and has brought all green objects in the scene into the attentional foreground. Their positions are represented in the reference field (sixth row). As explained in previous examples, the relative position of the target object (represented in the target field) with respect to all possible reference objects (represented in the reference field) is established by the spatial transformation, which projects into the relational candidates field (seventh row).[22] Projecting into the spatial relation CoS field (last row), the activation in the relational candidates field is transformed once more, essentially rotating the representation around the center of the field. The angle by which it is rotated is determined by the motion direction previously extracted from the object that is represented in the target field. Thus, the position of the bump input in the spatial relation CoS field always moves in a fixed direction. The activation in the spatial relation CoS field is plotted in Figure 4.10 such that objects moving in the scene move upward in the plot. Since the spatial relation CoS field is defined such that the position of the reference object is at the center, the positions of objects that move toward the reference object lie below it. Thus, the synaptic connection pattern that encodes the perceptual meaning of the spatial relation TOWARD is shaped to match a region below the center. Both bump inputs from the relational candidates field fall into that region in the spatial relation CoS field, but one of the bumps fits the relation better and forms a peak, suppressing the other. The peak that remains in the field is transformed back and projected into the spatial attention system, bringing the selected green object into the attentional foreground while suppressing the other green object.

Note that at time t_3 and t_4, the spatial relation production nodes for the spatial relation BELOW is active alongside the one for the spatial relation TOWARD. This is because the patterned synaptic connections between the spatial relation CoS field and the spatial relation production nodes are very similar for these two spatial relations and the peak in the field activates both nodes. However, only the spatial relation memory node for the spatial relation TOWARD is active. While multiple production nodes may be active at the same time, the selective memory nodes represent the grounding of the model.

At time t_4 (fourth column), the grounding of the phrase "the red object moving toward the green object" has been established. There is a peak in the target field, representing the spatial position of the selected target object, the lower of the two red objects in the scene. Analogously, the peak in the reference field represents the spatial position of the selected reference object, also the lower of the two green objects in the scene. The model has selected one of the two pairs of objects fitting the description. The fact that it selected the

[22] The relational candidates field was left out in the explanation of previous examples because for stationary scenes the activation there is largely similar to that of the spatial relation CoS field.

lower pair is likely due to the perspective distortion of the camera input, which makes objects in the lower part of the scene appear larger and thus more salient. However, in selecting the reference object, the influence of the specified spatial relation is a larger bias. Thus, the reference object is selected not only because it is more salient but rather because it fit the specified relation better than the other green object in the scene (which the red object is also moving toward, only less directly).

Example with no target

The following example describes the processes that unfold in the model when it grounds the phrase "the red object moving toward the green object" in a scene where no such object exists (test G78 in Table 4.6). The scene contains a red and a green object, but the red object is moving away from the (stationary) green object, instead of toward it.[23] Figure 4.11 shows how the activation in the model evolves over the course of the test. It is structured analogously to Figure 4.9.

[23]The video used in this test is video 2.02d. Please refer to Appendix C, in particular Figure C.15 for more information.

At the beginning of the test, the model is already being simulated and is receiving visual input from the video. As task input, the user activates the target color memory node for the color RED, the reference color memory node for the color GREEN, and the spatial relation memory node for the spatial relation TOWARD. This corresponds to the phrase "the red object moving toward the green object". The user then initiates the grounding process by activating the prior intention node of the 'ground relation process'. As in the previous example, the target process (blue line in the top panel of Figure 4.11) and spatial relation process (green line) activate first. At time t_2 (second column), the target process has brought the red object in the scene into the attentional foreground and has formed a representation of its spatial location in the target field (fifth row). The activation in the target field is projected onto the target IOR field (sixth row), which has strong self-excitation and forms a self-sustained peak. The spatial relation process activates the spatial relation production node for the relation TOWARD (fourth row), which projects into the spatial relation CoS field (eighth row) and spatial relation CoD field (last row) via patterned synaptic connections.

At time t_3 (third column), the reference process is active (top panel, yellow line) and has brought the green object into the attentional foreground. It forms a peak in the reference field (seventh row). The spatial relation CoS field (eighth row) and spatial relation CoD field (last row) receive input from the spatial transformations (both shift and rotation). As explained in the previous example, the input reflects the spatial position of the red object with respect to

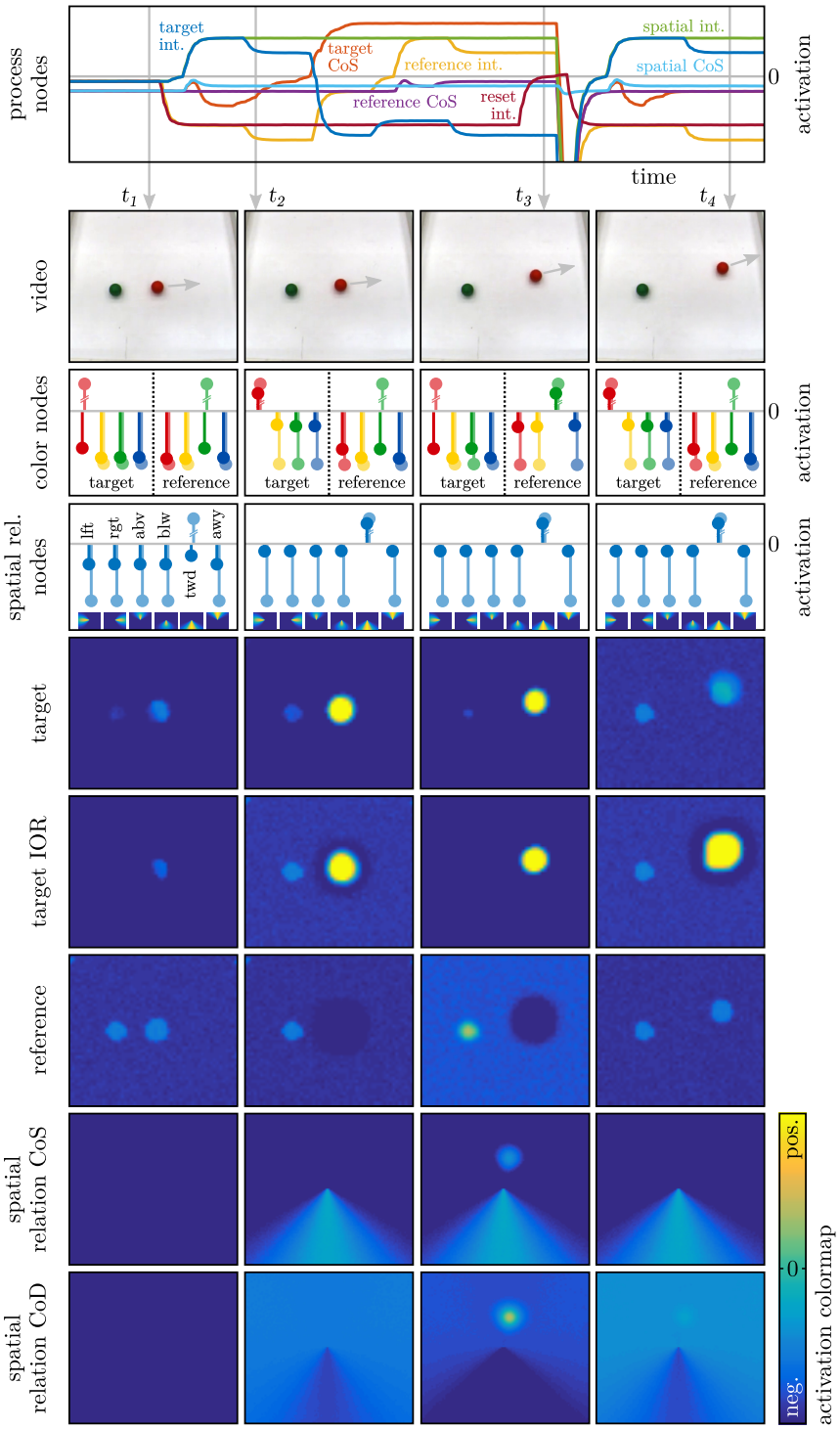


FIGURE 4.11: Grounding the phrase "the red object moving toward the green object" in a scene where no such object exists. See text for more detail.

the green object, rotated such that its motion direction points upwards in the plot. The red object is moving away from the green object. In the figure, this is visible by the bump input in the spatial relation CoS field and spatial relation CoD field (last two rows) being above the center. Thus, the subthreshold bump in the spatial

relation CoS field does not overlap with the input from the spatial relation production node and cannot form a peak. However, the input does form a peak in the spatial relation CoD field. As soon as there is a peak (shortly after t_3), the reset process is activated. As explained before, this activates a suppression node that inhibits all other processes (see top panel) and large parts of the model. Afterward, the grounding process begins anew, starting with the target process and the spatial relation process becoming active.

At time t_4 (fourth column), both processes are active and the target process is trying to bring red objects into the attentional foreground. However, the position of the only red object in the scene is inhibited by the peak in the target IOR field (sixth row). The field tracks the position of the moving red object and the model does not ground the phrase until the end of the test. Nevertheless, the model behaves as expected since it does not have a mechanism to detect that it has checked all potential objects in the scene.

4.2 Description tasks

In *description tasks* a symbolic description of a scene is generated. Depending on the scene and what is being attended to, this description may take on different forms and even focus on different aspects of the scene. The model is constrained in how it generates descriptions because in one description, it can only describe a single object. For scenes that only contain one object, the model extracts all available features (color and motion direction, if the object is moving) and generates task output that corresponds to a phrase such as "a red object moving upward". The model produces the task output by activating memory nodes that correspond to concepts in the phrase. In the example above, the model would activate the target color memory node for the color RED and the target motion memory node for the motion direction UPWARD. For scenes that contain multiple objects, the model describes the object that is most salient and uses a spatial relation (or movement relation) that the object has to one other object in the scene. The model produces relational phrases whenever there are multiple objects in the scene, even if the object that it is describing can unambiguously be referred to by its color or motion direction alone. An exemplary task output would correspond to a phrase such as "a red object that is to the left of a green object". The task output for this example would be given by activating the target color memory node for the color RED, the reference color memory node for the color GREEN, and the spatial relation memory node for the spatial relation TO THE LEFT OF.

The following section shows the results of systematically testing

ID	# objects	# moving objects	+/−	description
D1	0	0	−	no object
D2	1	0	+	red object, not moving
D3	1	0	+	green object, not moving
D4	1	1	+	red object, moving upward
D5	1	1	+	green object, moving leftward

Table 4.8: Tests of description tasks with scenes that contain at most a single object. See text for detail.

the model on description tasks with 15 qualitatively different scenes. Each of the 15 tests is referred to with an identifier (D1,…,D15). In all cases, the user only activates the describe process. Once activated, the model autonomously generates a symbolic description of the scene without further intervention by the user.

4.2.1 Single objects

The first set of tests establishes that the model is able to attend to single objects and describe their features. In all tests, there is at most a single object in the scene. The tests vary in the color of the object, its spatial position, its motion direction, and whether it moves at all. Table 4.8 lists all of these tests (D1,…,D5). In the columns from left to right, it shows the identifier of the test, the total number of objects, the number of moving objects, whether or not a target object can be described in the scene (marked with '+' or '−', respectively), as well as a language description of the visual scene.

The results of the tests show that if there is an object in the scene, the model correctly extracts all features of that object that are available in the scene (D2,…,D5). That is, if the object is stationary, the model extracts its color; if the object is moving, the model additionally extracts its motion direction. The model also forms representations of the object's spatial position, but does not extract it as part of the symbolic description. However, this is expected since the model does not have a way to express discrete concepts of absolute spatial positions.

If there is no object in the scene (D1), the model does not generate a symbolic description, nor does it do anything else. As previously mentioned, there is no mechanism to detect that the scene does not contain any objects. It is thus the expected behavior of

the model to remain in its state indefinitely until an object with a saturated color appears.

Example of describing a scene with a single object

The following example describes the processes that unfold in the model when it describes a scene that contains a single red object that is moving upward (test D4 in Table 4.8).[24] The task of the model is thus to bring the object into the attentional foreground and extract the available features (color and motion direction). Figure 4.12 shows how the activation in the model evolves over the course of the test. It is structured analogously to previous figures, in particular Figure 4.4. In addition, the top panel shows the activation of the intention node (yellow line) and CoS node (violet line) of the reference process. Moreover, an additional panel (seventh row) shows the activation of the selective spatial attention field at four points $t_1, \ldots, t_4$ in time. The activation of the field is visualized using the same color map (bottom right of Figure 4.12) as for all other plots.

[24] The video used in this test is video 1.01. Please refer to Appendix C, in particular Figure C.4 for more information.

At the beginning of the test, the model is already being simulated and is receiving visual input from the video. As task input, the user activates the prior intention node of the describe process. The intention node of the describe process, which activates shortly after, gives input to all processes on the next lower hierarchical level, as well as a few on even lower levels, each of which controls an aspect of describing the scene.[25] The intention node also gives excitatory input to the target color memory nodes and the target motion memory nodes (transparent bars in the third and fourth row, first column of Figure 4.12), bringing them into a dynamic regime where they can be activated by their corresponding production node.

[25] See Figure 3.14.

At time t_1 (first column), the target process is active (blue line in the top panel). It gives a homogeneous excitatory input to the selective spatial attention field[26] and the multi-peak spatial attention field, bringing them into a dynamic regime where they can create peaks from the localized input they receive from the color/space perception field and the motion/space perception field. The spatial position at which peaks form thus depends entirely on the strength of the input, which reflects the salience of the objects in the scene. The selective spatial attention field forms a peak at the spatial location of the red object (seventh row). This activation projects into the three-dimensional color/space attention field (fifth row) and motion/space attention field (sixth row). In these fields, the input is localized along the spatial dimensions x, y, centered on the spatial position of the red object, and it is homogeneous along the other feature dimensions of the fields (color c and motion direction ϕ, re-

[26] The selective spatial attention field also receives excitatory input from the spatial attention process, which the target process activates. Without that additional input, it would not be able to form peaks.

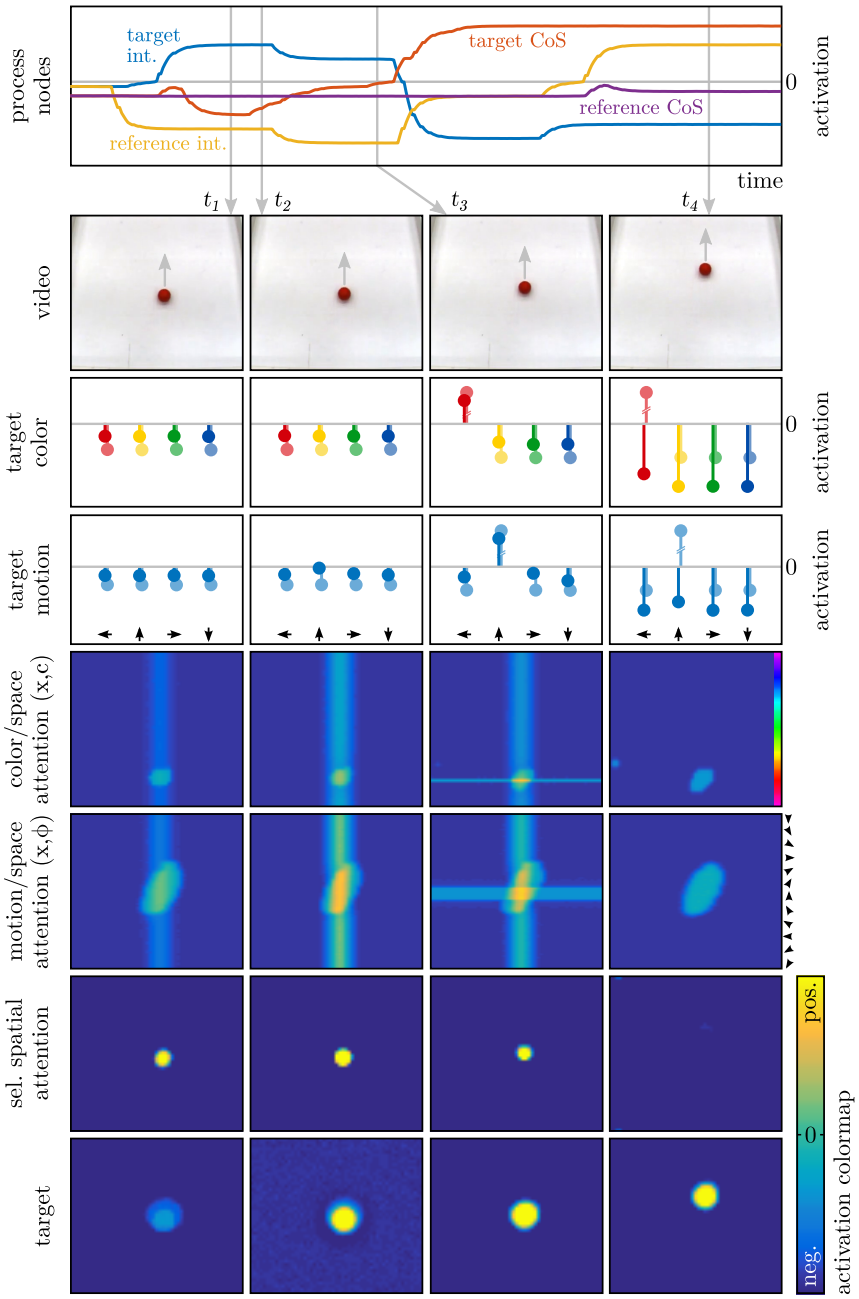


FIGURE 4.12: Generating a description of a single object in a scene; the object is red and moving upward. See text for more detail.

spectively). In the plots in Figure 4.12, the input shows up as thick, vertical lines. The input overlaps with the subthreshold bumps that stem from the red object in the scene.

At time t_2 (second column), both the color/space attention field and motion/space attention field have formed a peak at the spatial position and respective feature value of the red object. These peaks are projected onto the color CoS field (not shown), which is defined over color c and the motion CoS field (not shown), which is defined over motion direction ϕ and form peaks there. These fields project their activation onto the target color production nodes and

target motion production nodes, respectively, and activate them. At time t_2, the target motion production node for the motion direction UPWARD is just about to become active. The target color production nodes are still below threshold but the node for the color RED will be activated shortly after (see third row at time t_3). The spatial position of the red object is also represented by a peak in the target field (last row), which receives activation from the multi-peak spatial attention field (not shown).

At time t_3 (third column), both the target color production node for the color RED and the target motion production node for the motion direction UPWARD are active (third and fourth row). What happens next is very similar to the grounding tasks discussed earlier. The active production nodes brings up peaks in the color attention field and motion attention field (both not shown), which give input to the color/space attention field and motion/space attention field (visible as horizontal lines in the fifth and sixth row). Moreover, the active production nodes activate their respective memory nodes to form task output that corresponds to the phrase "a red object moving upward". At this moment, the model has generated a symbolic description of the scene. It has extracted the correct color and motion direction of the object and classified them by activating nodes that correspond to discrete concepts.

At time t_4 (fourth column), the target process has successfully finished and deactivated (top panel). As a result, all production nodes that are associated with the target have deactivated as well (third and fourth row). Most of the fields in the model have returned to below threshold. Only the target field (last row) still holds a stable peak that represents the spatial position of the red object. The same is true for the target IOR field, the color/space perception field, and the motion/space perception field (all not shown).

After the target process is deactivated (and the clean process is also successfully finished), the reference process is activated. Figure 4.12 shows that it remains active until the end of the test (yellow line in the top panel). This is the expected behavior of the model. There is no mechanism that detects that there is only a single object in the scene. Thus the model always activates the reference process in order to describe a relation. If the scene only contains a single object, like in the current example, the model activates the reference process and remains in this state indefinitely.

4.2.2 Relations between objects

The second set of tests establishes that the model is able to describe an object by a spatial relation or a movement relation it has with respect to another object that is present in the scene. In all tests,

ID	# objects	# moving objects	+/−	description
D6	2	0	+	relations between objects fit well into spatial terms
D7	2	0	+	relations between objects do not fit well into spatial terms
D8	3	1	−	moving toward a region without objects
D9	3	1	+	moving toward another object
D10	3	1	+	moving away from another object
D11	3	1	+	moving away from one and toward another object
D12	4	2	+	two objects moving toward two different objects
D13	4	2	+	two objects moving away from different objects
D14	4	2	+	two objects moving toward each other
D15	4	2	+	two objects moving away from each other

Table 4.9: Tests of description tasks with scenes that contain multiple objects. See text for detail.

there are multiple objects in the scene. The tests vary in the number of objects and their spatial configuration as well as their motion direction with respect to each other. Table 4.9 lists all of these tests (D6,…,D15). In the columns from left to right, it shows the identifier of the test, the total number of objects, the number of moving objects, whether or not a target object can be described by a relation in the scene (marked with '+' or '−', respectively), as well as a language description of the visual scene.

The results of the tests show that in all cases where the most salient object can be described by a spatial relation or movement relation, the model correctly selects both a target object and a reference object and extracts their available features as well as their relation. In case the target object is stationary, the model describes it in terms of a spatial relation (e.g., TO THE LEFT OF, TO THE RIGHT OF) to the reference object (D6,D7). In case the target object is moving, the model describes it using a movement relation (e.g., TOWARD, AWAY) with respect to the reference object (D9,…,D15). The model makes selection decisions as to what it describes in the scene: as target object, the model selects the most salient object in the scene; as reference object it selects the object that fits best to any spatial relation or movement relation with respect to the target object.

If there are multiple ways to describe the scene, the model selects one coherent form. For instance, in test D11, the scene could either be described as a red object moving toward a green object, or as a red object moving away from a blue object.

If the model has selected a target object and candidates for reference objects, but their relation does not match any of its relational concepts, the model does not generate a description and tries another object as the target object (D8).

Example of describing a static scene with multiple objects

The following example describes the processes that unfold in the model when it describes a scene that contains two stationary objects, a red object that is to the left of a green object (test D6 in Table 4.9).[27] The task of the model is to describe one of the objects in terms of its spatial relation to the other object. Figure 4.13 shows how the activation in the model evolves over the course of the test. It is structured analogously to Figure 4.8.

[27] The video used in this test is video 2.00a. Please refer to Appendix C, in particular Figure C.7 for more information.

At the beginning of the test, the model is already being simulated and is receiving visual input from the video. As task input, the user activates the prior intention node of the describe process. The intention node of the describe process, which activates shortly after, gives input to all processes on the next lower hierarchical level, as well as to a few on even lower levels, each of which controls an aspect of describing the scene.[28] The intention node also gives excitatory input to the target color memory nodes and the reference color memory nodes (transparent bars in the third row, first column), bringing them into a dynamic regime where they can be activated by their corresponding production node. Moreover, it gives input to all spatial relation production nodes (solid bars in the fourth row, first column). The spatial relation production nodes that represent spatial relations (i.e., TO THE LEFT OF, TO THE RIGHT OF, ABOVE, BELOW) receive excitatory input and are activated while the spatial relation production nodes that represent movement relations (i.e., TOWARD, AWAY) receive inhibitory input and remain below threshold. Please note that the spatial relation memory nodes (transparent bars, fourth row) are not yet in a dynamic regime where they can be activated by their respective production nodes.

[28] See Figure 3.14.

At time t_1 (first column), the target process and the spatial relation process are active (top panel in Figure 4.13). As in the previous example, the intention node of the target process gives excitatory input to all target color production nodes (third row, left side). Moreover, the intention node activates the perceptual boost process, which gives homogeneous excitatory input to both the selective spatial attention field and the multi-peak spatial attention field (neither

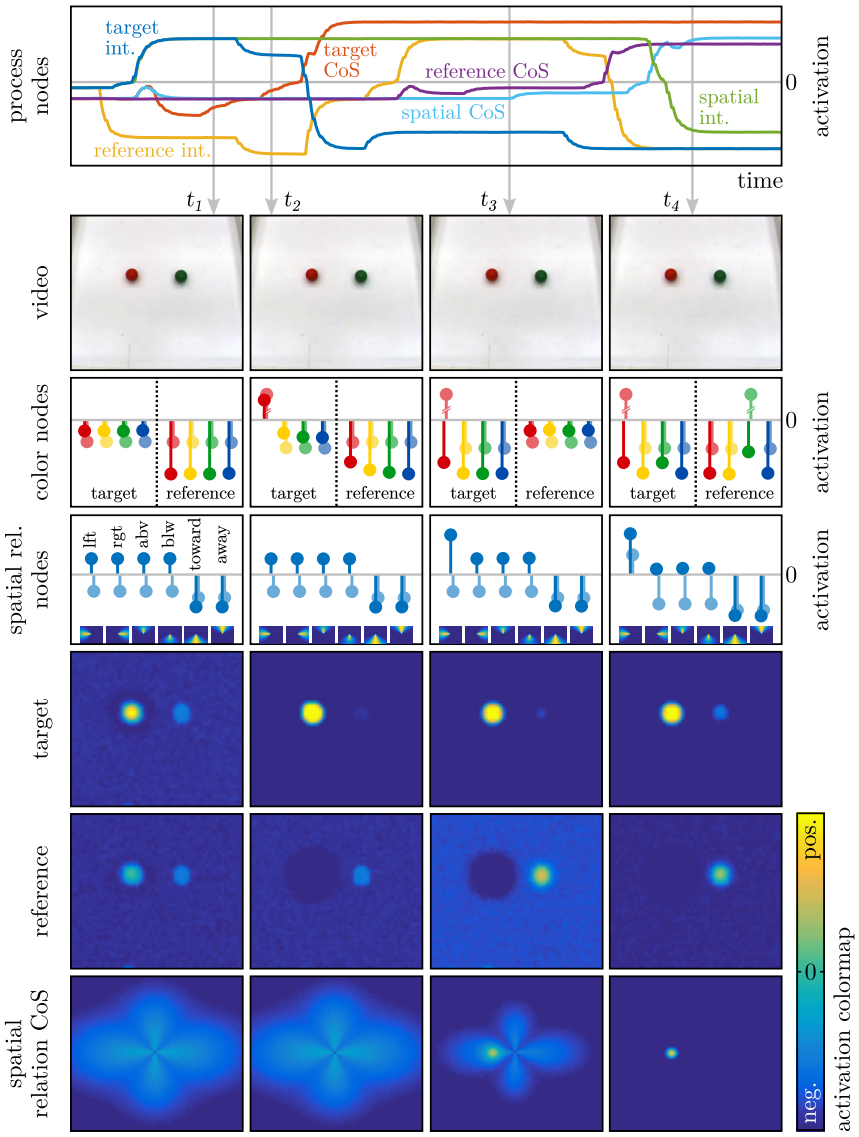


FIGURE 4.13: Generating a description of an object using a spatial relation with respect to to a second object. See text for more detail.

is shown). Because the spatial attention process is active, the selective spatial attention field is in a dynamic regime where it can form a peak. It makes a selection decision for the most salient object and forms a peak that is centered on the red object. The multi-peak spatial attention field forms a peak at the same spatial position and projects its activation into the target field (fifth row). At the same time, all active spatial relation production nodes project into the spatial relation CoS field via patterned synaptic weights, forming a subtreshold pattern of activation that has larger values in areas that match the known spatial relations (last row).

At time t_2 (second column), the target color production node for the color RED has been activated (third row, left side). This happened, as in the previous example, because the peak in the selective spatial attention field projects into the color/space attention field,

forming a peak there at the position and feature value of the red object. In turn, this peak projects into the color CoS field, forming a peak there at the position coding for red colors. The peak in this field activates the target color production node for the color RED. This in turn activates the target color memory node for the same color. Since the red object is not moving, its motion direction is not extracted and none of the target motion production nodes is activated (not shown).

At time t_3 (third column), the target process has been deactivated and the reference process is active (top panel). Its intention node gives excitatory input to all reference color production nodes, bringing them into a dynamic regime where they can be activated (third row, right side). Again, the perceptual boost process is activated and gives homogeneous, excitatory input to the selective spatial attention field and the multi-peak spatial attention field. However, since the reference process does not also activate the spatial attention process, the selective spatial attention field remains below threshold and only the multi-peak spatial attention field is able to form a peak. Since it also receives inhibitory input from the target IOR field (not shown), which has formed a peak at the position of the red object, it forms a peak at the position of the remaining green object. Its activation is projected into the reference field, which also forms a peak (sixth row). Since only the selective spatial attention field projects back into the color/space attention field (and motion/space attention field), the color feature of the green object can not yet be extracted. Through the spatial transformations, the spatial relation CoS field receives a subtreshold bump input that reflects the relative position of the red object with respect to the green object. At time t_3, a peak is forming at that position because the subthreshold bump overlaps with subtreshold input that encodes the meaning of the spatial relation TO THE LEFT OF (last row). Due to this overlap, the spatial relation production node for the same spatial relational concept receives excitatory input and gets a competitive advantage over the other spatial relation production nodes (fourth row). The peak in the spatial relation CoS field activates the CoS node of the spatial relational field process. This inhibits two precondition nodes. The first enables the 'spatial memory node process' to become active, giving homogeneous input to all spatial relation memory nodes. Since they are mutually coupled with inhibitory connections, this leads to a selection decision, where only the node with the strongest input is activated. At time t_3, the nodes have not yet received this input, but it is visible that the spatial relation production node for the relation TO THE LEFT OF is activated most strongly and will activate its corresponding spatial relation memory node. The second precondition node that is deactivated when the

spatial relation CoS field forms a peak enables the spatial attention process to become active, which brings the selective spatial attention field into a regime where it can form a peak. It receives input both from the color/space perception field and, indirectly, from the spatial relation CoS field (via the inverse transformations). Both inputs are localized and centered on the position of the green reference object, forming a peak in the selective spatial attention field. Its activation is projected into the color/space attention field, which in turn projects into the color CoS field. This activates the reference color production node for the color GREEN, which has not yet happened at time t_3.

At time t_4 (fourth column), the reference process is deactivated (top panel). Before the process deactivated, the reference color production node for the color GREEN activated its corresponding memory node, which now remains active. Similarly, the spatial relation memory node for the relation TO THE LEFT OF has activated because it received the strongest input from its corresponding spatial relation production node. At this moment, the model has generated a symbolic description of the scene by activating the target color memory node for the color RED, the reference color memory node for the color GREEN, and the spatial relation memory node for the relation TO THE LEFT OF. This task output corresponds to the phrase "a red object to the left of a green object".

Example of describing a dynamic scene with multiple objects

The following example describes the processes that unfold in the model when it describes a scene that contains a stationary blue object, a stationary green object, and a red object that is moving toward the green object (test D9 in Table 4.9).[29] Figure 4.14 shows how the activation in the model evolves over the course of the test. It is structured analogously to Figure 4.13 with an additional panel (seventh row) that shows the activation of the relational candidates field at four points $t_1, \ldots, t_4$ in time. The activation of the field is visualized using the same color map (bottom right of Figure 4.14) as for all other plots.

[29] The video used in this test is video 3.01b. Please refer to Appendix C, in particular Figure C.21 for more information.

At the beginning of the test, the model is already being simulated and is receiving visual input from the video. As task input, the user activates the prior intention node of the describe process. The processes that subsequently happen in the model are very similar to the previous example.

At time t_1 (first column), since the model detects motion in the scene, the spatial relation production nodes receive additional input to the one they are receiving from the intention node of the describe process. Inversely to that input, the spatial relation pro-

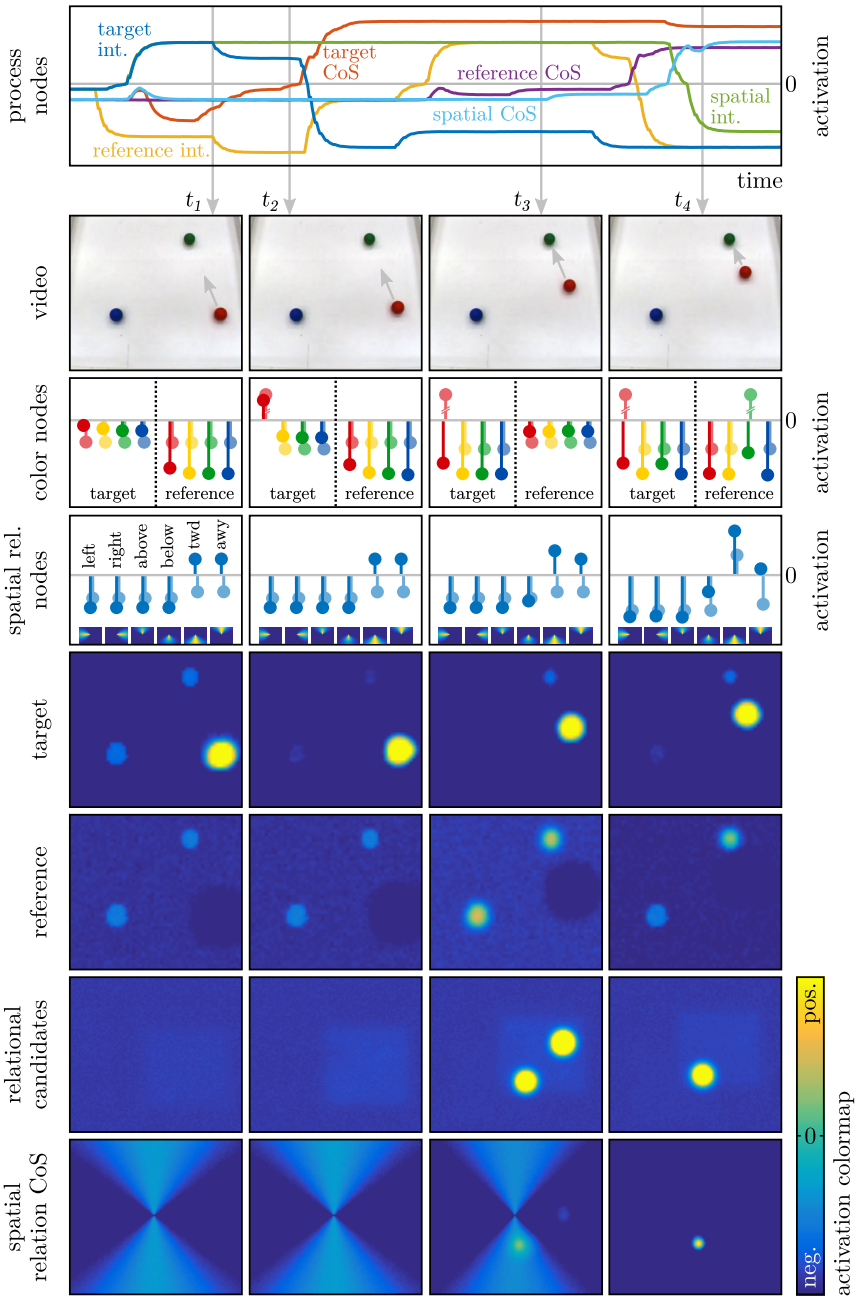


FIGURE 4.14: Generating a description of an object using a movement relation with respect to a second object. See text for more detail.

duction nodes that represent spatial relations (i.e., TO THE LEFT OF, TO THE RIGHT OF, ABOVE, BELOW) receive inhibitory input while the spatial relation production nodes that represent movement relations (i.e., TOWARD, AWAY) receive excitatory input.[30]. This activates the latter nodes while the former remain below threshold. The nodes project into the spatial relation CoS field via patterned synaptic connections, forming a subthreshold pattern of activation that has larger values in areas that match the known spatial relations (last row). The selective spatial attention field forms a peak at the position of the red object; it is the most salient object in the scene because it is mov-

[30] See Section 3.4.

ing. Its spatial position is then represented as a peak in the target field (fifth row).

At time t_2 (second column), the color feature of the red object has been extracted. This is visible by the active target color production node and target color memory node of the color RED (third row, left side). The model has also extracted the movement direction of the object and has activated the target motion production node and target motion memory node of the motion direction UPWARD (not shown). Even though the red object is not moving straight upward, that concept of motion direction fit best.

At time t_3 (third column), the target process is deactivated and the reference process is active. It is bringing all remaining objects in the scene into the attentional foreground, which form peaks first in the multi-peak spatial attention field (not shown) and later in the reference field (sixth row). Through the spatial transformations, the relational candidates field (seventh row) receives input that reflects the position of the target object (represented in the target field) with respect to all reference objects (represented in the reference field). In projecting into the spatial relation CoS field (last row), the representation in the relational candidates field are transformed further, essentially rotating them around the center of the field. The angle of rotation is determined by the motion direction of the target object, which has been extracted earlier. In the spatial relation CoS field, one of the subthreshold bumps of activation overlaps with the subthreshold pattern of activation that encodes the relational concept TOWARD (the lower, triangle-shaped area) and forms a peak. The field makes a selection decision for this spatial position and inhibits the other subthreshold bump. Through inverse transformations, the activation of the spatial relation CoS field is transformed back into the selective spatial attention field, where it overlaps with the spatial position of the green object in the scene. Since the peak in the spatial relation CoS field has also enabled the spatial attention process to become active,[31] the selective spatial attention field is able to form a peak and focus the attention of the model onto the green object alone. This leads to the peak that used to be at the spatial position of the blue object to disappear in the reference field.

[31] See previous example.

At time t_4, this also enables that the feature of the selected reference object is extracted through the color CoS field (which receives activation from the color/space attention field). The reference color production node for the color GREEN is activated (already inactive at time t_4), which in turn activates its corresponding reference color memory node (third row, right side). Similarly, the peak in the spatial relation CoS field has activated the spatial relation memory node for the relation TOWARD.

At this moment, the model has generated a symbolic description

of the scene by activating the target color memory node for the color RED, the target motion memory node for the motion direction UPWARD, the reference color memory node for the color GREEN, and the spatial relation memory node for the relation TOWARD. This task output corresponds to the phrase "a red object that is moving upward and toward a green object".

Discussion 5

The research objective of this thesis is to capture the neural processes of the following three aspects of perceptual grounding: (1) expressing spatial and movement relations (both basic and deictic) in continuous perceptual representations, (2) establishing a mapping between these continuous perceptual representations and discrete representations that may interface with language, and (3) organizing all grounding processes autonomously and based only on neural principles—a particular focus of this work.

The main contribution of this thesis is a neural process model (Section 3) that captures these three aspects of perceptual grounding. A first aspect of this contribution is conceptual work that refines the core component processes that are required for the grounding and describing of spatial and movement relations. This is discussed in detail in the first section of this chapter. A second section discusses more specific contributions of the model and this thesis. These consist of novel neural dynamic implementations that extend previous model of dynamic field theory (DFT) as well as further conceptual and methodological contributions. The chapter is concluded by a discussion of the limitations of this work, as well as suggestions for future research.

5.1 Core component processes

Grounding a phrase in a scene as well as generating a description of a scene require that a number of core component processes are performed. Each of these processes represents a fundamental problem that needs to be solved. The model presented in this thesis explicitly addresses all of these processes, except for the processing

of natural language. Related computational models of grounding also address some of these processes but tend to solve the underlying problems in a different way, most often algorithmically. This section discusses how the core component processes of grounding spatial and movement relations are addressed in the current model and in related work.

Grounding and description tasks consist of the same component processes but require a different sequential order and thus a different process organization. This is because grounding maps a given conceptual representation to a continuous representation while describing establishes the connection in the opposite direction. The following discussion of the core component processes assumes the sequential order of a grounding task, starting from language input.

5.1.1 Language processing

The process of grounding begins with language input, for instance a phrase such as "the red object to the left of the green object". If the language input were natural spoken language, it would require parsing the audio stream into recognized words and building up a grammatical structure for the sentence. This process is often strongly simplified or not addressed at all in models of grounding because it adds too much complexity; the model proposed here is no exception. To be able to work with natural language input, some models connect to (algorithmic) commercial speech recognition software (Gorniak & Roy, 2004; Lallee & Dominey, 2013) and use, for instance, context free grammar to describe the grammatical structure of the given sentence (Roy, 2005b). The words that result from such a speech recognition system can be connected to discrete concepts by learning connections in associative memory networks (Dominey & Boucher, 2005b; Steels & Belpaeme, 2005). Alternatively, the process can be simplified further by connecting written instead of spoken words with discrete concepts (Cangelosi & Riga, 2006). This is also the level at which recent approaches based on convolutional neural networks (CNNs) operate to generate image captions—learning the connection between images and written words (e.g., J. Johnson et al., 2016).

The proposed model simplifies the process of language processing to a similar degree. Here, it is assumed that natural language phrases have already been parsed and their discrete concepts and corresponding roles have been extracted. Concepts are represented bound to their roles (i.e., target or reference) in a structure of dynamic neural nodes. The process of grounding is thus modeled beginning from an amodal representation of concepts, not from language. Concepts that refer to features, such as color or motion direc-

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470; Lallee, S. & Dominey, P. F. (2013). Multi-modal convergence maps: From body schema and self-representation to mental imagery. *Adaptive Behavior*, *21*(4), 274–285

Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205

Dominey, P. F. & Boucher, J. D. (2005b). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1-2), 31–61; Steels, L. & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–489, 469–489

Cangelosi, A. & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, *30*(4), 673–689

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE

tion, correspond to open-class words, content words that are open to new members. Representing concepts by patterned connections between nodes and fields ensures that new concepts can easily be learned based on neural learning rules. The nodes would either have direct, patterned connections to fields, or they could be grounded indirectly based on other nodes (Cangelosi & Harnad, 2001). Some concepts that correspond to closed-class words can also be expressed by concept nodes, as shown here for the spatial and movement relations. However, it is assumed that most other closed-class words, such as conjunctions or determiners, would not be expressed as concept nodes. Instead, they would influence the process organization system or the attentional system. This is similar to how Dominey and Boucher (2005a) employ closed-class words to form a representation of a grammatical construction or how Roy (2005b) uses them to determine speech act classes: they shape how open-class words in a phrase are grounded.

Cangelosi, A. & Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, *4*(1)

Dominey, P. F. & Boucher, J. D. (2005a). Developmental stages of perception and language acquisition in a perceptually grounded robot. *Cognitive Systems Research*, *6*(3), 243–259

Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205

5.1.2 Concept grounding

Grounding language entails that a mapping is established between amodal representations of concepts, such as RED, and their perceptual meaning in continuous representations. In the current model, the mapping is encoded in bidirectional patterned synaptic connections between dynamic neural nodes, which by themselves correspond to amodal representations, and dynamic neural fields, which hold continuous perceptual representations. The dynamic neural nodes thus enable a categorization of the continuous representation into discrete concepts. The implementation corresponds to a generative model of categories (Roy, 2005a), as the patterned connections between node and field can be thought of as establishing a prototype of a concept that can be instantiated. Here, it is assumed that the patterned connections are fixed and already known to the system, but the substrate they are based on is open to learning. The model implements concepts of color, motion direction, spatial relations, and movement relations. In the results, this mapping is for instance shown in Figure 4.1 (page 88) for the color concept RED: the production node representing RED is active and projects onto the color attention field.

Roy, D. (2005a). Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences*, *9*(8), 389–396

In DFT models, concepts are typically encoded in this way, in particular in previous models of spatial language (e.g., Lipinski et al., 2012). Other models employ similar ideas. Steels and Belpaeme (2005) use feedforward adaptive networks, a modification of radial basis function networks, for the grounding of colors. Each color category is determined by a dedicated adaptive network, which maps the input in a three-dimensional color space to a discrete category

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511

Steels, L. & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–489, 469–489

Cangelosi, A. & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, *30*(4), 673–689

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; Karpathy, A. & Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 664–676

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205

Lallee, S. & Dominey, P. F. (2013). Multimodal convergence maps: From body schema and self-representation to mental imagery. *Adaptive Behavior*, *21*(4), 274–285

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470

Dominey, P. F. & Boucher, J. D. (2005a). Developmental stages of perception and language acquisition in a perceptually grounded robot. *Cognitive Systems Research*, *6*(3), 243–259; Madden, C., Hoen, M., & Dominey, P. F. (2010). A cognitive neuroscience perspective on embodied language for human-robot cooperation. *Brain and Language*, *112*(3), 180–188

Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205

via patterned connections. Similarly, Cangelosi and Riga (2006) employ feedforward neural networks to ground actions, where they learn the connections in the network to map discrete action words to joint angle values of a simulated robot.

Recent fast-paced development in object recognition and image captioning is largely driven by convolutional and recurrent neural networks that learn the mapping between perceptual representations and discrete labels (J. Johnson et al., 2016; Karpathy & Fei-Fei, 2017). The training data consists of large collections of images that have been labeled or captioned by humans. While inspired by neural ideas, the focus of these models is on solving problems rather than on keeping with neural realism. The models are often combined with probabilistic formulations and algorithmic solutions. Their impressive performance is a result of the availability of large training data sets and cheap computational power.

The Semantic Pointer Architecture Unified Network (SPAUN) model developed by Eliasmith et al. (2012) learns discrete concepts of handwritten digits from images using auto-encoders based on Restricted Boltzman Machines. The entire model is based on spiking neurons. Concepts (e.g., the number '5') are represented by semantic pointers, high-dimensional vectors that are compressed representations of much higher-dimensional input (e.g., images). The vectors retain compressed information of what they represent and can thus be compared to other semantic pointers to get a similarity measure. In this view, their representation of concepts still retains some perceptual information and is not entirely symbolic. Like the majority of models, they currently only use a single modality, vision. Lallee and Dominey (2013) link multiple representations of different modalities in an amodal convergence map, a winner-take-all pool of neurons, whose connection to the modality-specific representations is learned. This is similar to what is proposed here, where the amodal convergence map corresponds to the dynamic neural concept nodes. The difference is that their model contains multiple modalities, something not yet addressed here.

Some work employs similar ideas, even though the underlying implementations may be algorithmic and not neurally motivated. A purely algorithmic solution is used by Gorniak and Roy (2004), whose model expresses the perceptual meaning of colors by a probability density function in the three-dimensional RGB space (Gorniak & Roy, 2004). In other cases, the solution seems to be algorithmic but is not further specified (Dominey & Boucher, 2005a; Madden et al., 2010). Roy (2005b) uses an algorithmic implementation as well but describes his model based on the structural principle of schemas, which are composed of analog and categorical beliefs. *Analog beliefs* hold values defined over entire continuous feature di-

mensions and are a similar form of representation as dynamic neural fields. *Categorical beliefs* represent discrete categories or states, analogous to dynamic neural nodes. Different types of connections between analog beliefs, categorical beliefs, and the sensorimotor layer are established by *projections*. These correspond to different types of synaptic connection patterns in DFT. The parallels between the work by Roy (2005b) and DFT raise hope that similarly impressive architectures may be constructed based on neural dynamics.

Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205

5.1.3 Role-filler binding

Throughout the process of grounding, the binding between roles and their fillers must be maintained. In the model proposed here, roles are "target" and "reference", whereas fillers are certain features of objects or the objects themselves. Different parts of the model implement the binding between roles and fillers in different ways. In the memory nodes and production nodes that represent concepts of color and motion direction, the binding is achieved by conjunctive coding, explicitly representing a conjunction of role and filler. The target field and reference field use the same principle and bind the spatial position of the target and reference object to the role that is implicit to the fields. This is, for example, shown in Figure 4.8 (page 111), where the red target object is represented in the target field and the green object is represented in the reference field. The model establishes a connection between the representation in the nodes and the perceptual representation in the fields through the three-dimensional color/space attention field and motion/space attention field. However, since these fields do not encode the roles of objects, the binding between role and filler cannot be expressed in these fields alone. Because of this, the binding is maintained through simultaneous activation: the model sequentially brings first the target object into the attentional foreground, while *at the same time* bringing the target field into a dynamic regime where it can form a peak. It then repeats that process analogously for the reference object. The sequentiality is solved through the process organization system; it ensures that only the field for the currently active role (i.e., target or reference) is in a dynamic regime where it can form a peak.

In computational models of grounding, the role-filler binding problem is typically not discussed. This may be because a majority of the models build on algorithmic methods, where role-filler binding does not present itself as a problem because new variables can easily be created.

In symbolic accounts of cognition, the binding between roles and fillers is not a problem either. For instance, in the relation

LOVES(JOHN, MARY), JOHN and MARY are fillers and the slots in the function LOVES determine the roles. Here, roles and fillers are independent and any kind of filler can fill any slot, unless constrained otherwise. This role-filler independence is required for models to express both relations as well as symbolic representations (Hummel, 2011).

For neural networks, Hummel (2011) discusses different forms of role-filler binding. He argues that neural network approaches must rely primarily on dynamic binding (e.g., by synchronous firing) instead of conjunctive coding because it preserves role-filler independence. However, in these discussions, it is usually assumed that roles have a semantic content (Hummel & Holyoak, 2003, 2005; Doumas & Hummel, 2005). That is, in order to express LOVES(JOHN, MARY), the relation must be represented by the binding of JOHN+LOVER and MARY+BELOVED, where JOHN, MARY, but also LOVER, and BELOVED are expressed by some neural population. This way, the role LOVER, has to be bound to the filler JOHN. This type of binding is also used in the work by Eliasmith et al. (2012). Interestingly, he uses conjunctive coding to bind two semantic pointers together. The high-dimensional vector space, in which the semantic pointers are defined, enables that the bound representations are unbound at a later point in time. This shows that role-filler independence can be achieved even with conjunctive coding.

Although the model introduced in this thesis does face the problem of maintaining the binding between a role and an object, the problem is different from the role-filler binding problem as stated above. In the model proposed here, roles do not have a semantic content. They are thus much closer in spirit to the slots in symbolic architectures as they are to the connectionist ideas of how roles must be expressed (Doumas & Hummel, 2012). This is because roles are represented here by dedicated neural populations that can express the semantic content of a filler, for instance, the target field expressing the spatial position of the target object. The fact that they also represent a role is defined only implicitly, by the connections to other fields. In the model, roles do not have an additional semantic meaning; their meaning is determined by the relational template. For example, if the model expresses the spatial relation "the red object to the left of the green object", then the meaning of the target role is that it describes the object that is to the left of the reference object.

5.1.4 Attention

Apprehending spatial relations between objects requires that those objects are brought into the attentional foreground (Logan, 1994).

Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, *23*(2), 109–118

Hummel, J. E. & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–264; Hummel, J. E. & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture. An overview of the LISA project. *Current Directions in Psychological Science*, *14*(3), 153–157; Doumas, L. A. A. & Hummel, J. E. (2005). A symbolic-connectionist model of relation discovery. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 606–611). Austin, TX: Cognitive Science Society

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205

Doumas, L. A. A. & Hummel, J. E. (2012). Computational models of higher cognition. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (Chap. 5, pp. 52–66). Oxford University Press

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1015–1036

Given a phrase that describes an object, for instance "the red object to the left of the green object", the current model employs the given features and spatial relations for visual search. The given features of the objects are brought into the attentional foreground in a top-down manner, guiding attention to objects that have these features. This is shown in Figure 4.1 (page 88), where attention is focused on red colors and, as a consequence, the red object is attended to.

The attentional system is similar to those in previous models of spatial language (Lipinski et al., 2012; van Hengel et al., 2012) and represents a simplified version of part of a more complex model of scene representation (Schneegans, Spencer, & Schöner, 2015). It also corresponds to the "guidance" input in the Guided Search model for visual search (Wolfe, 2007), which guides the deployment of attention. The selective bottleneck he posits in visual attention is implemented by the selective spatial attention field.

Additionally to the top-down stream of input, which enables visual search, a bottom-up input reflects the saliency of objects. Figure 4.13 (page 128) shows that the model can bring an object into the attentional foreground based on bottom-up input alone. In the figure, the red object is perceived as more salient than the green object. Saliency is based on the color saturation of the object and also whether it is moving or not. Dominey and Boucher (2005b) also use recent motion as a primary measure of bottom-up attention in their model. The bottom-up input of this model's attention system is discussed in some more detail in Section 5.2.1.

When searching for multiple objects that adhere to a certain spatial relation, each object must be attended to individually and sequentially (Franconeri et al., 2012). The results show that the current model abides by this constraint: in Figure 4.8 (page 111) attention is first focused on the red object and then on the green object. The sequentiality is functionally required in the model because there is only a single three-dimensional attention field for each feature, which limits feature search to a single feature per feature dimension. Searching for multiple features (e.g., red and green) at the same time potentially creates binding errors, where the binding between the color and the associated role (e.g., target or reference) is lost. This is because the three-dimensional attention fields bind different features into a coherent object representation through shared spatial dimensions, a solution to the fundamental neural binding problem (Treisman & Gelade, 1980). In a previous version of the spatial language model, parallel search for multiple features was possible because there were independent feature search mechanisms for target and reference (Lipinski et al., 2009). Lipinski et al. (2012) introduced a single feature search mechanism and maintained the binding by searching for the target and reference object sequentially.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511; van Hengel, U., Sandamirskaya, Y., Schneegans, S., & Schöner, G. (2012). A neural-dynamic architecture for flexible spatial language: Intrinsic frames, the term "between", and autonomy. In *Robot and Human Interactive Communication, 2012 IEEE RO-MAN: The 21st IEEE International Symposium on* (pp. 150–157). IEEE

Schneegans, S., Spencer, J. P., & Schöner, G. (2015). Integrating "what" and "where": Visual working memory for objects in a scene. In G. Schöner & J. P. Spencer (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (Chap. 8, pp. 197–226). New York: Oxford University Press

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99–119). New York: Oxford University Press

Dominey, P. F. & Boucher, J. D. (2005b). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1-2), 31–61

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136

Lipinski, J., Sandamirskaya, Y., & Schöner, G. (2009). Swing it to the left, swing it to the right: Enacting flexible spatial language using a neurodynamic framework. *Cognitive Neurodynamics*, *3*(4), 373–400

This was done manually by the user giving inputs into the model during its performance. In the current model, the sequentiality that solves the binding problem is implemented through the principles of process organization, which is discussed later.

5.1.5 Working memory representations

In order to sequentially ground and attend to objects, stable representations of these objects have to be kept in working memory. In the model, a stable mental representation is built up for every object that is attended to. This is shown in Figure 4.8 (page 111), where the peak in the target field is sustained even when the attentional focus shifts to the reference object. Such sustained peaks are achieved through lateral interactions (local excitation and mid-range inhibition) within a field. Similarly, Eliasmith et al. (2012) use recurrent attractor neural networks to model working memory, where the recurrent connections lead to sustained activation in the absence of input.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205

Additionally, it is required that working memory representations are updated when changes occur in the scene. This is shown in Figure 4.10 (page 117), where the peak in the target field tracks the moving object it represents. It is achieved through continuous input from the color/space perception field and motion/space perception field, which drags the self-sustained peak along the current position of the object.

5.1.6 Reference frame transformation

One of the key challenges to apprehending spatial relations between objects is to form a representation of their relative positions (Logan & Sadler, 1996). It is a challenge because it requires that the reference frame of the object representations is adjusted. This is solved in the proposed model and shown, for instance, in Figure 4.8 (page 111), where the representation of the spatial position of the target object is brought into a space that is centered on the position of the reference object (last row, third column of the figure); this amounts to a transformation shifting the reference frame. In order to apprehend movement relations between objects, the reference frame is transformed further, essentially rotating it to align the spatial representation of the objects with the motion direction of the target object. This is shown in Figure 4.10 (page 117) (last row, third column). In the current model, the reference frame is adjusted using convolution operations that approximate steerable neural mappings (Schneegans & Schöner, 2012). Convolutions are used here only because they are computationally less expensive than

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13, pp. 493–529). Cambridge, MA, USA: MIT Press

Schneegans, S. & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109

explicit implementations of steerable neural mappings. The way the convolutions are used ensures they can be replaced by steerable neural mappings without impairing the functionality of the model. This is not true for previous DFT models, which is discussed in more detail in Section 5.2.5.

The current model does not address intrinsic relations between objects, where the spatial relation is established relative to the intrinsic reference frame of one of the objects. This requires that the intrinsic reference frame has to be extracted as well, which is captured by a previous DFT model (van Hengel et al., 2012). In their model, the rotation of the reference frame is based on the same mechanism used here, a shift in polar coordinates using a convolution. The angle by which the reference frame is rotated is determined by matching rotated versions of the reference object against a canonical view of it.

There is strikingly little work on the mechanisms by which spatial relations between separate objects are extracted (Franconeri et al., 2012). Most computational models of grounding that also address spatial relations implement the adjustment of the reference frame algorithmically and do not discuss that a neural implementation may represent a challenge (Regier, 1992, 1995; Gorniak & Roy, 2004; Roy, 2005b; Dominey & Boucher, 2005b). This is also true for recent architectures that employ CNNs. Take, for instance, the task of generating a caption or description for an image, which requires that relations between people or objects in the scene are described. Current architectures based on CNN that solve this task most typically do not explicitly address how relational information is extracted from the image (J. Johnson et al., 2016; Karpathy & Fei-Fei, 2017). The training data for the networks consists of images and human-generated image captions, either for the entire image or for a certain region. The given captions already contain descriptions that are based on relations, for instance "man playing tennis outside" (J. Johnson et al., 2016). It seems that the networks learn not the explicit relations but are able to express the similarity of an image with the description as a whole. An algorithmic solution is used by J. Johnson et al. (2015), who directly train their model with relational information, computed algorithmically from the position and size of the objects' bounding boxes. Their model finds images that fit complex descriptions that are given not in the form of language but as a scene graph.

Lu et al. (2016) extract relations with a CNN based on the spatial arrangement of objects' bounding boxes. This is not explained further but seems to imply that the relative position between the bounding boxes (and possibly their relative size) is learned. The similarity of relations between different pairs of objects, for instance

Regier, T. (1992). *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization* (tech. rep. No. TR-92-062). International Computer Science Institute. Berkeley; Regier, T. (1995). A model of the human capacity for categorizing spatial relations. *Cognitive Linguistics*, *6*(1), 63–88; Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470; Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205; Dominey, P. F. & Boucher, J. D. (2005b). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1-2), 31–61

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; Karpathy, A. & Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 664–676

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2015). Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3668–3678). IEEE

Lu, C., Krishna, R., Bernstein, M., & Fei-Fei, L. (2016). Visual relationship detection with language priors. In *European Conference on Computer Vision* (pp. 852–869). Springer International Publishing

"man-riding-horse" and "man-riding-elephant", is computed based on the similarity of the labels. This is done by comparing a vector representation of the labels.

Previous DFT models that deal with spatial relations build on the same or similar approaches as used here. The differences and the contribution of the current work in that regard are discussed in Section 5.2.5.

5.1.7 Matching relations

Grounding a relational phrase entails evaluating how well the relative position of objects fits with one or multiple relations (Logan & Sadler, 1996). This first requires a representation of the perceptual meaning of the relations themselves (Logan & Sadler, 1996). In the model proposed here, relations are represented by spatial templates that are encoded as patterned connections between discrete concept nodes and a neural field that is defined over continuous spatial dimensions. This is shown, for instance, in Figure 4.8 (page 111), where the spatial template for the relation TO THE LEFT OF is projected into the spatial relation CoS field. The shape of the patterned synaptic connections is inspired by behavioral data by Logan and Sadler (1996) and has been used in previous DFT models that address spatial relations (e.g., Lipinski et al., 2012).

In the majority of computational models that address spatial relations, the spatial templates are not explicitly represented. Instead they are tightly interwoven with algorithmic mechanisms that match object positions to relations. In some cases, the exact algorithmic method is not specified (Gorniak & Roy, 2004; Roy, 2005b). Other approaches state that they compute the match based on how much the target object deviates from a reference orientation, for instance 90 degrees for the relation ABOVE (Regier, 1992, 1995), or that the match is based on both orientation and distance (Dominey & Boucher, 2005b). The attentional vector sum (AVS) model introduced by Regier and Carlson (2001) uses a measure that is based on orientation and distance as well, but the orientation depends on two components, the orientation between the respective centers-of-mass of the target and reference object and their proximal orientation, determined by the angle of the vector between their closest points.

In the current model and previous DFT models of spatial language (Lipinski et al., 2009, 2012; van Hengel et al., 2012), matching object positions with relations requires a more elaborate process because it is based on explicit representations of both the relative object positions and the spatial relation. Here, a field (spatial relation CoS field) receives subthreshold input reflecting both of these

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13, pp. 493–529). Cambridge, MA, USA: MIT Press

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470; Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205

Regier, T. (1992). *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization* (tech. rep. No. TR-92-062). International Computer Science Institute. Berkeley; Regier, T. (1995). A model of the human capacity for categorizing spatial relations. *Cognitive Linguistics*, *6*(1), 63–88

Dominey, P. F. & Boucher, J. D. (2005b). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1-2), 31–61

representations. The field forms a peak when the two inputs overlap and signals that the objects match the spatial relation.

If the inputs do not overlap, the objects do not fit to the spatial template. In this case, other objects have to be selected and the process must be repeated. This is a form of hypothesis testing: a hypothesis is established that the selected objects match the specified relation; this hypothesis is tested and can either be accepted or rejected. Hypothesis testing is novel to the model proposed here and is discussed in more detail in Section 5.2.4.

5.1.8 Process organization

The biggest challenge in creating a model that can solve all the problems named above is to organize its processes based on neural principles. This is also a core component process, albeit one that organizes all other component processes.

Process organization includes the fundamental neural problem of generating discrete processing steps from dynamics that evolve in continuous real time. In DFT, instabilities in the time continuous dynamics give rise to discrete events that can be used to activate and deactivate processes (Sandamirskaya & Schöner, 2010). In algorithmic information-processing approaches, generating discrete processing steps is not a problem because the processing of finite amounts of data inherently has a defined beginning and end that can be used to trigger new processes. In neural approaches, the problem of generating discrete processing steps is also often not addressed, typically because this part is controlled algorithmically (e.g., Cangelosi & Harnad, 2001).[1]

In computational models of grounding, the problem of process organization is often ignored or not explicitly modeled. Processes are either organized directly through common algorithmic tools or based on structural principles. For example, in the work of Deb Roy, which is based on schema theory, aspects of process organization are in many cases explicitly addressed. Roy (2005b) builds its organization on top of categorical beliefs, which are discrete, conceptual representations similar to dynamic neural nodes. These categorical beliefs can represent the different outcomes of actions, similar to the condition of satisfaction (CoS) and condition of dissatisfaction (CoD) in the model presented here. This enables both sequential and parallel execution of processes. Roy (2008) states that he employs precondition rules for sequences of schemas, which is similar to the precondition nodes used in the model presented here. However, all of the principles of schema theory are implemented algorithmically and there is no connection to neural principles. This is sometimes augmented with purely algorithmic solutions, for in-

Sandamirskaya, Y. & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179

Cangelosi, A. & Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, *4*(1)

[1]Please note that in this case the discrete processing steps do not refer to the discrete time steps often used to implement the update of neural networks. They refer, instead, to more macroscopic processes, like beginning a new training phase or behavior.

Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205

Roy, D. (2008). A mechanistic model of three facets of meaning. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and Embodiment: Debates on Meaning and Cognition* (pp. 1–32). Oxford, UK: Oxford University Press

stance the use of decision trees for hypothesis testing (Mavridis & Roy, 2006).

But even models that are in part based on neural principles rely on algorithmic solutions for the organization of processes (e.g., Cangelosi & Harnad, 2001; Steels & Belpaeme, 2005; Dominey & Boucher, 2005b). While parts of these models are based on neural networks, it often remains unclear how the networks are controlled to exhibit different behaviors in time or how they would generate sequential behavior. In order to reduce the complexity of the models, these processes are organized by algorithmic tools that lie outside of the models themselves.

Shastri (1999), Shastri et al. (2002) give us a sense of how complex even simple systems for process organization can become when based on neural principles. With their SHRUTI system, they developed a connectionist implementation of x-schemas, which determine detailed procedures for actions such as grasping as well as more general relations. Each action or relation is represented by a neural structure that includes a representation of whether the relation applies or not, similar to the idea of CoS and CoD. In their system, this representation is structurally repeated for every relation, enabling the system to represent many relations simultaneously. The model presented here, in contrast, only has a single such structure and verifies relations between different objects in a sequential manner. While SHRUTI implements schemas based on neural principles, it is unclear how grounded its representations are; it seems that apart from a sampled spatial representation, other representations are symbolic.

In the fully neural SPAUN model by Eliasmith et al. (2012), processes are organized based on a model of the basal ganglia. Like the process organization system proposed here, it has connections into and from every part of the model. Its activation manipulates the flow of activation between different parts of the model.

For the model presented here, the process organization system is an integral part. Its specific contribution is further discussed in the next section, comparing it in more detail to previous DFT models of spatial language.

5.2 Specific contributions

This section discusses the specific contributions this thesis makes to the understanding of perceptual grounding in general and to DFT in particular. A first specific contribution regards the core component processes discussed in the previous section. They were in part previously established and implemented in DFT, although

Mavridis, N. & Roy, D. (2006). Grounded situation models for robots: Where words and percepts meet. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on* (pp. 4690–4697). IEEE

Cangelosi, A. & Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, *4*(1); Steels, L. & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–489, 469–489; Dominey, P. F. & Boucher, J. D. (2005b). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1-2), 31–61

Shastri, L. (1999). Advances in SHRUTI: A neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence*, *11*, 79–108; Shastri, L., Grannes, D., Narayanan, S. S., & Feldman, J. (2002). *A Connectionist Encoding of Parameterized Schemas and Reactive Plans* (tech. rep. No. TR-02-008). International Computer Science Institute. Berkeley

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205

sometimes in an ad-hoc way. This thesis thus contributes the conceptual work of refining the required core component processes of grounding and putting them in the context of the literature. Further specific contributions that include both conceptual work as well as novel implementations within the model are discussed next.

5.2.1 Grounding and describing

As a conceptual contribution, this thesis clarifies the three types of tasks relevant to grounding: *Grounding tasks* consist of matching a given phrase to a visual scene, where the phrase describes all necessary features of objects, for example "the red object to the left of the green object". *Description tasks* consist of generating such a phrase from a given visual scene, where no other information is given about what is to be described. In *mixed tasks*, partial information of a relation is given and the model must respond with the missing information. The model could, for instance, be given input such as "left of green" and would have to respond by activating concepts that best described the object to the left of the green object in the scene. There are different variants of mixed tasks, depending on whether or not the phrase specifies the target object, the reference object, and the spatial term. All mixed tasks contain elements of grounding tasks (i.e., searching for described objects) and description tasks (i.e., describing parts of the scene).

The model introduced here captures both grounding tasks and description tasks. Neither type of task was captured in previous DFT models; they instead addressed mixed tasks. For instance, Lipinski et al. (2012) show that given the question "Where is the red object relative to the blue object?", their model can respond by activating a node that corresponds to the relation ABOVE.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511

Compared to the mixed tasks performed by previous models, grounding tasks are easier because they do not require a response. In fact, it is likely that most previous models of spatial language could perform some simple grounding tasks as well, although this was not demonstrated. What makes grounding tasks particularly interesting is that they invite many cases in which the phrase does not match the scene well, or not at all. With relation to a visual scene, a phrase can either *unambiguously* specify a single object in the scene, it can *ambiguously* specify an object in the scene, such that multiple objects fit the description, or it can specify an object that cannot be found in the scene, thus leading to a *mismatch*. Handling these cases requires that the model is able to detect them and react accordingly. This is mostly a requirement of the model's process organization, which is a particular focus and specific contribution of this thesis, discussed later in this section. Cases where the given

phrase is ambiguous with respect to the scene or where it does not match are not addressed by previous models. In their demonstrated examples, the phrase always uniquely describes one of the objects in the scene. The model proposed here is able to handle cases where the phrase either unambiguously or ambiguously specifies an object in the scene. It is also tested on scenes where the phrase does not match any object. In these scenes, the model searches until it has tried all potential candidate objects, but it is currently missing a mechanism to detect that there are no more objects to test.

Additional to grounding tasks, the current model also captures description tasks, where a full conceptual representation (or phrase) is generated from visual input. This is also not demonstrated by previous models. Lipinski et al. (2012) show that they can generate a description for a *given* object, by selecting a fitting spatial term as well as a reference object. Generating an *entire* description additionally requires that objects are brought into the attentional foreground based on bottom-up input alone. To support this in the current model, a bottom-up saliency mechanism is added to its attentional system.

The bottom-up path of the model's attentional system covers some of the following aspects that are commonly addressed by computational models of visual attention (Itti & Koch, 2001). First, the attentional system is based on a pre-attentive computation of visual features (i.e., color and motion direction). Second, all feature representations feed into a unique saliency map, which is represented by the input into the spatial attention mechanism (multi-peak spatial attention field and selective spatial attention field). The spatial selectivity that is often associated with saliency maps is implemented by the lateral interaction in the selective spatial attention field as well as the field's coupling to the multi-peak spatial attention field. Without top-down input, the saliency of objects depends on the size of the objects, their color saturation, and whether they are moving or not. This part of the model is understood as a placeholder for a more realistic visual saliency mechanism, one that is guided most by feature contrast rather than absolute feature strength. Third, in order to sequentially shift the attentional focus to multiple salient objects, an inhibition-of-return (IOR) mechanism is employed to inhibit the location that is currently attended. In visual attention models such a mechanism explains the formation of attentional scanpaths. Here, it is implemented by the target IOR field. However, its function is more specialized because it is specific to the spatial position of the target object. Wolfe (2007) postulate that the IOR mechanism has a capacity limit. In the tests of the current model, the target IOR field never reached its capacity limit, but it has one due to the lateral interaction within the field. The model does not address the

Itti, L. & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99–119). New York: Oxford University Press

two remaining aspects of visual attention that Itti and Koch (2001) mention: how attention is connected to eye movements and how attention is constrained by scene understanding and object recognition.

Itti, L. & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203

5.2.2 Movement relations

The model proposed in this thesis is the first DFT model to capture how movement relations may be extracted from a scene and expressed in perceptual and amodal representations. This is the basis for expressing movement verbs and actions, which are a large part of language (Pulvermüller, 2005). The model captures movement relations in addition to spatial relations, which were addressed in previous models (e.g., Lipinski et al., 2012). Movement relations were used in a previous prototypical DFT architecture that is able to describe simple object-oriented actions like reaching, grasping, and dropping (Lobato et al., 2015). While parts of the model are based on DFT, the part that extracts the movement relations between objects is implemented algorithmically. In the model proposed here, movement relations are expressed in a neurally plausible way, based on the same principle as for spatial relations: each movement relation is expressed as a (static) template that is imposed on the relative position of objects. This requires that the representation of the objects' positions is not only transformed to center them on a reference position, but also to align them with the motion direction of the moving object, essentially rotating the frame of reference. This is shown in Figure 4.10 (page 117), where the selected target object is moving from right to left, toward a green object. The spatial template for the relational concept TOWARD is fixed, but the reference frame is translated and rotated depending on the position and movement direction of the target object. For stationary scenes, the reference frame is adjusted in the same way but with a fixed rotation of zero degrees. Both reference frame transformations can neurally be implemented by steerable neural mappings; in the implementation of the model they are solved by convolutions as a computational shortcut.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(July), 576–582

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511

Lobato, D., Sandamirskaya, Y., Richter, M., & Schöner, G. (2015). Parsing of action sequences: A neural dynamics approach. *Paladyn, Journal of Behavioral Robotics*, *6*, 119–135

In order to align the reference frame with the motion direction of an object, that motion direction must be extracted from the visual input. In the current model, this is done based on a neurally plausible implementation (Berger et al., 2012) of the counter change model of motion perception (Hock et al., 2009). The model integrates motion direction into the attentional system as an additional feature dimension. Where previous models of spatial language only use color as a feature to guide the attention of the model, here, motion direction may be used as well. This also requires that conjunc-

Berger, M., Faubel, C., Norman, J., Hock, H., & Schöner, G. (2012). The counter-change model of motion perception: An account based on dynamic field theory. In A. E. P. Villa (Ed.), *ICANN 2012, Part I, LNCS 7552* (pp. 579–586). Berlin Heidelberg: Springer

Hock, H. S., Schöner, G., & Gilroy, L. (2009). A counterchange mechanism for the perception of motion. *Acta Psychologica*, *132*(1), 1–21

tion searches are handled correctly, for instance searching for a red object that is moving rightward. Figure 4.4 (page 99) and other examples in Section 4.1.2 show that this is possible in the model.

Finally, dealing with scenes in which objects move requires that representations of the objects' spatial positions are continuously updated, even when they are not currently attended to. This is, for instance, shown in Figure 4.10 (page 117), where the peak in the target field tracks the moving object it represents. Tracking is achieved through continuous input from the color/space perception field and motion/space perception field, which drags the self-sustained peak along the current position of the object.

5.2.3 Process organization

A particular focus of the model proposed in this thesis is that it organizes all of its processes based on neural principles. The flexibility of the process organization system is demonstrated by the number of different tests that the model is able to capture, which differ along the following characteristics: grounding tasks and description tasks; basic relations and deictic relations; spatial relations and movement relations; different number of matching target and reference object (some require hypothesis testing, others do not). The model performs all of the tests autonomously, that is, without user interference and only based on visual input and an initial task input.

Previous models of spatial language have different approaches to solving the problem of process organization. One model is actually completely autonomous in that it does not require user input after the initial task input (Lipinski et al., 2009). However, this is only possible because the grounding of the target object is independent from the grounding of the reference object; the objects could even be grounded at the same time. This is because there are dedicated attention fields for the target object and the reference object. Additionally, the spatial transformations are based on convolutions of the spatial terms, which simplifies the process organization: since the match between spatial templates and object positions can be performed in camera coordinates, it only requires a single transformation. In contrast, the current model requires that objects are transformed into a space centered on a reference position (and aligned with motion direction) to match with spatial templates, and that they are then transformed back into camera coordinates to give a response. Overall, the model by Lipinski et al. (2009) is less complex and does not support as many different tasks as the model introduced here. Lipinski et al. (2012) improved upon these shortcomings, but their model in turn does not have any process organization.

Lipinski, J., Sandamirskaya, Y., & Schöner, G. (2009). Swing it to the left, swing it to the right: Enacting flexible spatial language using a neurodynamic framework. *Cognitive Neurodynamics*, *3*(4), 373–400

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511

Most of the processes involved in grounding different objects and their relations are manually activated and deactivated by the user. Another version by van Hengel et al. (2012) builds upon this work and uses the DFT model of serial order (Sandamirskaya & Schöner, 2010) to organize processes in a sequence. While this automates the processes and no longer requires manual input by the user, the different processes are not made explicit and cannot form complex sequences. To express more complex sequences, Durán et al. (2012) propose a hierarchical system that organizes processes and behaviors. Their hierarchical model enables chunks of elementary behaviors (EBs) to be organized in a sequence and activated by an EB on a higher hierarchical level. Since the sequence generation within each chunk is based on the serial order mechanism (Sandamirskaya & Schöner, 2010), it has the same drawbacks. First, it has little flexibility with regards to what sequences can be expressed. Within a chunk, only one EB can be active at the same time. Moreover, if EBs are organized into separate chunks, for instance to activate them in parallel, there is no mechanism to enforce a sequential constraint between any EBs in different chunks. Second, it requires changing synaptic weights in order to change established chunks. This is because the input to the CoS node on a higher hierarchical level is normalized depending on the number of EBs in a chunk (every connection going into the CoS node is divided by the number of EBs in a chunk). Thus, when a new EB is added to a chunk, the weights of all connections to the higher-level CoS node must be changed. Furthermore, since two EBs in a chunk have a direct synaptic connection, they can only ever be used in that specified sequential order. Expressing the opposite sequential order requires to change the synaptic connections or to have different copies of the EBs for different contexts.

van Hengel, U., Sandamirskaya, Y., Schneegans, S., & Schöner, G. (2012). A neural-dynamic architecture for flexible spatial language: Intrinsic frames, the term "between", and autonomy. In *Robot and Human Interactive Communication, 2012 IEEE RO-MAN: The 21st IEEE International Symposium on* (pp. 150–157). IEEE

Sandamirskaya, Y. & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179

Durán, B., Sandamirskaya, Y., & Schöner, G. (2012). A dynamic field architecture for the generation of hierarchically organized sequences. In A. E. P. Villa, W. Duch, P. Érdi, F. Masulli, & G. Palm (Eds.), *Artificial neural networks and machine learning - icann 2012* (pp. 25–32). Berlin, Heidelberg: Springer

Sandamirskaya, Y. & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179

The current model introduces a process organization system that enables the *reuse* of processes in different contexts. This is also achieved by creating a hierarchy (or heterarchy) of processes. The system proposed here is more flexible because it is based on the principles of behavioral organization (Section 2.2.7), which enables the expression of sequential constraints that can be activated and deactivated without changing synaptic weights. In addition, the structure of how processes are represented enables that the CoS of a process on a higher hierarchical level does not depend on the number of processes it activates on lower levels; the normalization of the input is achieved through the way processes are coupled.

5.2.4 Hypothesis testing

An additional novelty of the process organization system is the use of a CoD to assess the match of a relation. This supports a form of hypothesis testing: a hypothesis is established that the selected objects match a given relation; this hypothesis is tested and can either be accepted or rejected. Hypothesis testing is shown in Figure 4.9 (page 114), where the model first selects a target object that does not match the specified relation. Its relative position with respect to the reference object does not overlap with the spatial template (in the spatial relation CoS field); instead it has overlap in the spatial relation CoD field and forms a peak there. Selecting a different target object is solved by an IOR mechanism. Repeating the process of testing hypotheses for other objects is controlled based on the principles of process organization. Hypothesis testing has not been shown in this form in other computational models of grounding relations.

5.2.5 Matching of relational templates

Compared to all previous models of spatial language, the current model is further constrained with regards to how it deals with relational templates. Previous models employed reference frame transformations to transform the spatial templates directly into camera coordinates (Lipinski et al., 2009). This is only possible when using convolution operations to approximate steerable neural mappings because otherwise the spatial template would become distorted by the sigmoidal output function of multiple fields; it is thus neurally implausible. This could possibly be fixed by using linear sigmoidal output functions in multiple fields of the model but it is unclear whether this could capture the same functionality. A possible alternative is to induce a peak from the spatial template and transform that into camera coordinates (Lipinski et al., 2012). However, while this is neurally plausible, it has the disadvantage that the detailed shape of the spatial template is lost and cannot have an effect on the selection of the object.

The alternative solution implemented in the model proposed here is to transform the spatial representation of all relevant objects into the spatial relation CoS field, match them against the spatial template there, and transform the matching object back into the original camera coordinates to give a response. This enables the shape of the spatial template to guide the selection of the object in the spatial relation CoS field. This fits with data indicating that in selecting a reference object, people are influenced less by its saliency and rather by its alignment with spatial relations relative to the tar-

Lipinski, J., Sandamirskaya, Y., & Schöner, G. (2009). Swing it to the left, swing it to the right: Enacting flexible spatial language using a neurodynamic framework. *Cognitive Neurodynamics*, *3*(4), 373–400

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511

get object (Carlson & Hill, 2008). The solution has the drawback that it makes the process of grounding more complex; it requires additional fields (e.g., the multi-peak spatial attention field) and additional process organization to ensure that a response is only given once a peak has formed in the spatial relation CoS field.

Carlson, L. A. & Hill, P. L. (2008). Processing the presence, placement, and properties of a distractor in spatial language tasks. *Memory & Cognition*, *36*(2), 240–255

5.2.6 Roles and role-filler binding

The model proposed in this thesis introduces an architecture of dynamic neural nodes that holds an amodal representation of a relational phrase. The particular structure of this architecture provides a memory node and a production node for every concept in every role it could appear in. The memory node is required to represent that a concept is part of the current phrase, while the production node is required to gate the influence of the concept on the rest of the model. The copies for each role are required because the model organizes its own processes and must have access to the concrete binding between roles and fillers. This binding must be maintained throughout the whole grounding process.

In previous models, the binding between features (or objects) and roles was done implicitly, either by dedicated feature attention fields for each role (Lipinski et al., 2009), by manually activating feature concepts and the corresponding field for the role at the same time (Lipinski et al., 2012), or through a sequence generation model (van Hengel et al., 2012).

Lipinski, J., Sandamirskaya, Y., & Schöner, G. (2009). Swing it to the left, swing it to the right: Enacting flexible spatial language using a neurodynamic framework. *Cognitive Neurodynamics*, *3*(4), 373–400

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511

van Hengel, U., Sandamirskaya, Y., Schneegans, S., & Schöner, G. (2012). A neural-dynamic architecture for flexible spatial language: Intrinsic frames, the term "between", and autonomy. In *Robot and Human Interactive Communication, 2012 IEEE RO-MAN: The 21st IEEE International Symposium on* (pp. 150–157). IEEE

5.2.7 Extensive qualitative testing

Demonstrating the performance of DFT models is often hard, in particular if the models are complex and abstract. This is also true for the model proposed here. Statistical evaluations of its performance could be done but are not meaningful because in building the model, there are many degrees of freedom, many parameters, but only few constraints. Given enough time, the model could be parameterized to perform perfectly in all tested scenarios. Fitting empirical data and showing a close match is problematic for the same reason. What seems most meaningful is to give qualitative results that systematically demonstrate what the model is able to do and what it is not. Compared to previous models, the results showed here (Section 4) are extensive and clearly show how flexible the model is in performing various tasks in different visual scenes.

The capabilities of the model were demonstrated in 104 tests that systematically probed it in qualitatively different tasks and visual environments. For each test, it was presented with a video of colored balls on a white table surface. The task of the model was

to either match a given phrase to the corresponding objects in the scene or generate a phrase about the scene. First, the grounding of basic relations was examined, where single objects with simple features such as color and motion direction were grounded (tests G1–G57 in Section 4.1.1). Second, the grounding of deictic relations between pairs of objects was examined, both spatial relations (i.e., TO THE LEFT, TO THE RIGHT, ABOVE, BELOW; tests G58–G74 in Section 4.1.3) and movement relations (i.e., TOWARD, AWAY; tests G75–G89 in Section 4.1.3). Third, the description of basic relations was examined, where the model described the features of single objects (tests D1–D5 in Section 4.2.1). Fourth, the description of deictic relations was examined, where the model described the features and deictic relations between objects, both spatial relations and movement relations (tests D6–D15 in Section 4.2.2).

As an overall result, the tests show that the model is able to ground the given phrase or describe the scene in all cases where this is possible. After an initial task input is given, the model performs without user intervention. All tests listed in this thesis were performed on the same model, with the same set of parameters. The results thus show that the model captures the neural processes required to perceptually ground spatial relations as well as movement relations.

5.3 Limitations

This section covers potential limitations and weaknesses of the model as well as possible ways to remedy these limitations.

In order to reduce the complexity of the model, some of its parts represent simplified forms of already established DFT models. The following four simplifications could be addressed in future work and would mostly amount to integrative work.

First, early visual processing is reduced to a minimum; it is implemented algorithmically, and object recognition is not addressed. The model assumes that the objects are colored balls on a white background. The segmentation of the scene is based on the hue and saturation channel of the camera. In recognizing objects, neither the shape or scale of objects is taken into account. The representation of objects is limited to two non-spatial feature dimensions, color and motion direction, and the two-dimensional image space. This visual front-end of the model is a placeholder for a model of object recognition that captures how object instances may be learned in a neural dynamic model (Lomp et al., 2017).

Lomp, O., Faubel, C., & Schöner, G. (2017). A neural-dynamic architecture for concurrent estimation of object pose and identity. *Frontiers in Neurorobotics*, *11*(April), 1–17

Second, throughout the proposed model, the representation of the spatial position of objects is in retinal (camera) coordinates. This

is only possible here because the model assumes that the camera is fixed. The human eye, on the other hand, makes several saccades per second, thereby changing the spatial position of objects on the retina. The representation of the spatial position of objects must thus be in allocentric (world) coordinates. This is implemented in the scene representation model of DFT, where the retinal spatial representations are transformed into an allocentric spatial representation that holds objects in working memory (Schneegans, Spencer, & Schöner, 2015).

Schneegans, S., Spencer, J. P., & Schöner, G. (2015). Integrating "what" and "where": Visual working memory for objects in a scene. In G. Schöner & J. P. Spencer (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (Chap. 8, pp. 197–226). New York: Oxford University Press

Third, the current model simplifies visual search. When searching for an individual object, the attentional system always finds the correct object if it is in the scene; it never makes mistakes. This is ensured by inhibition from each feature attention field to the three-dimensional color/space attention field and motion/space attention field. The attentional system is created so reliably here to make the model more deterministic and thus easier to work with. However, it contradicts reaction time experiments for conjunction searches (simultaneous searches for multiple features), which show that reaction time increases with the number of objects in the scene (Treisman & Gelade, 1980). This data suggests that participants sequentially bring objects into the attentional foreground and only afterward check whether they satisfy all specified features. If an object does not match, another object is selected. To be consistent with the data, the model would have to be changed to select objects in a similar manner. This would additionally require a mechanism to detect whether the object satisfies all specified features. This is also captured in the model of scene representation (Schneegans, Spencer, & Schöner, 2015).

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136

Schneegans, S., Spencer, J. P., & Schöner, G. (2015). Integrating "what" and "where": Visual working memory for objects in a scene. In G. Schöner & J. P. Spencer (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (Chap. 8, pp. 197–226). New York: Oxford University Press

Fourth, the current model simplifies spatial relations since it does not address intrinsic relations (Logan & Sadler, 1996), where the intrinsic reference frame of the reference object has an effect on how the relation is perceived. In the visual scenes used here, this is not necessary since none of the objects have an intrinsic reference frame. However, many natural objects do, in particular elongated ones. Modeling intrinsic relations requires that the spatial representation of objects is aligned with the intrinsic reference frame of the reference object. van Hengel et al. (2012) propose a DFT model that uses a convolution to rotate the object representations to align with the intrinsic reference frame. This idea could be integrated into the current model, possibly even using the same rotational transformation used here to align object representations with the motion direction of an object.

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13, pp. 493–529). Cambridge, MA, USA: MIT Press

van Hengel, U., Sandamirskaya, Y., Schneegans, S., & Schöner, G. (2012). A neural-dynamic architecture for flexible spatial language: Intrinsic frames, the term "between", and autonomy. In *Robot and Human Interactive Communication, 2012 IEEE RO-MAN: The 21st IEEE International Symposium on* (pp. 150–157). IEEE

The following minor simplifications in the model require work that may be addressed more quickly.

The model currently does not have a mechanism to detect that

a specified object does not exist in the scene. Neither can it detect that the scene currently does not have any object in it. A possible solution could be akin to a timer, where activation builds up and reaches the threshold after some time of inactivity. If such a mechanism were introduced, for instance for the selective spatial attention field, the model may be able to detect that it had already tried all objects or that the scene did not contain any object at all.

The synaptic connections that encode the meaning of concepts, such as RED, are defined between the production nodes and the feature attention fields, but also between the production nodes and the feature CoS attention fields. In the model, these weights patterns are implemented as independent connection weights that are manually set to the same weight patterns. Such independent connection weights are implausible given that the meaning of concepts must be learned from perceptual experience. This could be solved by encoding the meaning of concepts in patterned synaptic connections between the production nodes and an additional field, which is defined over the same dimension as the feature attention field and feature CoS attention field and is connected to both fields by static one-to-one connections. This way, the meaning of concepts could be learned by adjusting only these patterned connections.

The estimation of the motion direction of objects is currently imprecise and could be improved. This may be due to the low resolution of the camera image,[2] where pixel-changes have a large influence on the estimation of motion direction. Additionally, the motion direction is extracted only at a single point in time and is then represented in a field by a self-sustained peak. If the target object does not follow a straight trajectory, the represented motion direction is not correct over time.

[2]The original video resolution of 640x480 pixels is cropped and resized to 55x50 pixels to reduce computational load.

A practical issue is the computational load of the model. During simulation, the model currently needs to be slowed down because the CPU load is too high for an average computer. In the short term, this could be alleviated by running it on better hardware or distributing the model onto multiple networked computers. In the long term, it could be addressed by implementing operations that are computationally costly on GPUs or dedicated neuromorphic hardware.

5.4 Further research

The model introduced in this thesis represents a first step toward a comprehensive and neurally plausible model of relational processing. A next step could be to scale the model to include more feature dimensions, more concepts, and more complex processing steps. This

should be possible without additional conceptual work, as care was taken that scaling is possible. For instance, even though a copy of a concept must be provided for each role in which it may appear, the representation of relations only requires very few roles, perhaps only one or two additional ones to express tertiary relations such as BETWEEN. Expressing more complex relations may be accomplished by some form of compositionality that would reuse the few available roles.

The model is also open to more pervasive extensions in various directions. A first direction could be to incorporate learning processes. Currently, the entire model is designed by hand, but some aspects lend themselves to learning, in particular the patterned synaptic connections that encode concepts as well as the sequential constraints and CoS of the process organization system (Luciw et al., 2015). Incorporating learning processes is a challenge because these processes would have to be governed by the same principles as the rest of the model. That means that they have to be continuously updated and that the learning controlled by the model itself—an autonomous learning.

Luciw, M., Kazerounian, S., Lahkman, K., Richter, M., & Sandamirskaya, Y. (2015). Learning the condition of satisfaction of an elementary behavior in dynamic field theory. *Paladyn, Journal of Behavioral Robotics*, *6*(1), 180–190

A second direction to extend the model is to systematically explore how other relations may be realized in neural dynamics. The spatial relations implemented in the current model may be viewed as instantiations of the image schemas LEFT-RIGHT and UP-DOWN, while the movement relations may correspond to the PATH schema (M. Johnson, 1987). Exploring other schemas will likely uncover that the model must be extended to incorporate additional transformations or other mechanisms. For example, the schema CONTAINMENT cannot simply be implemented by introducing a new spatial template and comparing it to the relative position of two objects. Instead, even a minimal version requires that the absolute position of the target object is compared to the absolute position of the reference object to detect overlap. A more comprehensive model would also capture that the meaning of CONTAINMENT depends on the shape and common use of the container. Compare, for instance, the difference in the perceptual meaning of "The apple is *inside* the bowl." and "The person is *inside* the house."

Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago: University of Chicago Press

A third extension of the model is to capture how mental models may be established from descriptions. This task may thus be viewed as the inverse process to describing a scene. For instance, given a description such as "There is a red object to the left of a green object.", the task is to build up a perceptual representation of that scene in working memory. After populating the representation with multiple objects, it can be utilized to make inferences on them. A central issue is where objects are to be placed in space, in particular when there are multiple possible positions because the given description

is ambiguous. Ragni and Knauff (2013) show that people establish a preferred mental model that arises by adding objects to an imagined scene such that only minimal change is required. Their model covers problems of where to place objects but is based on abstract symbolic representations and an algorithmic implementation. Nevertheless, their ideas could be transferred to the model introduced here to build a grounded neural process model of mental imagery. Building further on this idea, the model could be extended to express more abstract relations, such as "Gerhard Schröder was more popular than Angela Merkel is." by mapping abstract dimensions such as popularity onto space (Ragni & Knauff, 2013).

Ragni, M. & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588

A fourth, ambitious extension would be to address productivity and compositionality (Fodor & Pylyshyn, 1988). Currently, the model is only able to ground or describe single (simple) phrases. However, language is filled with complex sentences that include multiple internal references, relative clauses, and conjunctions. Capturing these types of sentences requires, at the very least, a (much) more complex organization of processes, and most likely also demanding conceptual work. But only by facing these challenges can we hope to establish a comprehensive account of the grounding of cognition and language.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71

Independently from extending the model, a line of research could be established with the goal of testing assumptions and possible predictions of the model with behavioral experiments. This is a challenge because the model is both complex and abstract and many of its assumptions do not have obvious behavioral signatures. Nevertheless, in a first step toward such a line of research, Lins and Schöner (2017) employed mouse tracking to investigate the attentional processes during the grounding of spatial relations. They find that the mouse trajectory is attracted toward distractor objects, toward the reference object, as well as that it is biased by the spatial term itself. A comprehensive analysis of the findings and additional experimental setups may lead to conclusions about the underlying attentional processes.

Lins, J. & Schöner, G. (2017). Mouse tracking shows attraction to alternative targets while grounding spatial relations. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 2586–2591). Austin, TX: Cognitive Science Society

Finally, mapping the conceptual ideas that are at the basis of the proposed model to literature from other related fields, for example to verbal theories of language and cognition (e.g., Langacker, 1986; Barsalou, 1999) may lead to a much needed dialog and synthesis between these distant fields.

Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive Science*, *10*, 1–40; Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609, 577–609

Conclusion 6

This thesis examined the perceptual grounding of spatial relations and movement relations. It proposes a model based on dynamic field theory (DFT) that captures the neural processes of how such relations may be extracted from visual scenes and how they may connect to conceptual representations close to language. The capabilities of the model were demonstrated in an extensive set of computer simulations that probed it in qualitatively different tasks. The visual scenes used for these tests were recorded with a real camera and real objects, simplifying object segmentation by using a white background; they were created specifically to test the proposed model. In all 104 tests, the model successfully either grounded the given phrase or described the relation of two objects in the given scene.

This thesis is making several contributions, both on a conceptual level and in terms of novel neural dynamic implementations. Most notably, it establishes neural dynamic principles by which DFT models may flexibly organize their own processes and behaviors. A notable conceptual contribution consists of refining the core component processes required for grounding spatial relations. Overall, the thesis shows how the perceptual grounding of spatial and movement relations may be captured based on neural principles.

The most direct impact of this work may stem from the principles of process organization. They are formulated in a generic form and can be applied to organize the processes of any DFT model. The concepts and principles of behavioral organization on which the process organization system is built are already pervasively used in recent DFT models and have become a useful tool to think about neural dynamic models. The more flexible and generic process organization system proposed here will become increasingly important as models become more integrated, grow larger and more complex.

This development will hopefully demonstrate that process organization is a problem that can and should be addressed by neural approaches, something currently often overlooked.

The proposed model itself may have an impact on a longer time scale, as the foundation of an ambitious research project toward understanding higher cognition. It has many entry points to integrate with related DFT models, for instance those of object recognition, scene representation, and motor control. It is open to extension into related research areas like mental imagery, image schemas, abstract relations, or metaphors. With the novel structure of discrete conceptual representations, it has created an interface to language that may invite work in cognitive linguistics, possibly tackling the challenging characteristic of compositionality. In short, the model could become one of the central pieces in an integrated, neural process model of higher cognition. While this thesis is only laying some of the ground work, the prospects and future potential of this ambitious project are truly exciting.

Appendices

Appendices

A Implementation details

A.1 Model parameters

The model introduced in Section 3 is specified using a using a number of parameters. This section gives a rough overview of the parameter values used. However, please keep in mind that this parameter set may be specific to the camera input used and to the specific implementations of the dynamics in *cedar*.

Some parameters are the same for all dynamics in the model. This includes the strength $w_\xi = 0.1$ of the noise term as well as the steepness parameter $\beta = 100$ of the sigmoidal output function. Other parameters also have values that are used for many dynamics, but in some cases other parameter values had to be used.

Time scale

The default time scale is $\tau = 50$, which is used in most of the dynamics of the model. The following list contains all time scale parameters that deviate from this default.

$$\begin{aligned}
\tau_{\text{v}} &= 200 \\
\tau_{\text{AS}} &= 25 \\
\tau_{\text{ASm}} &= 100 \\
\tau_{\text{RC}} &= 100 \\
\tau_{\text{Scs}} &= 100 \\
\tau_{\text{Scd}} &= 500 \\
\tau_{\text{ROTs}} &= 20
\end{aligned}$$

$$\tau_{\text{ROTd}} = 100$$
$$\tau_{\text{ROT}} = 20$$
$$\tau_{\text{TCP}} = 20$$
$$\tau_{\text{TMP}} = 20$$
$$\tau_{\text{RCP}} = 20$$
$$\tau_{\text{SP}} = 20$$

Resting levels

The default resting level used in most nodes and fields is $h = -5$. A few fields use different resting level parameters.

$$h_{\text{AS}} = -10$$
$$h_{\text{ASm}} = -10$$
$$h_{\text{T}} = -7$$
$$h_{\text{IR}} = -8$$
$$h_{\text{Scs}} = 7.5$$
$$h_{\text{Scd}} = 3.75$$
$$h_{\text{SM}} = -10$$

Kernels and connection weights

In the equations in Section 3 and Section B, I simplified the connections between different dynamic elements. When two fields (or nodes) are coupled such that field A gives input to field B, then the output of field A is usually convolved with a kernel. In the implementation, the result is multiplied with a scalar weight. In writing down the equations, I dropped the scalar weights and only wrote down the kernels. For more detailed information on the connection weights and kernels, please consult the *cedar* configuration file of the model, which is humanly readable and available online at https://www.ini.rub.de/pages/publications/richterphdthesis.

Processes

All processes in the process organization system have the same parameters. This is a list of parameters that deviate from the default parameters listed above for the four nodes that make up a process. The self-excitation of the prior intention node is $w_{\text{P,P}} = 1$. The parameters of the intention node are

$$h_{\text{I}} = -1,$$
$$w_{\text{I,I}} = 4,$$
$$w_{\text{I,P}} = 1,$$

$$w_{\mathrm{I,M}} = 6.$$

The CoS node has the following parameters

$$\begin{aligned} h_{\mathrm{C}} &= -20, \\ w_{\mathrm{C,C}} &= 1, \\ w_{\mathrm{C,I}} &= 20. \end{aligned}$$

The parameters of the CoS memory node are

$$\begin{aligned} h_{\mathrm{M}} &= -10, \\ w_{\mathrm{M,M}} &= 6, \\ w_{\mathrm{M,C}} &= 6. \end{aligned}$$

The weight from the intention node of a process on a higher hierarchical level to the prior intention node and CoS memory node on a lower hierarchical level is $w_{\mathrm{AI}} = 6$.

A.2 Visual preprocessing

The visual preprocessing is performed on videos from a camera that delivers BGR images (blue, green, red) of 640x480 pixels with 8 bits per pixel, which allows for 256 different color values. Each incoming image is cropped to remove surroundings that are not to be fed to the model. After removing 124 pixels on the left, 114 pixels on the right, 45 pixels on the bottom, and 73 pixels on top, the remaining image is of size 402x362 pixels. Each cropped image is then scaled down to size 55x50 pixels, the size at which the spatial dimensions x and y of the image are sampled in the fields. The scaled image is converted to the hue, saturation, value (HSV) color space.

This is done by an algorithm that sets the color saturation value of each pixel into a three-dimensional matrix, where the first two indices are given by the pixel coordinates and the third index is the hue value at the pixel, scaled to a range between 0 and 49.

A.3 Software

The model introduced in this thesis (Section 3) was implemented, parameterized, and all of the tests were conducted using the software framework *cedar*. *cedar* is an open-source C++ library that is freely available under the license LGPL version 3. Its source code and documentation can be accessed at http://cedar.ini.rub.de.

The model was built using *cedar* version 5.0.1 (debug build) with the third-party libraries boost (1.54), Qt (4.86), OpenCV (2.4.11),

and FFTW (3.3.3). Most of the processing steps are available in the core of cedar, except for two processing steps (SpatialPattern and ShiftedAddition) that are part of a plugin.[1] The plugin is available online at http://bitbucket.org/cedar/plugins.

[1] In more recent versions of *cedar*, the processing step "SpatialPattern" is available within *cedar* under the name "SpatialTemplate".

Loading the *cedar* architecture that implements the model moreover requires configuration files that are available online at https://www.ini.rub.de/pages/publications/richterphdthesis. One of these configuration files holds all parameter values for the model.

In order to visualize both the state of activation as well as how it evolved in time during the tests, the activation of neural fields and nodes was written to files during experiments. These files were processed and visualized using Matlab scripts, which are available online at https://www.ini.rub.de/pages/publications/richterphdthesis.

B Process organization system: equations

All individual processes are governed by the generic equations listed in Section 3.5. However, since their interconnection is only shown graphically in the main text (Figure 3.14 on page 75), the equations of all processes are listed here in detail.

B.1 Ground object process

The differential equations governing the nodes of the process are as follows. The activation u_{GOP} of the prior intention node evolves in time based on

$$\begin{aligned}\tau \dot{u}_{\mathrm{GOP}}(t) = &- u_{\mathrm{GOP}}(t) + h + w_\xi \cdot \xi_{\mathrm{GOP}}(t) \\ &+ w_{\mathrm{P,P}}\, g(u_{\mathrm{GOP}}(t)) \\ &+ s_{\mathrm{GOP,U}}(t),\end{aligned} \tag{1}$$

where $s_{\mathrm{GOP,U}}(t)$ is user input that activates the prior intention node.

The activation u_{GOI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\mathrm{GOI}}(t) = &- u_{\mathrm{GOI}}(t) + h_{\mathrm{I}} + w_\xi \cdot \xi_{\mathrm{GOI}}(t) \\ &+ w_{\mathrm{I,I}}\, g(u_{\mathrm{GOI}}(t)) \\ &+ w_{\mathrm{I,P}}\, g(u_{\mathrm{GOP}}(t)) \\ &- w_{\mathrm{I,M}}\, g(u_{\mathrm{GOM}}(t)),\end{aligned} \tag{2}$$

where the third line formalizes the excitatory input from the prior intention node and the fourth line is the inhibitory input from the CoS node with the activation variable u_{GOC}.

The activation u_{GOC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GOC}}(t) = &- u_{\text{GOC}}(t) + h_{\text{C}} + w_{\xi} \cdot \xi_{\text{GOC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{GOC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{GOI}}(t)) \\ &- g(u_{\text{TP}}(t)) + g(u_{\text{TM}}(t)),\end{aligned} \tag{3}$$

where the third line is input from the intention node. The last line is inhibitory input from the prior intention node of the target process as well as excitatory input from the CoS memory node of the target process. This makes the condition of satisfaction (CoS) of the ground object process dependent on a previously established CoS of the target process on the lower level.

The activation u_{GOM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GOM}}(t) = &- u_{\text{GOM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{GOM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{GOM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{GOC}}(t)) \\ &+ s_{\text{GOM,U}}(t),\end{aligned} \tag{4}$$

where the third line is input from the CoS node and the fourth line $s_{\text{GOM,U}}(t) = s_{\text{GOP,U}}(t)$ is the same user input that activates the prior intention node.

B.2 Ground relation process

The activation u_{GRP} of the prior intention node evolves in time based on the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GRP}}(t) = &- u_{\text{GRP}}(t) + h + w_{\xi} \cdot \xi_{\text{GRP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{GRP}}(t)) \\ &+ s_{\text{GRP,U}}(t),\end{aligned} \tag{5}$$

where $s_{\text{GRP,U}}(t)$ is user input that activates the prior intention node.

The activation u_{GRI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{GRI}}(t) = &- u_{\text{GRI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{GRI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{GRP}}(t)) \\ &- w_{\text{I,M}}\, g(u_{\text{GRM}}(t)),\end{aligned} \tag{6}$$

which is structured analogous to Equation 3.54.

The activation u_{GRC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GRC}}(t) = &- u_{\text{GRC}}(t) + h_{\text{C}} + w_{\xi} \cdot \xi_{\text{GRC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{GRC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{GRI}}(t)) \\ &- g(u_{\text{TP}}(t)) + g(u_{\text{TM}}(t)) \\ &- g(u_{\text{RP}}(t)) + g(u_{\text{RM}}(t)) \\ &- g(u_{\text{SP}}(t)) + g(u_{\text{SM}}(t)),\end{aligned} \tag{7}$$

where the last three lines are input from the target process, the reference process, and the spatial relation process, which give inhibitory input from their prior intention node and excitatory input from their CoS memory node. This makes the CoS of the ground relation process dependent on the CoS of these three processes. The CoS node does not receive input from the clean process or the reset process because they do not necessarily have to be activated to successfully complete the task.

The activation u_{GRM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GRM}}(t) = &- u_{\text{GRM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{GRM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{GRM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{GRC}}(t)) \\ &+ s_{\text{GRM,U}}(t),\end{aligned} \tag{8}$$

where the third line is input from the CoS node and the fourth line $s_{\text{GRM,U}}(t) = s_{\text{GRP,U}}(t)$ is the same user input that activates the prior intention node.

B.3 Describe process

The neural nodes implementing this process evolve in time based on differential equations analogous to the processes explained above. The activation u_{DP} of the prior intention node follows the equation

$$\begin{aligned}\tau \dot{u}_{\text{DP}}(t) = &- u_{\text{DP}}(t) + h + w_{\xi} \cdot \xi_{\text{DP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{DP}}(t)) \\ &+ s_{\text{DP,U}}(t),\end{aligned} \tag{9}$$

where $s_{\text{DP,U}}(t)$ is user input that activates the prior intention node.

The activation u_{DI} of the intention node evolves in time based on the following differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{DI}}(t) = &- u_{\mathrm{DI}}(t) + h_{\mathrm{I}} + w_{\xi} \cdot \xi_{\mathrm{DI}}(t) \\ &+ w_{\mathrm{I,I}}\, g(u_{\mathrm{DI}}(t)) \\ &+ w_{\mathrm{I,P}}\, g(u_{\mathrm{DP}}(t)) \\ &- w_{\mathrm{I,M}}\, g(u_{\mathrm{DM}}(t)).\end{aligned} \tag{10}$$

The activation u_{DC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{DC}}(t) = &- u_{\mathrm{DC}}(t) + h_{\mathrm{C}} + w_{\xi} \cdot \xi_{\mathrm{DC}}(t) \\ &+ w_{\mathrm{C,C}}\, g(u_{\mathrm{DC}}(t)) \\ &+ w_{\mathrm{C,I}}\, g(u_{\mathrm{DI}}(t)) \\ &- g(u_{\mathrm{TP}}(t)) + g(u_{\mathrm{TM}}(t)) \\ &- g(u_{\mathrm{RP}}(t)) + g(u_{\mathrm{RM}}(t)) \\ &- g(u_{\mathrm{SP}}(t)) + g(u_{\mathrm{SM}}(t)),\end{aligned} \tag{11}$$

which is structured analogous to Equation 3.55.

The activation u_{DM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{DM}}(t) = &- u_{\mathrm{DM}}(t) + h_{\mathrm{M}} + w_{\xi} \cdot \xi_{\mathrm{DM}}(t) \\ &+ w_{\mathrm{M,M}}\, g(u_{\mathrm{DM}}(t)) \\ &+ w_{\mathrm{M,C}}\, g(u_{\mathrm{DC}}(t)) \\ &+ s_{\mathrm{DM,U}}(t),\end{aligned} \tag{12}$$

where the third line is input from the CoS node and the fourth line $s_{\mathrm{DM,U}}(t) = s_{\mathrm{DP,U}}(t)$ is the same user input that activates the prior intention node.

B.4 Target process

The neural nodes of the target process are structured analogously to those processes explained above. The activation u_{TP} of the prior intention node is governed by the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{TP}}(t) = &- u_{\mathrm{TP}}(t) + h + w_{\xi} \cdot \xi_{\mathrm{TP}}(t) \\ &+ w_{\mathrm{P,P}}\, g(u_{\mathrm{TP}}(t)) \\ &+ w_{\mathrm{TP,GOI}}\, g(u_{\mathrm{GOI}}(t)) \\ &+ w_{\mathrm{TP,GRI}}\, g(u_{\mathrm{GRI}}(t)) \\ &+ w_{\mathrm{TP,DI}}\, g(u_{\mathrm{DI}}(t)),\end{aligned} \tag{13}$$

where the last three lines formalize the excitatory input from the intention nodes of all processes on the next higher hierarchical level,

the ground object process, the ground relation process, and the describe process.

The activation u_{TI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\mathrm{TI}}(t) = &- u_{\mathrm{TI}}(t) + h_{\mathrm{I}} + w_{\xi} \cdot \xi_{\mathrm{TI}}(t) \\ &+ w_{\mathrm{I,I}}\, g(u_{\mathrm{TI}}(t)) \\ &+ w_{\mathrm{I,P}}\, g(u_{\mathrm{TP}}(t)) \\ &- w_{\mathrm{I,M}}\, g(u_{\mathrm{TM}}(t)) \\ &- w_{\mathrm{TI,SR}}\, g(u_{\mathrm{SR}}(t)),\end{aligned} \tag{14}$$

where the first four lines are analogous to Equation 3.54. The last line formalizes inhibitory input from a suppression node. The node is activated by the reset process in order to restart the grounding process of the target object and reference object. Please refer to Section B.22 for more information on the reset process and to Equation 90 for the differential equation that governs the suppression node.

The activation u_{TC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{TC}}(t) = &- u_{\mathrm{TC}}(t) + h_{\mathrm{C}} + w_{\xi} \cdot \xi_{\mathrm{TC}}(t) \\ &+ w_{\mathrm{C,C}}\, g(u_{\mathrm{TC}}(t)) \\ &+ w_{\mathrm{C,I}}\, g(u_{\mathrm{TI}}(t)) \\ &- g(u_{\mathrm{GPP}}(t)) + g(u_{\mathrm{GPM}}(t)) \\ &- g(u_{\mathrm{GAP}}(t)) + g(u_{\mathrm{GAM}}(t)) \\ &- g(u_{\mathrm{GFP}}(t)) + g(u_{\mathrm{GFM}}(t)) \\ &- g(u_{\mathrm{TIP}}(t)) + g(u_{\mathrm{TIM}}(t)) \\ &- g(u_{\mathrm{TNP}}(t)) + g(u_{\mathrm{TNM}}(t)) \\ &- g(u_{\mathrm{TMP}}(t)) + g(u_{\mathrm{TMM}}(t)) \\ &- g(u_{\mathrm{TTP}}(t)) + g(u_{\mathrm{TTM}}(t)),\end{aligned} \tag{15}$$

which is structured analogously to Equation 7. Lines 4–10 formalize inhibitory input from the prior intention nodes and excitatory input from the CoS nodes of all processes on the next lower hierarchical level that the target process is associated with (dark blue lines in Figure 3.14): from top to bottom, the inputs come from the perceptual boost process (GP), the spatial attention process (GA), the feature process (GF), the target IOR process (TI), the target memory node process (TN), the target motion field process (TM), and the target field process (TT).

The activation u_{TM} of the CoS memory node follows the differ-

ential equation

$$\begin{aligned}\tau \dot{u}_{\text{TM}}(t) = &- u_{\text{TM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{TM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{TM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{TC}}(t)) \\ &+ w_{\text{TM,GOI}}\, g(u_{\text{GOI}}(t)) \\ &+ w_{\text{TM,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{TM,DI}}\, g(u_{\text{DI}}(t)) \\ &- w_{\text{TM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{16}$$

where the third line is input from the CoS node, lines 4–6 are the same inputs from the intention nodes of processes on the next higher hierarchical level that activate the prior intention node. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.5 Reference process

The dynamic neural nodes that represent the reference process are governed by the following equations. The activation u_{RP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{RP}}(t) = &- u_{\text{RP}}(t) + h + w_{\xi} \cdot \xi_{\text{RP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{RP}}(t)) \\ &+ w_{\text{RP,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{RP,DI}}\, g(u_{\text{DI}}(t)),\end{aligned} \tag{17}$$

where the last two lines formalize the excitatory input from the intention nodes of the ground relation process and the describe process, both of which are on the next higher hierarchical level.

The activation u_{RI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{RI}}(t) = &- u_{\text{RI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{RI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{RI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{RP}}(t)) \\ &- w_{\text{I,M}}\, g(u_{\text{RM}}(t)) \\ &- w_{\text{RI,PRC}}\, g(u_{\text{PRC}}(t)) \\ &- w_{\text{RI,SR}}\, g(u_{\text{SR}}(t))\end{aligned} \tag{18}$$

where the first four lines are analogous to Equation 3.54. The fifth line is inhibitory input from a precondition node that ensures the reference process is only activated once the clean process is finished.[2] The last line formalizes inhibitory input from the suppression node activated by the reset process.

[2] Please refer to Section B.22, in particular Equation 87.

The activation u_{RC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{RC}}(t) = &- u_{\text{RC}}(t) + h_{\text{C}} + w_{\xi} \cdot \xi_{\text{RC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{RC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{RI}}(t)) \\ &- g(u_{\text{GPP}}(t)) + g(u_{\text{GPM}}(t)) \\ &- g(u_{\text{GAP}}(t)) + g(u_{\text{GAM}}(t)) \\ &- g(u_{\text{GFP}}(t)) + g(u_{\text{GFM}}(t)) \\ &- g(u_{\text{RNP}}(t)) + g(u_{\text{RNM}}(t)) \\ &- g(u_{\text{RFP}}(t)) + g(u_{\text{RFM}}(t)),\end{aligned} \tag{19}$$

which is structured analogously to Equation 7. Lines 4–8 formalize inhibitory input from the prior intention nodes and excitatory input from the CoS nodes of all processes on the next lower hierarchical level that the reference process is associated with: from top to bottom, the inputs come from the perceptual boost process (GP), the spatial attention process (GA), feature process (GF), the reference memory node process (RN), and the reference field process (RF).

The activation u_{RM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{RM}}(t) = &- u_{\text{RM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{RM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{RM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{RC}}(t)) \\ &+ w_{\text{RM,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{RM,DI}}\, g(u_{\text{DI}}(t)) \\ &- w_{\text{RM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{20}$$

where the third line is input from the CoS node, lines four and five are the same inputs from the intention nodes of processes on the next higher hierarchical level that activate the prior intention node. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.6 Spatial relation process

The spatial relation process activates multiple processes on the next lower hierarchical level (dark red lines in Figure 3.14): the spatial memory node process (abbreviated SN in Figure 3.14) and the spatial relational field process (SR). All of these processes control an aspect of the grounding of the spatial relation and will be explained later.

The dynamic neural nodes that represent the spatial relation process are governed by the following equations. The activation u_{SP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SP}}(t) = &- u_{\text{SP}}(t) + h + w_{\xi} \cdot \xi_{\text{SP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{SP}}(t)) \\ &+ w_{\text{SP,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{SP,DI}}\, g(u_{\text{DI}}(t)),\end{aligned} \tag{21}$$

where the last two lines formalize the excitatory input from the intention nodes of the ground relation process and the describe process, both of which are on the next higher hierarchical level.

The activation u_{SI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{SI}}(t) = &- u_{\text{SI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{SI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{SI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{SP}}(t)) \\ &- w_{\text{I,M}}\, g(u_{\text{SM}}(t)) \\ &- w_{\text{SI,SR}}\, g(u_{\text{SR}}(t))\end{aligned} \tag{22}$$

where the first four lines are analogous to Equation 3.54. The last line formalizes inhibitory input from the suppression node activated by the reset process.

The activation u_{SC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SC}}(t) = &- u_{\text{SC}}(t) + h_{\text{C}} + w_{\xi} \cdot \xi_{\text{SC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{SC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{SI}}(t)) \\ &- g(u_{\text{SNP}}(t)) + g(u_{\text{SNM}}(t)) \\ &- g(u_{\text{SRP}}(t)) + g(u_{\text{SRM}}(t)),\end{aligned} \tag{23}$$

where lines four and five formalize inhibitory input from the prior intention nodes and excitatory input from the CoS nodes of all processes on the next lower hierarchical level that the spatial relation process is associated with: the inputs come from the spatial memory node process (SN) and the spatial relational field process (SR).

The activation u_{SM} of the CoS memory node evolves in time based on the following differential equation, where the third line is input from the CoS node, lines four and five are the same inputs from the intention nodes of processes on the next higher hierarchi-

cal level that activate the prior intention node

$$\begin{aligned}\tau \dot{u}_{\mathrm{SM}}(t) = &- u_{\mathrm{SM}}(t) + h_{\mathrm{M}} + w_{\xi} \cdot \xi_{\mathrm{SM}}(t) \\ &+ w_{\mathrm{M,M}}\, g(u_{\mathrm{SM}}(t)) \\ &+ w_{\mathrm{M,C}}\, g(u_{\mathrm{SC}}(t)) \\ &+ w_{\mathrm{SM,GRI}}\, g(u_{\mathrm{GRI}}(t)) \\ &+ w_{\mathrm{SM,DI}}\, g(u_{\mathrm{DI}}(t)) \\ &- w_{\mathrm{SM,SR}}\, g(u_{\mathrm{SR}}(t)).\end{aligned} \tag{24}$$

The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.7 Clean process

The dynamic neural nodes that represent the clean process are governed by the following equations. The activation u_{CP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{CP}}(t) = &- u_{\mathrm{CP}}(t) + h + w_{\xi} \cdot \xi_{\mathrm{CP}}(t) \\ &+ w_{\mathrm{P,P}}\, g(u_{\mathrm{CP}}(t)) \\ &+ w_{\mathrm{CP,GRI}}\, g(u_{\mathrm{GRI}}(t)) \\ &+ w_{\mathrm{CP,DI}}\, g(u_{\mathrm{DI}}(t)),\end{aligned} \tag{25}$$

where the last two lines formalize the excitatory input from the intention nodes of the ground relation process and the describe process, both of which are on the next higher hierarchical level.

The activation u_{CI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\mathrm{CI}}(t) = &- u_{\mathrm{CI}}(t) + h_{\mathrm{I}} + w_{\xi} \cdot \xi_{\mathrm{CI}}(t) \\ &+ w_{\mathrm{I,I}}\, g(u_{\mathrm{CI}}(t)) \\ &+ w_{\mathrm{I,P}}\, g(u_{\mathrm{CP}}(t)) \\ &- w_{\mathrm{I,M}}\, g(u_{\mathrm{CM}}(t)) \\ &- w_{\mathrm{CI,PCT}}\, g(u_{\mathrm{PCT}}(t)) \\ &- w_{\mathrm{CM,SR}}\, g(u_{\mathrm{SR}}(t)),\end{aligned} \tag{26}$$

[3]Please refer to Section B.22, in particular Equation 86.

where the first four lines are analogous to Equation 3.54. The fifth line is inhibitory input from a precondition node that ensures the clean process is only activated once the target process is finished.[3] The last line formalizes inhibitory input from the suppression node activated by the reset process.

The activation u_{CC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{CC}}(t) = &- u_{\mathrm{CC}}(t) + h_{\mathrm{C}} + w_{\xi} \cdot \xi_{\mathrm{CC}}(t) \\ &+ w_{\mathrm{C,C}}\, g(u_{\mathrm{CC}}(t)) \\ &+ w_{\mathrm{C,I}}\, g(u_{\mathrm{CI}}(t)) \\ &- w_{\mathrm{CC,ACcs}}\, \max_{c}(g(u_{\mathrm{ACcs}}(c,t))) \\ &- w_{\mathrm{CC,AMcs}}\, \max_{c}(g(u_{\mathrm{AMcs}}(c,t))),\end{aligned} \tag{27}$$

where lines four and five formalize inhibitory input from the color CoS field and the motion CoS field, respectively. Their output is contracted to a scalar value and multiplied with a weight. This inhibitory input ensures that the CoS node of the clean process is only activated once there is no peak in either of these fields.

The activation u_{CM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{CM}}(t) = &- u_{\mathrm{CM}}(t) + h_{\mathrm{M}} + w_{\xi} \cdot \xi_{\mathrm{CM}}(t) \\ &+ w_{\mathrm{M,M}}\, g(u_{\mathrm{CM}}(t)) \\ &+ w_{\mathrm{M,C}}\, g(u_{\mathrm{CC}}(t)) \\ &+ w_{\mathrm{CM,GRI}}\, g(u_{\mathrm{GRI}}(t)) \\ &+ w_{\mathrm{CM,DI}}\, g(u_{\mathrm{DI}}(t)) \\ &- w_{\mathrm{CM,SR}}\, g(u_{\mathrm{SR}}(t)),\end{aligned} \tag{28}$$

where the third line is input from the CoS node, lines four and five are the same inputs from the intention nodes of processes on the next higher hierarchical level that activate the prior intention node. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.8 Reset process

The dynamic neural nodes that represent the reset process are governed by the following equations. The activation u_{EP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{EP}}(t) = &- u_{\mathrm{EP}}(t) + h + w_{\xi} \cdot \xi_{\mathrm{EP}}(t) \\ &+ w_{\mathrm{P,P}}\, g(u_{\mathrm{EP}}(t)) \\ &+ w_{\mathrm{EP,GRI}}\, g(u_{\mathrm{GRI}}(t)) \\ &+ w_{\mathrm{EP,DI}}\, g(u_{\mathrm{DI}}(t)),\end{aligned} \tag{29}$$

where the last two lines formalize the excitatory input from the intention nodes of the ground relation process and the describe process, both of which are on the next higher hierarchical level.

The activation u_{EI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{EI}}(t) = &- u_{\text{EI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{EI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{EI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{EP}}(t)) \\ &- w_{\text{I,M}}\, g(u_{\text{EM}}(t)) \\ &- w_{\text{EI,PRD}}\, g(u_{\text{PRD}}(t)),\end{aligned} \tag{30}$$

where the first four lines are analogous to Equation 3.54. The last line formalizes inhibitory input from a precondition node that ensures the reset process is only activated if there is a peak in the spatial relation CoD field.[4]

[4]Please refer to Section B.22, in particular Equation 91.

The activation u_{EC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{EC}}(t) = &- u_{\text{EC}}(t) + h_{\text{C}} + w_{\xi} \cdot \xi_{\text{EC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{EC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{EI}}(t)),\end{aligned} \tag{31}$$

which means it does not get any input except from the intention node. The reset process does not necessarily need a CoS, because it only activates the suppression node. The suppression node inhibits large parts of the model, including the spatial relation CoS field, which leads to the activation of the reset process in the first place. Thus, if the spatial relation CoS field is inhibited, the reset process is deactivated.

The activation u_{EM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{EM}}(t) = &- u_{\text{EM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{EM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{EM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{EC}}(t)) \\ &+ w_{\text{EM,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{EM,DI}}\, g(u_{\text{DI}}(t)),\end{aligned} \tag{32}$$

where the third line is input from the CoS node, lines four and five are the same inputs from the intention nodes of processes on the next higher hierarchical level that activate the prior intention node. Since the CoS node is never going to be activated, the CoS memory node is not necessarily required either.

B.9 Perceptual boost process

The dynamic neural nodes that represent the perceptual boost process are governed by the following equations. The activation u_{GPP}

of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GPP}}(t) = &- u_{\text{GPP}}(t) + h + w_{\xi} \cdot \xi_{\text{GPP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{GPP}}(t)) \\ &+ w_{\text{GPP,TI}}\, g(u_{\text{TI}}(t)) \\ &+ w_{\text{GPP,RI}}\, g(u_{\text{RI}}(t)) \\ &+ w_{\text{GPP,DI}}\, g(u_{\text{DI}}(t)),\end{aligned} \tag{33}$$

where the last three lines formalize the excitatory input from the intention nodes of the target process, the reference process, and the describe process. The weights of these connections are set such that the prior intention node can only be activated when the target process or the reference process are active at the same time as the describe process.

The activation u_{GPI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{GPI}}(t) = &- u_{\text{GPI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{GPI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{GPI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{GPP}}(t)),\end{aligned} \tag{34}$$

which is analogous to Equation 3.54. However, please note that the intention node of this process is not inhibited by the CoS memory node. This is because the boost of the spatial attention fields must remain active until the higher level process (target process or reference process) is completed.

The activation u_{GPC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GPC}}(t) = &- u_{\text{GPC}}(t) + h_{\text{GPC}} + w_{\xi} \cdot \xi_{\text{GPC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{GPC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{GPI}}(t)) \\ &+ w_{\text{GPC,ASm}} \max_{x,y}(g(u_{\text{ASm}}(x, y, t))),\end{aligned} \tag{35}$$

where the last line is excitatory input from the multi-peak spatial attention field. Please note that since this CoS node receives input from a field of the model, rather than from other processes, its resting level is lowered from the default value used in the processes described previously.

The activation u_{GPM} of the CoS memory node follows the dif-

ferential equation

$$\begin{aligned}\tau \dot{u}_{\text{GPM}}(t) = &- u_{\text{GPM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{GPM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{GPM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{GPC}}(t)) \\ &+ w_{\text{GPM,TI}}\, g(u_{\text{TI}}(t)) \\ &+ w_{\text{GPM,RI}}\, g(u_{\text{RI}}(t)) \\ &- w_{\text{GPM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{36}$$

where the third line is input from the CoS node, lines four and five are the same inputs from the intention nodes of processes on the next higher hierarchical level that activate the prior intention node. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.10 Spatial attention process

The dynamic neural nodes that represent the spatial attention process are governed by the following equations. The activation u_{GAP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GAP}}(t) = &- u_{\text{GAP}}(t) + h + w_{\xi} \cdot \xi_{\text{GAP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{GAP}}(t)) \\ &+ w_{\text{GAP,TI}}\, g(u_{\text{TI}}(t)) \\ &+ w_{\text{GAP,RI}}\, g(u_{\text{RI}}(t)) \\ &+ w_{\text{GAP,GRI}}\, g(u_{\text{GRI}}(t)),\end{aligned} \tag{37}$$

where the last three lines formalize the excitatory input from the intention nodes of the target process, the reference process, and the ground relation process. The weights of these connections are set such that the prior intention node can only be activated when the target process or the reference process are active at the same time as the ground relation process.

The activation u_{GAI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{GAI}}(t) = &- u_{\text{GAI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{GAI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{GAI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{GAP}}(t)) \\ &- w_{\text{GAI,PAR}}\, g(u_{\text{PAR}}(t)),\end{aligned} \tag{38}$$

which is analogous to Equation 3.54. However, please note that the intention node of this process is not inhibited by the CoS memory node. This is because the boost of the selective spatial attention field must remain active until the higher level process is completed.

The last line is inhibitory input from a precondition node that ensures the spatial attention process is only activated once the spatial relational field process is finished.[5] This precondition node is only activated by the reference process, not by the target process.

[5] Please refer to Section B.22, in particular Equation 89.

The activation u_{GAC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GAC}}(t) = &- u_{\text{GAC}}(t) + h_{\text{GAC}} + w_{\xi} \cdot \xi_{\text{GAC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{GAC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{GAI}}(t)) \\ &+ w_{\text{GAC,AS}} \max_{x,y}(g(u_{\text{AS}}(x,y,t))),\end{aligned} \tag{39}$$

where the last line is excitatory input from the selective spatial attention field. Please note that since this CoS node receives input from a field of the model, rather than from other processes, its resting level is lowered from the default value used in the processes described previously.

The activation u_{GAM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GAM}}(t) = &- u_{\text{GAM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{GAM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{GAM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{GAC}}(t)) \\ &+ w_{\text{GAM,TI}}\, g(u_{\text{TI}}(t)) \\ &+ w_{\text{GAM,RI}}\, g(u_{\text{RI}}(t)) \\ &- w_{\text{GAM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{40}$$

where lines four and five are inputs from the intention nodes of the target process and the reference process. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.11 Feature process

The dynamic neural nodes that represent the feature process are governed by the following equations. The activation u_{GFP} of the prior intention node is governed by the following equation, where the last four lines formalize the excitatory input from the intention nodes of the target process, the reference process, the ground object process,

and the ground relation process

$$\begin{aligned}\tau \dot{u}_{\text{GFP}}(t) = &- u_{\text{GFP}}(t) + h + w_{\xi} \cdot \xi_{\text{GFP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{GFP}}(t)) \\ &+ w_{\text{GFP,TI}}\, g(u_{\text{TI}}(t)) \\ &+ w_{\text{GFP,RI}}\, g(u_{\text{RI}}(t)) \\ &+ w_{\text{GFP,GOI}}\, g(u_{\text{GOI}}(t)) \\ &+ w_{\text{GFP,GRI}}\, g(u_{\text{GRI}}(t)).\end{aligned} \tag{41}$$

The weights of these connections are set such that the prior intention node can only be activated when the target process or the reference process are active in conjunction with the ground object process or the ground relation process.

The activation u_{GFI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{GFI}}(t) = &- u_{\text{GFI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{GFI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{GFI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{GFP}}(t)) \\ &- w_{\text{I,M}}\, g(u_{\text{GFM}}(t)),\end{aligned} \tag{42}$$

which is analogous to Equation 3.54.

The activation u_{GFC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GFC}}(t) = &- u_{\text{GFC}}(t) + h_{\text{C}} + w_{\xi} \cdot \xi_{\text{GFC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{GFC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{GFI}}(t)) \\ &- \max_{c}(g(u_{\text{AC}}(c,t))) + \max_{c}(g(u_{\text{ACcs}}(c,t))) \\ &- \max_{\phi}(g(u_{\text{AM}}(\phi,t))) + \max_{\phi}(g(u_{\text{AMcs}}(\phi,t))),\end{aligned} \tag{43}$$

where the fourth line formalizes inhibitory input from the color attention field and excitatory input from the color CoS field. If a color is specified by the user, the color attention field holds a peak representation of that color and thereby inhibits the CoS node. Once an object of that color is found in the scene, the color CoS field holds a peak as well and thereby may activate the CoS node. The fifth line formalizes the same connection structure for the motion attention field and the motion CoS field. If a color as well as a motion direction is specified, peaks in both the color CoS field and the motion CoS field are required to activate the CoS node of the feature process.

The activation u_{GFM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{GFM}}(t) = &- u_{\text{GFM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{GFM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{GFM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{GFC}}(t)) \\ &+ w_{\text{GFM,TI}}\, g(u_{\text{TI}}(t)) \\ &+ w_{\text{GFM,RI}}\, g(u_{\text{RI}}(t)) \\ &- w_{\text{GFM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{44}$$

where lines four and five are inputs from the intention nodes of the target process and the reference process. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.12 Target IOR process

The dynamic neural nodes that represent the target IOR process are governed by the following equations. The activation u_{TIP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TIP}}(t) = &- u_{\text{TIP}}(t) + h + w_{\xi} \cdot \xi_{\text{TIP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{TIP}}(t)) \\ &+ w_{\text{TIP,TI}}\, g(u_{\text{TI}}(t)),\end{aligned} \tag{45}$$

where the last line formalizes the excitatory input from the intention node of the target process.

The activation u_{TII} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{TII}}(t) = &- u_{\text{TII}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{TII}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{TII}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{TIP}}(t)) \\ &- w_{\text{I,M}}\, g(u_{\text{TIM}}(t)),\end{aligned} \tag{46}$$

which is analogous to Equation 3.54.

The activation u_{TIC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TIC}}(t) = &- u_{\text{TIC}}(t) + h_{\text{TIC}} + w_{\xi} \cdot \xi_{\text{TIC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{TIC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{TII}}(t)) \\ &+ w_{\text{TIC,IRcs}} \max_{x,y}(g(u_{\text{IRcs}}(x,y,t))),\end{aligned} \tag{47}$$

where the fourth line formalizes excitatory input from the target IOR CoS field, which checks whether the target IOR field represents the object currently represented in the target field. Please note

that since this CoS node receives input from a field of the model, rather than from other processes, its resting level is lowered from the default value.

The activation u_{TIM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TIM}}(t) = &- u_{\text{TIM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{TIM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{TIM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{TIC}}(t)) \\ &+ w_{\text{TIM,TI}}\, g(u_{\text{TI}}(t)) \\ &- w_{\text{TIM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{48}$$

where line four formalizes input from the intention node of the target process. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.13 Target memory node process

The dynamic neural nodes that represent the target memory node process are evolve in time based on the following differential equations. The activation u_{TNP} of the prior intention node follows the equation

$$\begin{aligned}\tau \dot{u}_{\text{TNP}}(t) = &- u_{\text{TNP}}(t) + h + w_{\xi} \cdot \xi_{\text{TNP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{TNP}}(t)) \\ &+ w_{\text{TNP,TI}}\, g(u_{\text{TI}}(t)),\end{aligned} \tag{49}$$

where the last line formalizes the excitatory input from the intention node of the target process.

The activation u_{TNI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{TNI}}(t) = &- u_{\text{TNI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{TNI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{TNI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{TNP}}(t)),\end{aligned} \tag{50}$$

which is analogous to Equation 3.54. However, please note that the intention node of this process is not inhibited by the CoS memory node.

The activation u_{TNC} of the CoS node evolves in time based on the following equation, where lines four and five formalize inhibitory input from the prior intention nodes and excitatory input from the CoS nodes of the two processes on lowest hierarchical

level (light blue lines in Figure 3.14)

$$\begin{aligned}\tau \dot{u}_{\text{TNC}}(t) = &- u_{\text{TNC}}(t) + h_{\text{C}} + w_{\xi} \cdot \xi_{\text{TNC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{TNC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{TNI}}(t)) \\ &- g(u_{\text{TNCP}}(t)) + g(u_{\text{TNCM}}(t)) \\ &- g(u_{\text{TNMP}}(t)) + g(u_{\text{TNMM}}(t)).\end{aligned} \tag{51}$$

The activation u_{TNM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TNM}}(t) = &- u_{\text{TNM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{TNM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{TNM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{TNC}}(t)) \\ &+ w_{\text{TNM,TI}}\, g(u_{\text{TI}}(t)) \\ &- w_{\text{TNM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{52}$$

where line four formalizes input from the intention node of the target process. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.14 Target motion field process

The dynamic neural nodes that represent the target motion field process are governed by the following equations. The activation u_{TMP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TMP}}(t) = &- u_{\text{TMP}}(t) + h + w_{\xi} \cdot \xi_{\text{TMP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{TMP}}(t)) \\ &+ w_{\text{TMP,TI}}\, g(u_{\text{TI}}(t)) \\ &+ w_{\text{TMP,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{TMP,DI}}\, g(u_{\text{DI}}(t)) \\ &+ w_{\text{TMP,MT}}\, g(u_{\text{MT}}(t)),\end{aligned} \tag{53}$$

where lines 3–5 formalize the excitatory input from the intention nodes of the target process, the ground relation process, and the describe process (from top to bottom). The weights of these connections are set such that the prior intention node can only be activated when the target process is active in conjunction with the ground relation process or the describe process. The last line is input from the *motion term node*, which is active whenever one of the spatial relation memory nodes is active that code for dynamic spatial relations (i.e., TOWARD or AWAY). The motion term node with activation u_{MT}

is governed by the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{MT}}(t) = &- u_{\text{MT}}(t) + h + w_{\xi} \cdot \xi_{\text{MT}}(t) \\ &+ w_{\text{MT,MT}}\, g(u_{\text{MT}}(t)) \\ &+ \max_{i=1,\ldots,N_{\text{R}}} (w_{\text{MT,SM}_i}\, g(u_{\text{SM}i}(t))),\end{aligned} \tag{54}$$

where the third line is input from the spatial relation memory nodes with the maximum activation $\vec{u}_{\text{SM}}$ weighted by the vector $\vec{w}_{\text{MT,SM}} = (0, 0, 0, 0, a, a)^T$. The constant $a > 0$ is set such that the motion term node is activated whenever one of the last two spatial relation memory nodes is active.

The activation u_{TMI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{TMI}}(t) = &- u_{\text{TMI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{TMI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{TMI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{TMP}}(t)) \\ &- w_{\text{I,M}}\, g(u_{\text{TMM}}(t)),\end{aligned} \tag{55}$$

which is analogous to Equation 3.54.

The activation u_{TMC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TMC}}(t) = &- u_{\text{TMC}}(t) + h_{\text{TMC}} + w_{\xi} \cdot \xi_{\text{TMC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{TMC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{TMI}}(t)) \\ &+ w_{\text{TMC,ROTs}} \max_{x,y}(g(u_{\text{ROTs}}(x, y, t))),\end{aligned} \tag{56}$$

where the last line formalizes excitatory input from the rotation selection field, which holds a peak if a motion direction has been extracted by the motion detection system. Please note that since the CoS node receives input from a field of the model, rather than from other processes, its resting level is lowered from the default value.

The activation u_{TMM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TMM}}(t) = &- u_{\text{TMM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{TMM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{TMM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{TMC}}(t)) \\ &+ w_{\text{TMM,TI}}\, g(u_{\text{TI}}(t)) \\ &- w_{\text{TMM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{57}$$

where line four formalizes input from the intention node of the target process. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.15 Target field process

The dynamic neural nodes that represent the target field process are governed by the following equations. The activation u_{TTP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{TTP}}(t) = &- u_{\mathrm{TTP}}(t) + h + w_{\xi} \cdot \xi_{\mathrm{TTP}}(t) \\ &+ w_{\mathrm{P,P}}\, g(u_{\mathrm{TTP}}(t)) \\ &+ w_{\mathrm{TTP,TI}}\, g(u_{\mathrm{TI}}(t)),\end{aligned} \tag{58}$$

where the last line formalizes the excitatory input from the intention nodes of the target process.

The activation u_{TTI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\mathrm{TTI}}(t) = &- u_{\mathrm{TTI}}(t) + h_{\mathrm{I}} + w_{\xi} \cdot \xi_{\mathrm{TTI}}(t) \\ &+ w_{\mathrm{I,I}}\, g(u_{\mathrm{TTI}}(t)) \\ &+ w_{\mathrm{I,P}}\, g(u_{\mathrm{TTP}}(t)) \\ &- w_{\mathrm{I,M}}\, g(u_{\mathrm{TTM}}(t)),\end{aligned} \tag{59}$$

which is analogous to Equation 3.54.

The activation u_{TTC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{TTC}}(t) = &- u_{\mathrm{TTC}}(t) + h_{\mathrm{TTC}} + w_{\xi} \cdot \xi_{\mathrm{TTC}}(t) \\ &+ w_{\mathrm{C,C}}\, g(u_{\mathrm{TTC}}(t)) \\ &+ w_{\mathrm{C,I}}\, g(u_{\mathrm{TTI}}(t)) \\ &+ w_{\mathrm{TTC,T}} \max_{x,y}(g(u_{\mathrm{T}}(x,y,t))),\end{aligned} \tag{60}$$

where the last line formalizes excitatory input from the target field. Please note that since the CoS node receives input from a field of the model, rather than from other processes, its resting level is lowered from the default value.

The activation u_{TTM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\mathrm{TTM}}(t) = &- u_{\mathrm{TTM}}(t) + h_{\mathrm{M}} + w_{\xi} \cdot \xi_{\mathrm{TTM}}(t) \\ &+ w_{\mathrm{M,M}}\, g(u_{\mathrm{TTM}}(t)) \\ &+ w_{\mathrm{M,C}}\, g(u_{\mathrm{TTC}}(t)) \\ &+ w_{\mathrm{TTM,TI}}\, g(u_{\mathrm{TI}}(t)) \\ &- w_{\mathrm{TTM,SR}}\, g(u_{\mathrm{SR}}(t)),\end{aligned} \tag{61}$$

where line four formalizes input from the intention node of the target process. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.16 Reference memory node process

The dynamic neural nodes representing the reference memory node process evolve in time based on the following differential equations. The activation u_{RNP} of the prior intention node follows the equation

$$\begin{aligned}\tau \dot{u}_{\text{RNP}}(t) = &- u_{\text{RNP}}(t) + h + w_\xi \cdot \xi_{\text{RNP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{RNP}}(t)) \\ &+ w_{\text{RNP,RI}}\, g(u_{\text{RI}}(t)),\end{aligned} \tag{62}$$

where the last line formalizes the excitatory input from the intention node of the reference process.

The activation u_{RNI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{RNI}}(t) = &- u_{\text{RNI}}(t) + h_{\text{I}} + w_\xi \cdot \xi_{\text{RNI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{RNI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{RNP}}(t)),\end{aligned} \tag{63}$$

which is analogous to Equation 3.54. However, please note that the intention node of this process is not inhibited by the CoS memory node.

The activation u_{RNC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{RNC}}(t) = &- u_{\text{RNC}}(t) + h_{\text{C}} + w_\xi \cdot \xi_{\text{RNC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{RNC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{RNI}}(t)) \\ &+ w_{\text{RNC,RCM}} \max_{i=1,\ldots,N_{\text{C}}} (g(u_{\text{RCM}i}(t))),\end{aligned} \tag{64}$$

where the fourth line formalizes excitatory input from all reference color memory nodes with activation $\vec{u}_{\text{RCM}}$.

The activation u_{RNM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{RNM}}(t) = &- u_{\text{RNM}}(t) + h_{\text{M}} + w_\xi \cdot \xi_{\text{RNM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{RNM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{RNC}}(t)) \\ &+ w_{\text{RNM,RI}}\, g(u_{\text{RI}}(t)) \\ &- w_{\text{RNM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{65}$$

where line four formalizes input from the intention node of the reference process. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.17 Reference field process

The dynamic neural nodes that represent the reference field process are governed by the following equations. The activation u_{RFP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{RFP}}(t) = &- u_{\text{RFP}}(t) + h + w_\xi \cdot \xi_{\text{RFP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{RFP}}(t)) \\ &+ w_{\text{RFP,RI}}\, g(u_{\text{RI}}(t)),\end{aligned} \tag{66}$$

where the last line formalizes the excitatory input from the intention nodes of the reference process.

The activation u_{RFI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{RFI}}(t) = &- u_{\text{RFI}}(t) + h_{\text{I}} + w_\xi \cdot \xi_{\text{RFI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{RFI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{RFP}}(t)) \\ &- w_{\text{I,M}}\, g(u_{\text{RFM}}(t)),\end{aligned} \tag{67}$$

which is analogous to Equation 3.54.

The activation u_{RFC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{RFC}}(t) = &- u_{\text{RFC}}(t) + h_{\text{RFC}} + w_\xi \cdot \xi_{\text{RFC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{RFC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{RFI}}(t)) \\ &+ w_{\text{RFC,R}} \max_{x,y}(g(u_{\text{R}}(x,y,t))),\end{aligned} \tag{68}$$

where the last line formalizes excitatory input from the reference field. Please note that since the CoS node receives input from a field of the model, rather than from other processes, its resting level is lowered from the default value.

The activation u_{RFM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{RFM}}(t) = &- u_{\text{RFM}}(t) + h_{\text{M}} + w_\xi \cdot \xi_{\text{RFM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{RFM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{RFC}}(t)) \\ &+ w_{\text{RFM,RI}}\, g(u_{\text{RI}}(t)) \\ &- w_{\text{RFM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{69}$$

where line four formalizes input from the intention node of the reference process. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.18 Spatial memory node process

The dynamic neural nodes that represent the spatial memory node process are governed by the following differential equations. The activation u_{SNP} of the prior intention node follows the equation

$$\begin{aligned}\tau \dot{u}_{\text{SNP}}(t) = &- u_{\text{SNP}}(t) + h + w_{\xi} \cdot \xi_{\text{SNP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{SNP}}(t)) \\ &+ w_{\text{SNP,SI}}\, g(u_{\text{SI}}(t)),\end{aligned} \tag{70}$$

where the last line formalizes the excitatory input from the intention node of the spatial relation process.

The activation u_{SNI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{SNI}}(t) = &- u_{\text{SNI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{SNI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{SNI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{SNP}}(t)) \\ &- w_{\text{SNI,PNR}}\, g(u_{\text{PNR}}(t)),\end{aligned} \tag{71}$$

which is analogous to Equation 3.54. However, please note that the intention node of this process is not inhibited by the CoS memory node. The last line is inhibitory input from a precondition node that ensures the spatial memory node process is only activated once the spatial relational field process is finished.[6]

[6] Please refer to Section B.22, in particular Equation 88.

The activation u_{SNC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SNC}}(t) = &- u_{\text{SNC}}(t) + h_{\text{SNC}} + w_{\xi} \cdot \xi_{\text{SNC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{SNC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{SNI}}(t)) \\ &+ w_{\text{SNC,SM}} \max_{i=1,\ldots,N_{\text{R}}} (g(u_{\text{SM}i}(t))),\end{aligned} \tag{72}$$

where the fourth line formalizes the excitatory input from all spatial relation memory nodes with activation $\vec{u}_{\text{SM}}$. Please note that since the CoS node receives input from a field of the model, rather than from other processes, its resting level is lowered from the default value.

The activation u_{SNM} of the CoS memory node is governed by the following differential equation, where line four formalizes input from the intention node of the spatial relation process and the last line formalizes inhibitory input from the suppression node activated

by the reset process

$$\begin{aligned}\tau \dot{u}_{\text{SNM}}(t) = &- u_{\text{SNM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{SNM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{SNM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{SNC}}(t)) \\ &+ w_{\text{SNM,SI}}\, g(u_{\text{SI}}(t)) \\ &- w_{\text{SNM,SR}}\, g(u_{\text{SR}}(t)).\end{aligned} \tag{73}$$

B.19 Spatial relational field process

The dynamic neural nodes that represent the spatial relational field process are governed by the following differential equations. The activation u_{SRP} of the prior intention node follows the equation

$$\begin{aligned}\tau \dot{u}_{\text{SRP}}(t) = &- u_{\text{SRP}}(t) + h + w_{\xi} \cdot \xi_{\text{SRP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{SRP}}(t)) \\ &+ w_{\text{SRP,SI}}\, g(u_{\text{SI}}(t)) \\ &+ w_{\text{SRP,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{SRP,DI}}\, g(u_{\text{DI}}(t)),\end{aligned} \tag{74}$$

where the last line formalizes the excitatory input from the intention nodes of the spatial relation process, the ground relation process, and the describe process. The weights of these connections are set such that the prior intention node can only be activated when the spatial relation process is active in conjunction with the ground relation process or the describe process.

The activation u_{SRI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{SRI}}(t) = &- u_{\text{SRI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{SRI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{SRI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{SRP}}(t)) \\ &- w_{\text{I,M}}\, g(u_{\text{SRM}}(t)),\end{aligned} \tag{75}$$

which is analogous to Equation 3.54.

The activation u_{SRC} of the CoS node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SRC}}(t) = &- u_{\text{SRC}}(t) + h_{\text{SRC}} + w_{\xi} \cdot \xi_{\text{SRC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{SRC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{SRI}}(t)) \\ &+ w_{\text{SRC,Scs}} \max_{x,y}(g(u_{\text{Scs}}(x, y, t))),\end{aligned} \tag{76}$$

where the last line formalizes excitatory input from the spatial relation CoS field. Please note that since the CoS node receives input

from a field of the model, rather than from other processes, its resting level is lowered from the default value.

The activation u_{SRM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{SRM}}(t) = &- u_{\text{SRM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{SRM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{SRM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{SRC}}(t)) \\ &+ w_{\text{SRM,SI}}\, g(u_{\text{SI}}(t)) \\ &- w_{\text{SRM,SR}}\, g(u_{\text{SR}}(t)),\end{aligned} \tag{77}$$

where line four formalizes input from the intention node of the spatial relation process. The last line formalizes inhibitory input from the suppression node activated by the reset process.

B.20 Target memory node color process

The dynamic neural nodes that represent the target memory node color process are governed by the following equations. The activation u_{TNCP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TNCP}}(t) = &- u_{\text{TNCP}}(t) + h + w_{\xi} \cdot \xi_{\text{TNCP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{TNCP}}(t)) \\ &+ w_{\text{TNCP,TNI}}\, g(u_{\text{TNI}}(t)),\end{aligned} \tag{78}$$

where the last line formalizes the excitatory input from the intention node of the target memory node process.

The activation u_{TNCI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{TNCI}}(t) = &- u_{\text{TNCI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{TNCI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{TNCI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{TNCP}}(t)),\end{aligned} \tag{79}$$

which is analogous to Equation 3.54. However, please note that the intention node of this process is not inhibited by the CoS memory node.

The activation u_{TNCC} of the CoS node evolves in time based on the following differential equation, where the fourth line formalizes excitatory input from all target color memory nodes with activation $\vec{u}_{\text{TCM}}$. Please note that since the CoS node receives input from a field of the model, rather than from other processes, its

resting level is lowered from the default value

$$\begin{aligned}\tau \dot{u}_{\text{TNCC}}(t) = &- u_{\text{TNCC}}(t) + h_{\text{TNCC}} + w_{\xi} \cdot \xi_{\text{TNCC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{TNCC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{TNCI}}(t)) \\ &+ w_{\text{TNCC,TCM}} \max_{i=1,\ldots,N_{\text{C}}} (g(u_{\text{TCM}i}(t))).\end{aligned} \tag{80}$$

The activation u_{TNCM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TNCM}}(t) = &- u_{\text{TNCM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{TNCM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{TNCM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{TNCC}}(t)) \\ &+ w_{\text{TNCM,TNI}}\, g(u_{\text{TNI}}(t)),\end{aligned} \tag{81}$$

where line four formalizes input from the intention node of the target memory node process.

B.21 Target memory node motion process

The dynamic neural nodes that represent the target memory node motion process are governed by the following equations. The activation u_{TNMP} of the prior intention node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TNMP}}(t) = &- u_{\text{TNMP}}(t) + h + w_{\xi} \cdot \xi_{\text{TNMP}}(t) \\ &+ w_{\text{P,P}}\, g(u_{\text{TNMP}}(t)) \\ &+ w_{\text{TNMP,TNI}}\, g(u_{\text{TNI}}(t)),\end{aligned} \tag{82}$$

where the last line formalizes the excitatory input from the intention node of the target memory node process.

The activation u_{TNMI} of the intention node is governed by

$$\begin{aligned}\tau \dot{u}_{\text{TNMI}}(t) = &- u_{\text{TNMI}}(t) + h_{\text{I}} + w_{\xi} \cdot \xi_{\text{TNMI}}(t) \\ &+ w_{\text{I,I}}\, g(u_{\text{TNMI}}(t)) \\ &+ w_{\text{I,P}}\, g(u_{\text{TNMP}}(t)),\end{aligned} \tag{83}$$

which is analogous to Equation 3.54. However, please note that the intention node of this process is not inhibited by the CoS memory node.

The activation u_{TNMC} of the CoS node is governed by the following differential equation, where the fourth line formalizes excitatory

input from all target motion memory nodes with activation $\vec{u}_{\text{TMM}}$

$$\begin{aligned}\tau \dot{u}_{\text{TNMC}}(t) = &- u_{\text{TNMC}}(t) + h_{\text{TNMC}} + w_{\xi} \cdot \xi_{\text{TNMC}}(t) \\ &+ w_{\text{C,C}}\, g(u_{\text{TNMC}}(t)) \\ &+ w_{\text{C,I}}\, g(u_{\text{TNMI}}(t)) \\ &+ w_{\text{TNMC,TMM}} \max_{i=1,\ldots,N_{\text{M}}} (g(u_{\text{TMM}i}(t))).\end{aligned} \tag{84}$$

Please note that since the CoS node receives input from a field of the model, rather than from other processes, its resting level is lowered from the default value.

The activation u_{TNMM} of the CoS memory node follows the differential equation

$$\begin{aligned}\tau \dot{u}_{\text{TNMM}}(t) = &- u_{\text{TNMM}}(t) + h_{\text{M}} + w_{\xi} \cdot \xi_{\text{TNMM}}(t) \\ &+ w_{\text{M,M}}\, g(u_{\text{TNMM}}(t)) \\ &+ w_{\text{M,C}}\, g(u_{\text{TNMC}}(t)) \\ &+ w_{\text{TNMM,TNI}}\, g(u_{\text{TNI}}(t)),\end{aligned} \tag{85}$$

where line four formalizes input from the intention node of the target memory node process.

B.22 Sequentiality

Target process, reference process, and clean process

The sequential order in which the target process, the clean process, and the reference process are activated is implemented using two precondition nodes.

The first ensures that the clean process is activated after the target process. Its activation u_{PCT} is governed by the differential equation

$$\begin{aligned}\tau_{\text{PCT}} \dot{u}_{\text{PCT}}(t) = &- u_{\text{PCT}}(t) + h_{\text{PCT}} + w_{\xi} \cdot \xi_{\text{PCT}}(t) \\ &+ w_{\text{PCT,PCT}}\, g(u_{\text{PCT}}(t)) \\ &+ w_{\text{PCT,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{PCT,DI}}\, g(u_{\text{DI}}(t)) \\ &- w_{\text{PCT,TM}}\, g(u_{\text{TM}}(t)),\end{aligned} \tag{86}$$

where the third and fourth line are excitatory inputs from the intention nodes of the ground relation process and the describe process, respectively. Both of these inputs can activate the precondition node by itself. The last line is inhibitory input from the CoS memory node of the target process, which deactivates the node. The precondition node itself has an inhibitory connection to the intention node of the clean process.[7]

The second precondition node ensures that the reference process is activated after the clean process. Its activation u_{PRC} evolves in time based on the differential equation

$$\begin{aligned}\tau_{\text{PRC}}\dot{u}_{\text{PRC}}(t) = &- u_{\text{PRC}}(t) + h_{\text{PRC}} + w_{\xi} \cdot \xi_{\text{PRC}}(t) \\ &+ w_{\text{PRC,PRC}}\, g(u_{\text{PRC}}(t)) \\ &+ w_{\text{PRC,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{PRC,DI}}\, g(u_{\text{DI}}(t)) \\ &- w_{\text{PRC,TM}}\, g(u_{\text{TM}}(t)),\end{aligned} \tag{87}$$

where the third and fourth line are analogous to Equation 86. The last line is inhibitory input from the CoS memory node of the clean process, which deactivates the node. The precondition node itself has an inhibitory connection to the intention node of the reference process.[8]

[8] See Equation 18 in Section B.

Producing a response after the relation has been evaluated

The precondition node that ensures that the spatial memory node process is only activated once the spatial relational field process is completed has the activation u_{PNR}, which follows the differential equation

$$\begin{aligned}\tau_{\text{PNR}}\dot{u}_{\text{PNR}}(t) = &- u_{\text{PNR}}(t) + h + w_{\xi} \cdot \xi_{\text{PNR}}(t) \\ &+ w_{\text{PNR,PNR}}\, g(u_{\text{PNR}}(t)) \\ &+ w_{\text{PNR,SI}}\, g(u_{\text{SI}}(t)) \\ &- w_{\text{PNR,SRM}}\, g(u_{\text{SRM}}(t)),\end{aligned} \tag{88}$$

where the third line is excitatory input from the intention node of the spatial relation process, which activates the precondition node. The last line is inhibitory input from the CoS memory node of the spatial relational field process, which deactivates the node. The precondition node itself has an inhibitory connection to the intention node of the spatial memory node process.[9]

[9] See Equation 71 in Section B.

A second precondition node follows the equation

$$\begin{aligned}\tau_{\text{PAR}}\dot{u}_{\text{PAR}}(t) = &- u_{\text{PAR}}(t) + h + w_{\xi} \cdot \xi_{\text{PAR}}(t) \\ &+ w_{\text{PAR,PAR}}\, g(u_{\text{PAR}}(t)) \\ &+ w_{\text{PAR,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{PAR,DI}}\, g(u_{\text{DI}}(t)) \\ &+ w_{\text{PAR,RI}}\, g(u_{\text{RI}}(t)) \\ &- w_{\text{PAR,SRM}}\, g(u_{\text{SRM}}(t)),\end{aligned} \tag{89}$$

where u_{PAR} is its own activation. It ensures that the spatial attention process is only activated once the spatial relational field process

is completed is governed by the equation Lines 3–5 are excitatory input from the intention nodes of the ground relation process, the describe process, and the reference process. The connection weights are set such that the precondition node is only activated when the reference process is active in conjunction with the ground relation process or the describe process. The last line is inhibitory input from the CoS memory node of the spatial relational field process, which deactivates the node. The precondition node itself has an inhibitory connection to the intention node of the spatial attention process.[10]

[10]See Equation 38 in Section B.

Reset process

The activation u_{SR} of the suppression node that implements the inhibition of the reset process is governed by the differential equation

$$\begin{aligned}\tau_{\text{SR}}\dot{u}_{\text{SR}}(t) = &- u_{\text{SR}}(t) + h + w_\xi \cdot \xi_{\text{SR}}(t) \\ &+ w_{\text{SR,SR}}\, g(u_{\text{SR}}(t)) \\ &+ w_{\text{SR,EI}}\, g(u_{\text{EI}}(t)),\end{aligned} \tag{90}$$

where the third line is excitatory input from the intention node of the reset process, which directly activates the suppression node. The suppression node itself has an inhibitory connection to a large number of nodes and fields of the model, too many to list here.

The activation u_{PRD} of the precondition node that inhibits the reset process evolves in time based on the differential equation

$$\begin{aligned}\tau_{\text{PRD}}\dot{u}_{\text{PRD}}(t) = &- u_{\text{PRD}}(t) + h_{\text{PRD}} + w_\xi \cdot \xi_{\text{PRD}}(t) \\ &+ w_{\text{PRD,PRD}}\, g(u_{\text{PRD}}(t)) \\ &+ w_{\text{PRD,GRI}}\, g(u_{\text{GRI}}(t)) \\ &+ w_{\text{PRD,DI}}\, g(u_{\text{DI}}(t)) \\ &- w_{\text{PRD,Scd}} \max_{x,y}(g(u_{\text{Scd}}(x,y,t))),\end{aligned} \tag{91}$$

where lines three and four are excitatory input from the intention nodes of the ground relation process and the describe process. Both processes directly activate the precondition node. The last line is inhibitory input from the spatial relation CoD field, which deactivates the node. The precondition node itself has an inhibitory connection to the intention node of the reset process.[11]

[11]See Equation 30 in Section B.

C Video data set

The video data set consists of 82 video clips, all of which feature colored balls on a white background. The videos differ in the number of balls in the scene, the number of balls that have the same color, the number of moving balls, as well as the spatial arrangement of the balls in the scene.

All balls are 6.5 cm in diameter and are made of a matte, uniformly colored rubber that is either red, orange, yellow, green, or blue. The balls in the video are placed in a wooden area with a matte white finish that is 1 m^2 in size. The camera[12] that was used to record the videos was placed at a height of 130 cm, at a horizontal distance of 70 cm from the center of the arena (Figure C.1). The camera was set up with a negative inclination of 50° from horizontal. This setup leads to some perspective distortion. As a result, objects that are located in the lower part of the video appear larger than objects in the upper part.

[12] Sony XCD-V60CR firewire camera

All videos were recorded with a resolution of 640x480 pixels at 60 frames per second. Videos were then compressed with the video codec H.264/MPEG-4 AVC. Most videos were only a couple of seconds long, all were shorter than 30 s (average 5.72 s, standard deviation 6.49 s).

The full video data set is freely available online at https://www.ini.rub.de/pages/publications/richterphdthesis.

Table C.1 lists the video file that was used for each test described in Section 4.

The following figures show snapshots of all videos in the data set. For videos of moving objects, snapshots were taken at every second, unless noted otherwise. For videos of static objects, only a single frame is shown. A description of what is visible in the videos can be found in the figure captions.

test ID	video ID
G1	0.00
G2	4.00
G3	4.01a
G4	4.02a
G5	4.03
G6	4.04a
G7	4.05a
G8	4.06a
G9	4.07
G10	4.08
G11	4.09a
G12	4.10a
G13	4.11a
G14	4.12a
G15	4.21
G16	4.14e
G17	4.23
G18	4.15a
G19	0.00
G20	4.00
G21	4.01a
G22	4.10a
G23	4.15c
G24	4.06b
G25	0.00
G26	4.00
G27	4.01b
G28	4.01a
G29	4.02a
G30	4.02b
G31	4.03
G32	4.05a
G33	4.05b
G34	4.04a
G35	4.04b

test ID	video ID
G36	4.16a
G37	4.16b
G38	4.06b
G39	4.06a
G40	4.06c
G41	4.06d
G42	4.08
G43	4.09a
G44	4.09b
G45	4.10a
G46	4.10b
G47	4.11a
G48	4.11b
G49	4.12d
G50	4.12c
G51	4.12b
G52	4.12a
G53	4.14e
G54	4.14f
G55	4.15a
G56	4.15b
G57	4.15c
G58	2.00a
G59	2.00c
G60	2.01a
G61	2.01b
G62	2.02a
G63	2.02b
G64	2.03a
G65	2.03b
G66	3.04a
G67	3.04b
G68	3.04c
G69	3.05a
G70	3.05b

test ID	video ID
G71	3.05c
G72	4.18a
G73	4.18b
G74	4.18c
G75	2.00a
G76	2.01
G77	2.02c
G78	2.02d
G79	2.03c
G80	2.03d
G81	3.06a
G82	3.06b
G83	3.06c
G84	3.07a
G85	3.07b
G86	3.07c
G87	4.19a
G88	4.19b
G89	4.19c
P1	0.00
P2	1.00
P3	1.02
P4	1.01
P5	1.03
P6	2.00a
P7	2.00b
P8	3.01a
P9	3.01b
P10	3.01c
P11	3.01d
P12	4.06e
P13	4.06f
P14	4.06g
P15	4.06h

Table C.1: Identifier of the video file used for each test.

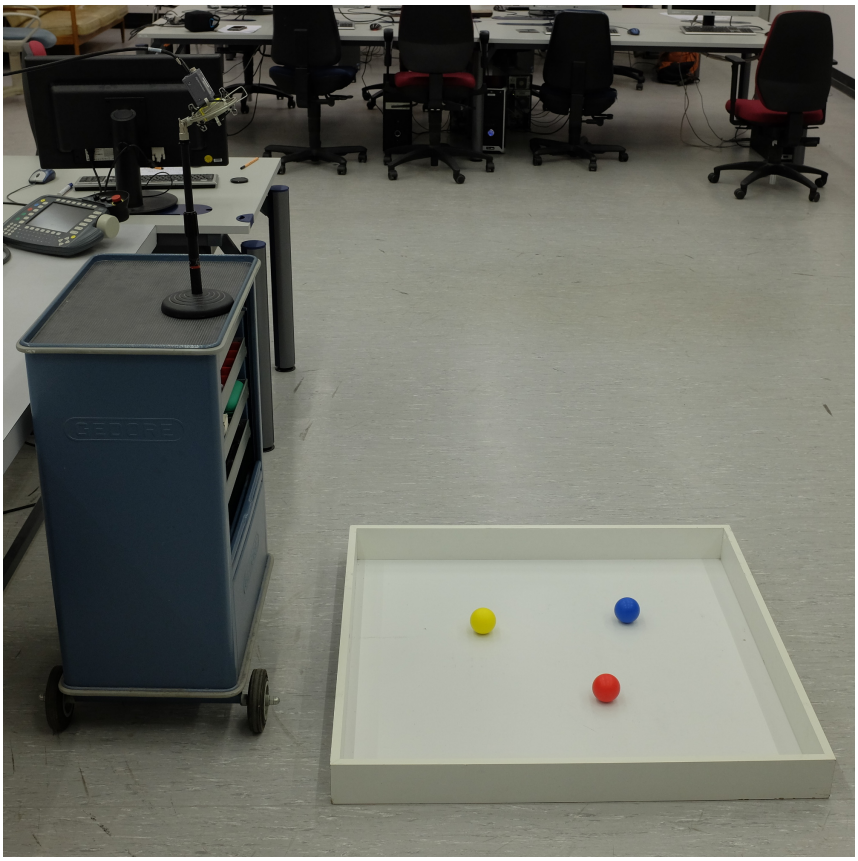


Figure C.1: Photo of the setup in which the videos were recorded.

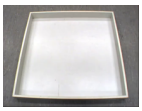


Figure C.2: Snapshot of the video 0.00: no objects in the scene, just the white background.

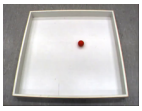


Figure C.3: Snapshot of the video 1.00: a single static red object in the scene.

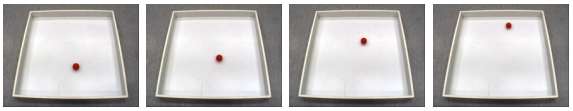


FIGURE C.4: Snapshots of the video 1.01: a single red object moving upward.

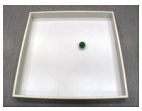


FIGURE C.5: Snapshot of the video 1.02: a single static green object.

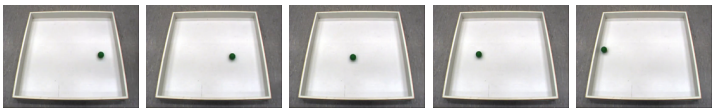


FIGURE C.6: Snapshots of the video 1.03: a single green object moving leftward.

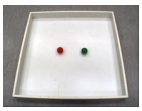


FIGURE C.7: Snapshot of the video 2.00a: a static red object to the left of a static green object.

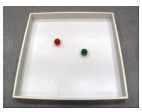


FIGURE C.8: Snapshot of the video 2.00b: a static red object to the left and slightly above a static green object.

FIGURE C.9: Snapshot of the video 2.00c: a static red object to the right of a static green object.

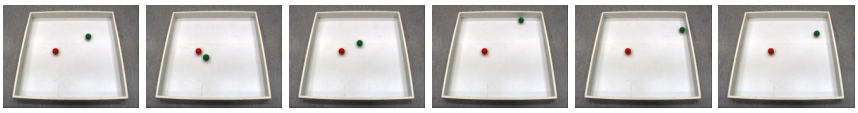


FIGURE C.10: Snapshots (3 s apart) of the video 2.01a: a static red object to the left of a moving green object.

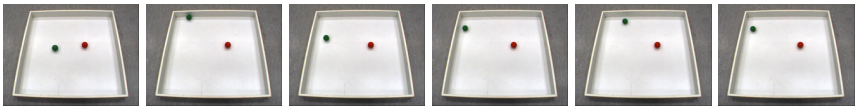


FIGURE C.11: Snapshots (4 s apart) of the video 2.01b: a static red object to the right of a moving green object.

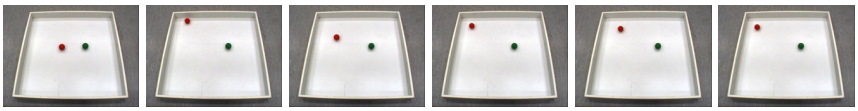


FIGURE C.12: Snapshots (4 s apart) of the video 2.02a: a moving red object to the left of a static green object.

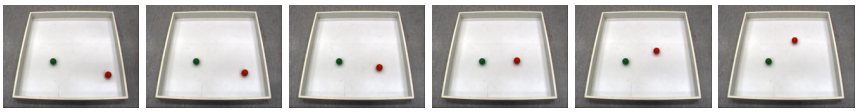


FIGURE C.13: Snapshots of the video 2.02b: a moving red object to the right of a static green object.

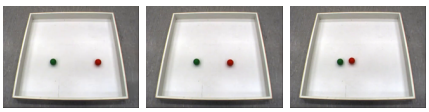


FIGURE C.14: Snapshots of the video 2.02c: a red object moving toward a static green object.

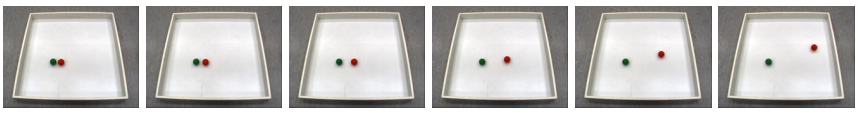


FIGURE C.15: Snapshots of the video 2.02d: a red object moving away from a static green object.

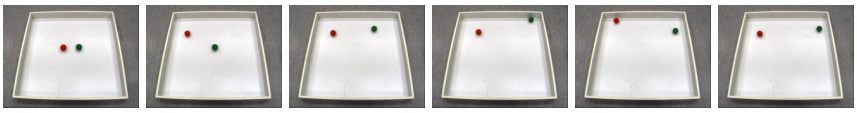


FIGURE C.16: Snapshots (4 s apart) of the video 2.03a: a moving red object to the left of a moving green object.

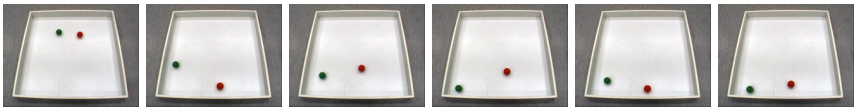


FIGURE C.17: Snapshots (4 s apart) of the video 2.03b: a moving red object to the right of a moving green object.

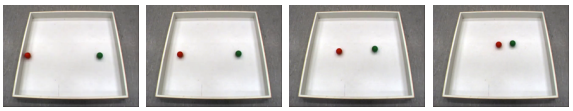


FIGURE C.18: Snapshots of the video 2.03c: a red object moving toward a moving green object.

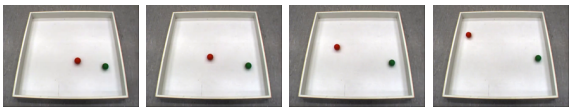


FIGURE C.19: Snapshots of the video 2.03d: a red object moving away from a moving green object.

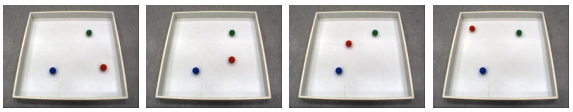


FIGURE C.20: Snapshots of the video 3.01a: a red object moving in between a static green object and a static blue object.

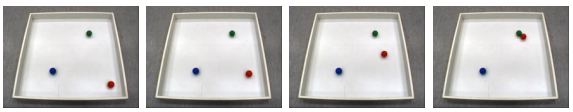


FIGURE C.21: Snapshots of the video 3.01b: a red object moving toward a static green object; an additional blue object is in the scene as well.

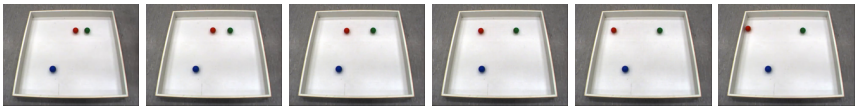


FIGURE C.22: Snapshots of the video 3.01c: a red object moving away from a static green object; an additional blue object is in the scene as well.

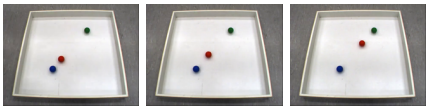


FIGURE C.23: Snapshots of the video 3.01d: a red object moving away from a static blue object and toward a static green object.

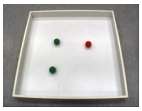


FIGURE C.24: Snapshot of the video 3.04a: a static red object to the right of two static green objects.

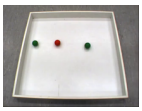


FIGURE C.25: Snapshot of the video 3.04b: a static red object to the right of a static green object and to the left of another static green object.

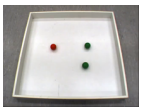


FIGURE C.26: Snapshot of the video 3.04c: a static red object to the left of two static green objects.

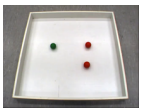


FIGURE C.27: Snapshot of the video 3.05a: two static red objects to the right of a static green object.

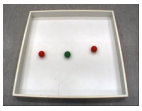


FIGURE C.28: Snapshot of the video 3.05b: a static red object to the left of a static green object; a second static red object to the right of the green object.

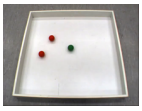


FIGURE C.29: Snapshot of the video 3.05c: two static red objects to the left of a static green object.

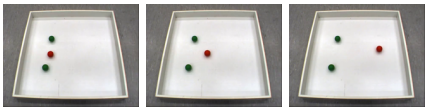


FIGURE C.30: Snapshots of the video 3.06a: a red object moving away from two static green objects.

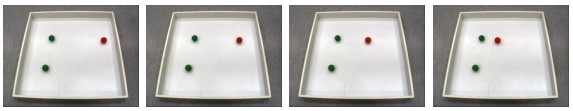


FIGURE C.31: Snapshots of the video 3.06b: a red object moving toward a static green object; another static green object is in the scene.

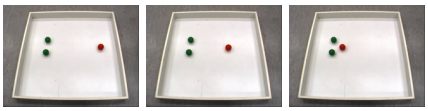


FIGURE C.32: Snapshots of the video 3.06c: a red object moving toward two static green objects.

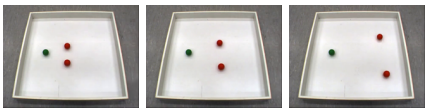


FIGURE C.33: Snapshots of the video 3.07a: two red objects moving away from a static green object.

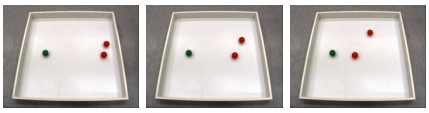


FIGURE C.34: Snapshots of the video 3.07b: a red object moving toward a static green object; another red object moving, but not toward the green object.

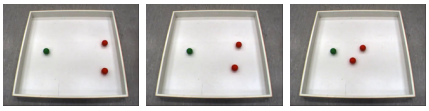


FIGURE C.35: Snapshots of the video 3.07c: two red objects moving toward a static green object.

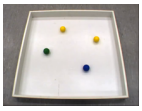


FIGURE C.36: Snapshot of the video 4.00: four static objects (yellow, orange, blue, green) in the scene.

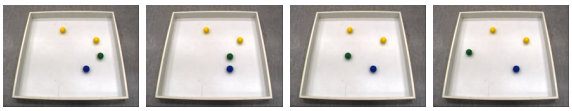


FIGURE C.37: Snapshots of the video 4.01a: a green object moving leftward; three other static objects (yellow, orange, blue) are in the scene.

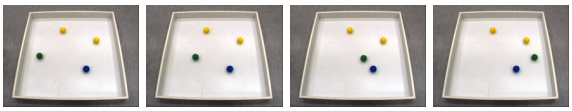


FIGURE C.38: Snapshots of the video 4.01b: a green object moving rightward; three other static objects (yellow, orange, blue) are in the scene.

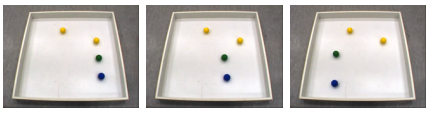


FIGURE C.39: Snapshots of the video 4.02a: a blue and a green object moving leftward; two other static objects (yellow, orange) are in the scene.

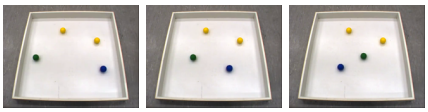


FIGURE C.40: Snapshots of the video 4.02b: a blue object moving rightward; a green object moving leftward; two other static objects (yellow, orange) are in the scene.

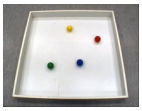


FIGURE C.41: Snapshot of the video 4.03: four static objects (red, yellow, blue, green) in the scene.

FIGURE C.42: Snapshots of the video 4.04a: a red object moving rightward; three other static objects (yellow, blue, green) are in the scene.

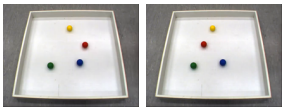


FIGURE C.43: Snapshots of the video 4.04b: a red object moving leftward; three other static objects (yellow, blue, green) are in the scene.

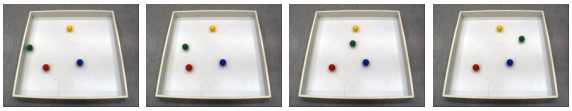


FIGURE C.44: Snapshots of the video 4.05a: a green object moving rightward; three other static objects (red, yellow, blue) are in the scene.

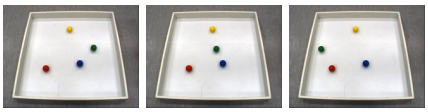


FIGURE C.45: Snapshots of the video 4.05b: a green object moving leftward; three other static objects (red, yellow, blue) are in the scene.

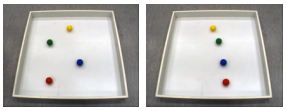


FIGURE C.46: Snapshots of the video 4.06a: a red and a green object moving rightward; two other static objects (yellow, blue) are in the scene.

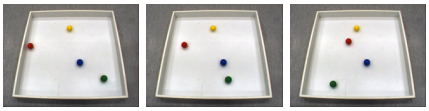


FIGURE C.47: Snapshots of the video 4.06b: a red object moving rightward; a green object moving leftward; two other static objects (yellow, blue) are in the scene.

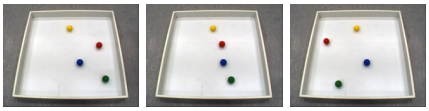


FIGURE C.48: Snapshots of the video 4.06c: a red and a green object moving leftward; two other static objects (yellow, blue) are in the scene.

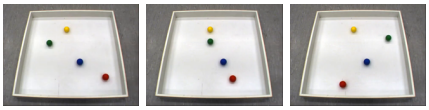


FIGURE C.49: Snapshots of the video 4.06d: a red object moving leftward; a green object moving rightward; two other static objects (yellow, blue) are in the scene.

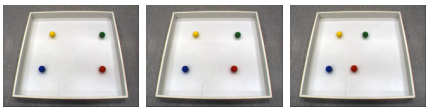


FIGURE C.50: Snapshots of the video 4.06e: a red object moving toward a static blue object; a green object moving toward a static yellow object.

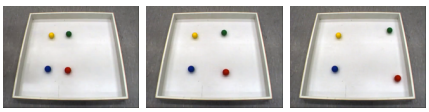


FIGURE C.51: Snapshots of the video 4.06f: a red object moving away from a static blue object; a green object moving away from a static yellow object.

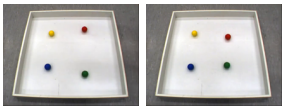


FIGURE C.52: Snapshots of the video 4.06g: a red and a green object moving toward each other; two other static objects (yellow, blue) are in the scene.

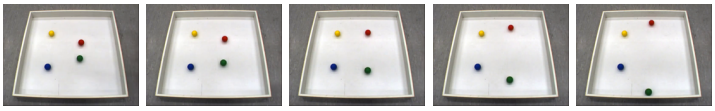


FIGURE C.53: Snapshots of the video 4.06h: a red and a green object moving away from each other; two other static objects (yellow, blue) are in the scene.

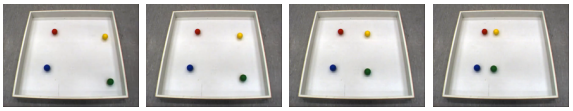


FIGURE C.54: Snapshots of the video 4.07: a green object moving toward a static blue object; a yellow object moving toward a static red object.

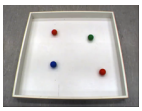


FIGURE C.55: Snapshot of the video 4.08: two static red objects, a static green object, and a static blue object.

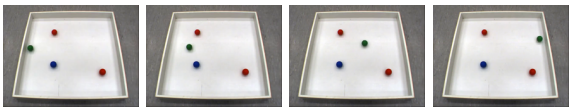


FIGURE C.56: Snapshots of the video 4.09a: a green object moving rightward; two static red objects and a static blue object are also in the scene.

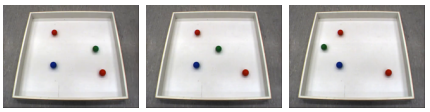


FIGURE C.57: Snapshots of the video 4.09b: a green object moving leftward; two static red objects and a static blue object are also in the scene.

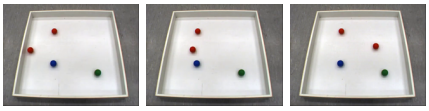


FIGURE C.58: Snapshots of the video 4.10a: a red object moving rightward; three other static objects (red, blue, green) are in the scene.

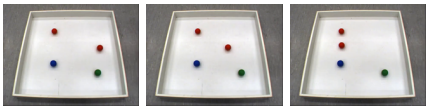


FIGURE C.59: Snapshots of the video 4.10b: a red object moving leftward; three other static objects (red, blue, green) are in the scene.

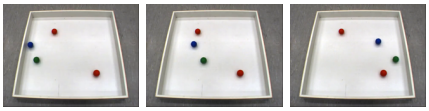


FIGURE C.60: Snapshots of the video 4.11a: a blue and a green object moving rightward; two red static objects are in the scene.

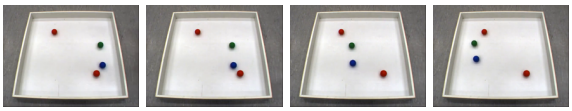


FIGURE C.61: Snapshots of the video 4.11b: a blue and a green object moving leftward; two red static objects are in the scene.

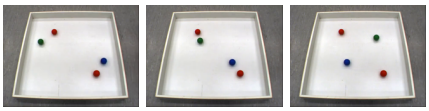


FIGURE C.62: Snapshots of the video 4.11c: a blue object moving leftward; a green object moving rightward; two red static objects are in the scene.

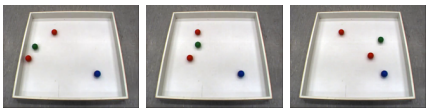


FIGURE C.63: Snapshots of the video 4.12a: a red and a green object moving rightward; two other static objects (red, blue) are in the scene.

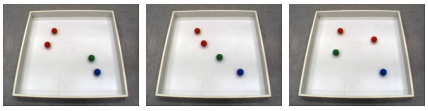


FIGURE C.64: Snapshots of the video 4.12b: a red object moving rightward; a green object moving leftward; two other static objects (red, blue) are in the scene.

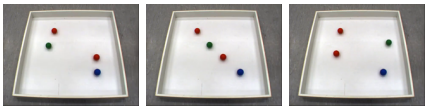


FIGURE C.65: Snapshots of the video 4.12c: a red object moving leftward; a green object moving rightward; two other static objects (red, blue) are in the scene.

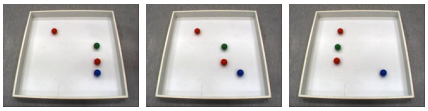


FIGURE C.66: Snapshots of the video 4.12d: a red and a green object moving leftward; two other static objects (red, blue) are in the scene.

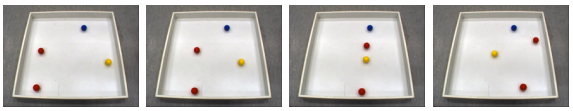


FIGURE C.67: Snapshots of the video 4.14e: two red objects moving rightward; a yellow object moving leftward; a blue static object is in the scene.

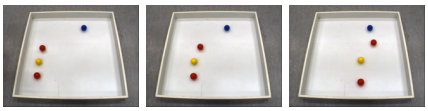


FIGURE C.68: Snapshots of the video 4.14f: two red objects and a yellow object moving rightward; a blue static object is in the scene.

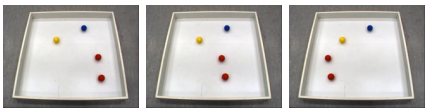


FIGURE C.69: Snapshots of the video 4.15a: two red objects moving leftward; two other static objects (blue, yellow) are in the scene.

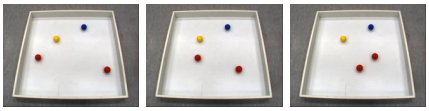


FIGURE C.70: Snapshots of the video 4.15b: a red object moving leftward, another red object moving rightward; two other static objects (blue, yellow) are in the scene.

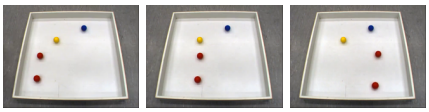


FIGURE C.71: Snapshots of the video 4.15c: two red objects moving rightward; two other static objects (blue, yellow) are in the scene.

FIGURE C.72: Snapshots of the video 4.16a: a green and a blue object moving leftward; two other static objects (red, yellow) are in the scene.

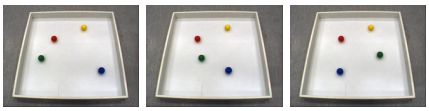


FIGURE C.73: Snapshots of the video 4.16b: a green object moving rightward; a blue object moving leftward; two other static objects (red, yellow) are in the scene.

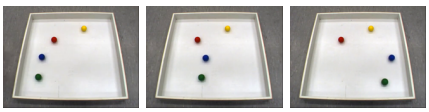


FIGURE C.74: Snapshots of the video 4.16c: a green and a blue object moving rightward; two other static objects (red, yellow) are in the scene.

FIGURE C.75: Snapshot of the video 4.18a: four static objects (two red, two green); all red objects are to the right of all green objects.

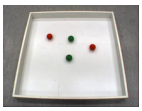


FIGURE C.76: Snapshot of the video 4.18b: four static objects (two red, two green); a red object is to the left of a green object; the other red object is to the right of both green objects.

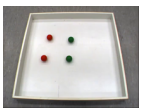


FIGURE C.77: Snapshot of the video 4.18c: four static objects (two red, two green); all red objects are to the left of all green objects.

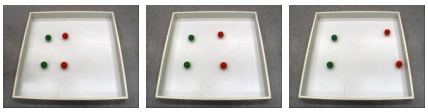


FIGURE C.78: Snapshot of the video 4.19a: two red objects moving away from two static green objects, respectively.

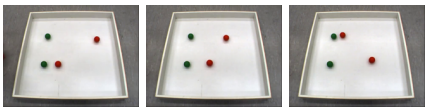


FIGURE C.79: Snapshot of the video 4.19b: a red object moving toward a static green object; another red object moving away from another static green object.

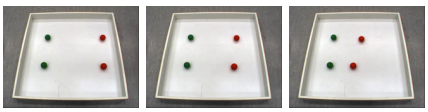


FIGURE C.80: Snapshot of the video 4.19c: two red objects moving toward two static green objects, respectively.

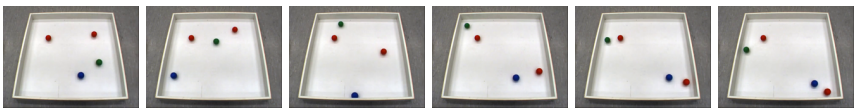


FIGURE C.81: Snapshots (2 s apart) of the video 4.21: three moving objects (red, green, blue); one static red object.

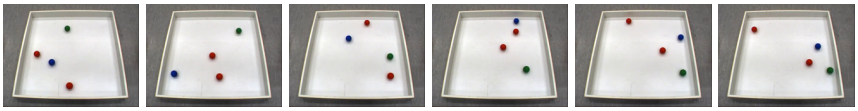


FIGURE C.82: Snapshots (2 s apart) of the video 4.23: four moving objects (two red, one green, one blue).

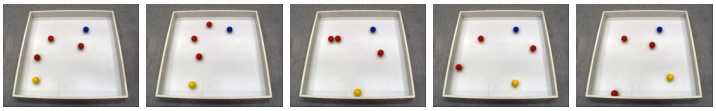


FIGURE C.83: Snapshots of the video 5.00: three moving objects (two red, one yellow); two static objects (red, blue).

Bibliography

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87.

Ascher, U. M. & Petzold, L. R. (1998). *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–767.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–609, 577–609.

Bergen, B. K. & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In J.-O. Östman & M. Fried (Eds.), *Construction Grammars: Cognitive Grounding and Theoretical Extensions* (Chap. 6, pp. 147–190). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Berger, M., Faubel, C., Norman, J., Hock, H., & Schöner, G. (2012). The counter-change model of motion perception: An account based on dynamic field theory. In A. E. P. Villa (Ed.), *ICANN 2012, Part I, LNCS 7552* (pp. 579–586). Berlin Heidelberg: Springer.

Cangelosi, A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, *7*(2), 139–151.

Cangelosi, A. & Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, *4*(1).

Cangelosi, A. & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, *30*(4), 673–689.

Carlson, L. A. & Hill, P. L. (2008). Processing the presence, placement, and properties of a distractor in spatial language tasks. *Memory & Cognition*, *36*(2), 240–255.

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, *3*(9), 345–351.

Dennett, D. C. & Viger, C. D. (1999). Sort-of symbols? [Peer commentary on "Perceptual symbol systems" by Lawrence W. Barsalou]. *Behavioral and Brain Sciences*, *22*(4), 613.

Dominey, P. F. (2007). Towards a construction-based framework for development of language, event perception and social cognition: Insights from grounded robotics and simulation. *Neurocomputing*, *70*(13-15), 2288–2302.

Dominey, P. F. & Boucher, J. D. (2005a). Developmental stages of perception and language acquisition in a perceptually grounded robot. *Cognitive Systems Research*, *6*(3), 243–259.

Dominey, P. F. & Boucher, J. D. (2005b). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, *167*(1-2), 31–61.

Doumas, L. A. A. & Hummel, J. E. (2005). A symbolic-connectionist model of relation discovery. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 606–611). Austin, TX: Cognitive Science Society.

Doumas, L. A. A. & Hummel, J. E. (2012). Computational models of higher cognition. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (Chap. 5, pp. 52–66). Oxford University Press.

Durán, B., Sandamirskaya, Y., & Schöner, G. (2012). A dynamic field architecture for the generation of hierarchically organized sequences. In A. E. P. Villa, W. Duch, P. Érdi, F. Masulli, & G. Palm (Eds.), *Artificial neural networks and machine learning - icann 2012* (pp. 25–32). Berlin, Heidelberg: Springer.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227.

Gallese, V. & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3), 455–479.

Georgopolous, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal Population Coding of Movement Direction. *Science*, *233*(March), 1416–1419.

Gibbs, R. W. & Colston, H. L. (1995). The cognitive psychological reality of image schemas and their transformations. *Cognitive Linguistics*, *6*(4), 347–378.

Glenberg, A. M. & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*(3), 558–565.

Goldberg, A. (1992). The inherent semantics of argument structure: The case of English ditransitive construction. *Cognitive Linguistics*, *3*(1), 37–74.

Gorniak, P. & Roy, D. (2004). Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, *21*, 429–470.

Guerra-Filho, G. & Aloimonos, Y. (2012). The syntax of human actions and interactions. *Journal of Neurolinguistics*, *25*(5), 500–514.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001, June). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, *84*(6), 401–10.

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, *14*(11), 497–505.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

Henson, R. N. A. & Burgess, N. (1997). Representations of serial order. In J. A. Bullinaria, D. W. Glasspool, & G. Houghton (Eds.), *4th Neural Computation and Psychology Workshop, London 9-11 April 1997: Connectionist Representations* (pp. 283–300). London, UK: Springer.

Hock, H. S., Schöner, G., & Gilroy, L. (2009). A counterchange mechanism for the perception of motion. *Acta Psychologica*, *132*(1), 1–21.

Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, *23*(2), 109–118.

Hummel, J. E. & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–264.

Hummel, J. E. & Holyoak, K. J. (2005). Relational reasoning in a neurally plausible cognitive architecture. An overview of the LISA project. *Current Directions in Psychological Science*, *14*(3), 153–157.

Itti, L. & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. New York: Oxford University Press.

Jancke, D., Erlhagen, W., Dinse, H. R., Akhavan, A. C., Giese, M., Steinhage, A., & Schöner, G. (1999). Parametric population representation of retinal location: neuronal interaction dynamics in cat primary visual cortex. *Journal of Neuroscience*, *19*(20), 9016–9028.

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2015). Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3668–3678). IEEE.

Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago: University of Chicago Press.

Karpathy, A. & Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 664–676.

Kaschak, M. P. & Jones, J. L. (2014). Grounding language in our bodies and the world. In Thomas Holtgraves (Ed.), *The Oxford Handbook of Language and Social Psychology* (Chap. 20, pp. 317–329). London: Oxford University Press.

Knips, G., Zibner, S. K. U., Reimann, H., Popova, I., & Schöner, G. (2017). A neural dynamics architecture for grasping that integrates perception and movement generation and enables on-line updating. *Frontiers in Neurorobotics*, *11*(9).

Lallee, S. & Dominey, P. F. (2013). Multi-modal convergence maps: From body schema and self-representation to mental imagery. *Adaptive Behavior*, *21*(4), 274–285.

Langacker, R. W. (1986). An introduction to cognitive grammar. *Cognitive Science*, *10*, 1–40.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141–160.

Lins, J. & Schöner, G. (2017). Mouse tracking shows attraction to alternative targets while grounding spatial relations. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 2586–2591). Austin, TX: Cognitive Science Society.

Lipinski, J., Sandamirskaya, Y., & Schöner, G. (2009). Swing it to the left, swing it to the right: Enacting flexible spatial language using a neurodynamic framework. *Cognitive Neurodynamics*, *3*(4), 373–400.

Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J. P., & Schöner, G. (2012). A neurobehavioral model of flexible spatial language behaviors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1490–1511.

Lobato, D., Sandamirskaya, Y., Richter, M., & Schöner, G. (2015). Parsing of action sequences: A neural dynamics approach. *Paladyn, Journal of Behavioral Robotics*, *6*, 119–135.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1015–1036.

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13, pp. 493–529). Cambridge, MA, USA: MIT Press.

Lomp, O., Faubel, C., & Schöner, G. (2017). A neural-dynamic architecture for concurrent estimation of object pose and identity. *Frontiers in Neurorobotics*, *11*(April), 1–17.

Lomp, O., Richter, M., Zibner, S. K. U., & Schöner, G. (2016). Developing dynamic field theory architectures for embodied cognitive systems with cedar. *Frontiers in Neurorobotics*, *10*, 1–18.

Lu, C., Krishna, R., Bernstein, M., & Fei-Fei, L. (2016). Visual relationship detection with language priors. In *European Conference on Computer Vision* (pp. 852–869). Springer International Publishing.

Luciw, M., Kazerounian, S., Lahkman, K., Richter, M., & Sandamirskaya, Y. (2015). Learning the condition of satisfaction of an elementary behavior in dynamic field theory. *Paladyn, Journal of Behavioral Robotics*, *6*(1), 180–190.

Madden, C., Hoen, M., & Dominey, P. F. (2010). A cognitive neuroscience perspective on embodied language for human-robot cooperation. *Brain and Language*, *112*(3), 180–188.

Mavridis, N. & Roy, D. (2006). Grounded situation models for robots: Where words and percepts meet. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on* (pp. 4690–4697). IEEE.

Pastra, K. & Aloimonos, Y. (2012). The minimalist grammar of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1585), 103–117.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(July), 576–582.

Ragni, M. & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, *89*(4), 1009–1023.

Regier, T. (1992). *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization* (tech. rep. No. TR-92-062). International Computer Science Institute. Berkeley.

Regier, T. (1995). A model of the human capacity for categorizing spatial relations. *Cognitive Linguistics*, *6*(1), 63–88.

Regier, T. & Carlson, L. A. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, *130*(2), 273–298.

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2847–2852). Austin, TX: Cognitive Science Society.

Richter, M., Lins, J., Schneegans, S., & Schöner, G. (2014). A neural dynamic architecture resolves phrases about spatial relations in visual scenes. In S. Wermter (Ed.), *Artificial Neural Networks and Machine Learning: ICANN 2014, 24th International Conference on Artificial Neural Networks, Lecture Notes in Computer Science 8681* (pp. 201–208).

Richter, M., Lins, J., & Schöner, G. (2016). A neural dynamic model parses object-oriented actions. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1931–1936). Austin, TX: Cognitive Science Society.

Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Science*, *9*(1), 35–47.

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2457–2464). New York, NY: Institute of Electrical and Electronics Engineers (IEEE).

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–25.

Roy, D. (2005a). Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciences*, *9*(8), 389–396.

Roy, D. (2005b). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, *167*(1-2), 170–205.

Roy, D. (2008). A mechanistic model of three facets of meaning. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and Embodiment: Debates on Meaning and Cognition* (pp. 1–32). Oxford, UK: Oxford University Press.

Sandamirskaya, Y., Richter, M., & Schöner, G. (2011). A neural-dynamic architecture for behavioral organization of an embodied agent. In *IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL EPIROB 2011)* (pp. 1–7). IEEE.

Sandamirskaya, Y. & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179.

Schneegans, S., Lins, J., & Spencer, J. P. (2015). Integration and selection in multidimensional neural fields. In G. Schöner & J. P. Spencer (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (Chap. 5, pp. 121–150). New York: Oxford University Press.

Schneegans, S. & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109.

Schneegans, S., Spencer, J. P., & Schöner, G. (2015). Integrating "what" and "where": Visual working memory for objects in a scene. In G. Schöner & J. P. Spencer (Eds.), *Dynamic Thinking: A Primer on Dynamic Field Theory* (Chap. 8, pp. 197–226). New York: Oxford University Press.

Schöner, G., Spencer, J. P., & the DFT Research Group. (2015). *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York: Oxford University Press.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–457.

Shastri, L. (1999). Advances in SHRUTI: A neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence*, *11*, 79–108.

Shastri, L., Grannes, D., Narayanan, S. S., & Feldman, J. (2002). *A Connectionist Encoding of Parameterized Schemas and Reactive Plans* (tech. rep. No. TR-02-008). International Computer Science Institute. Berkeley.

Spencer, J. P. & Schöner, G. (2003). Bridging the representational gap in the dynamic systems approach to development. *Developmental Science*, *6*(4), 392–412.

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7*(7), 308–312.

Steels, L. (2008). The symbol grounding problem has been solved. So what's next? In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and Embodiment: Debates on Meaning and Cognition* (pp. 223–244). New York: Oxford University Press.

Steels, L. & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*(4), 469–489, 469–489.

Steels, L. & Kaplan, F. (2002). AIBO's first words. The social learning of language and meaning. *Evolution of Communication*, *4*(1), 3–32.

Stramandinoli, F., Marocco, D., & Cangelosi, A. (2012). The grounding of higher order concepts in action and language: A cognitive robotics model. *Neural Networks*, *32*, 165–173.

Talmy, L. (1988). The relation of grammar to cognition. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics* (pp. 165–205). Amsterdam/Philadelphia: John Benjamins.

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

van der Velde, F. & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*(1), 37–108.

van Hengel, U., Sandamirskaya, Y., Schneegans, S., & Schöner, G. (2012). A neural-dynamic architecture for flexible spatial language: Intrinsic frames, the term "between", and autonomy. In *Robot and Human Interactive Communication, 2012 IEEE RO-MAN: The 21st IEEE International Symposium on* (pp. 150–157). IEEE.

Wilson, H. R. & Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Biological Cybernetics*, *13*(2), 55–80.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–36.

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99–119). New York: Oxford University Press.

Acronyms

AVS attentional vector sum. 144

CNN convolutional neural network. 136, 143

CoD condition of dissatisfaction. xv, 42, 58–60, 79, 82, 113, 115, 119–121, 145, 146, 152, 176, 194

CoS condition of satisfaction. xv, 34, 42, 48, 56, 58–62, 64, 66, 68–75, 78, 79, 81, 82, 84, 87, 90, 93, 95, 96, 100, 110, 112, 113, 115–121, 123, 124, 128–132, 144–146, 151–153, 156, 157, 165–173, 175–194

DFT dynamic field theory. vii, 2, 3, 5, 7, 9, 11, 13, 18–21, 24, 26, 27, 31, 34, 37–39, 135, 137, 139, 143–147, 149, 151, 153–155, 159, 160

DPA distribution of population activation. 19

EB elementary behavior. 34–37, 71, 82, 151

GPU graphics processing unit. 156

HSV hue, saturation, value. 43, 165, 223

IOR inhibition-of-return. 42, 51, 55, 56, 78, 90, 93, 94, 100, 102, 113–115, 119, 121, 125, 129, 148, 152, 170, 181

RGB red-green-blue. 138

SPAUN Semantic Pointer Architecture Unified Network. 17, 138, 146

Glossary

u activation of a field or node. 21–31, 35, 36, 38, 42–49, 51, 52, 54–57, 59–63, 65–69, 71–74, 166–194, 223

c hue feature dimension of the hue, saturation, value (HSV) color space. 42–48, 51, 52, 54–56, 66, 67, 69, 88–90, 98, 99, 123, 124, 175, 180, 223

W static weights between a neural node and a neural field that encode the perceptual meaning of a concept (e.g., the color concept RED). 47, 48, 59, 60, 66–70, 223

s external input into a dynamic neural field or node. 21, 23–31, 35–37, 43–45, 61, 65, 67–69, 71–74, 166–169, 223

k kernel; patterned weight function that determines the interaction between different points within a field (lateral interaction) or between different fields. 21, 22, 24, 27, 29, 30, 36, 37, 43–49, 51, 52, 54–57, 59–63, 71, 223

ϕ feature dimension of the direction of motion. 28, 42, 43, 45–49, 51, 52, 54–56, 58–62, 68–71, 92–94, 98, 100, 123, 124, 180, 223

ξ noise. 21, 24, 38, 43–49, 51, 52, 54–57, 59–62, 65–69, 71, 72, 163, 166–194, 223

h (negative) resting level of a field or node. 21, 23, 24, 29, 30, 35, 36, 43–49, 51, 52, 54–57, 59–62, 65–69, 71, 72, 164–194, 223

r feature dimension of scale. 43, 45, 58–62, 70, 223

g sigmoid function; non-linear function that determines the output of all dynamic neural fields and dynamic neural nodes; a common choice is the logistic function (Equation 2.5); see Figure 2.5 for an exemplary plot. 21–27, 29–31, 35, 36, 43–49, 51, 52, 54–57, 59–63, 65–69, 71, 72, 74, 166–194, 223

x feature space; in the description of the model (Section 3), x refers to the horizontal spatial dimension of the camera image. 21–31, 35, 36, 42–60, 62, 63, 69–71, 88–90, 92–94, 98–100, 123, 165, 177, 179, 181, 184, 185, 187, 189, 194, 224

y feature space; in the description of the model (Section 3), y refers to the vertical spatial dimension of the camera image. 26, 29, 30, 42–60, 62, 63, 69–71, 89, 90, 92–94, 98–100, 123, 165, 177, 179, 181, 184, 185, 187, 189, 194, 224

t (continuous) time. 21, 23–31, 35, 36, 38, 42–49, 51, 52, 54–57, 59–63, 65–69, 71–74, 166–194, 224

τ time scale of dynamics. 21, 24, 28–30, 35, 36, 38, 43–49, 51, 52, 54–57, 59–62, 65–69, 71, 72, 93, 163, 164, 166–194, 224

w weight; usually a constant scalar. 21, 22, 24–28, 35, 36, 38, 43–49, 51, 52, 54–57, 59–62, 65–72, 74, 163–194, 224