Proceedings of the 40th Annual Conference of the Cognitive Science Society (2018)

Sequences of discrete attentional shifts emerge from a neural dynamic architecture for conjunctive visual search that operates in continuous time

Raul Grieben[1], Jan Tekülve[1], Stephan K.U. Zibner[1], Sebastian Schneegans[2] and Gregor Schöner[1]

[1]Institut für Neuroinformatik, Ruhr-Universität Bochum
44780 Bochum, Germany
{name.surname}@ini.rub.de

[2]Department of Psychology, University of Cambridge
Cambridge CB2 3EB, United Kingdom
sebastian@schneegans.de

Abstract

The goal of conjunctive visual search is to attentionally select a location at which the visual array matches a set of cued feature values. Here we present a neural dynamic architecture in which all neural processes operate in parallel in continuous time, but in which discrete sequences of processing steps emerge from dynamic instabilities. When biased competition selects an object location at which not all conjunctive feature values match the cue, the neural representation of a condition of dissatisfaction is activated and induces an attentional shift. Successful match activates the neural representation of a condition of satisfaction that ends the search. The search takes place in the current visual array but takes into account an autonomously acquired feature-space scene memory.

Keywords: neural dynamic architecture; visual search; binding; visual working memory;

Introduction

Bringing an object into the attentional foreground **through visual search is a critical first step in almost all actions that are directed at the outer world (Tatler & Land, 2016).** The search may be based on a set of visual features that we are familiar with (for example, search for a large, blueish, vertically aligned shape when looking for a bottle of Skyy vodka). A large literature in visual cognition addresses many aspects of visual search (Wolfe, 2015). Since Anne Treisman's seminal work on feature integration theory (Treisman, 1998), the organization of visual search guided by the combination of multiple feature dimensions (or conjunctions) into a parallel and a serial stream has been a dominant theme. How the time needed to find a searched item scales with the number of distractors, but also with the metric differences between targets and distractors, is used to diagnose the organization of the underlying processes.

In its most recent variant, the guided search hypothesis accounts for a wealth of data by postulating that an early parallel stage of search is followed by a serial examination of candidate items (Wolfe, 2007). At the core of this theory is an information processing algorithm that starts diffusions races for each examined item to determine the match to the search criteria. Competitive guided search (Moran, Zehetleitner, Muller, & Usher, 2013) puts a stronger emphasis on neural mechanisms by introducing inhibition into the selection processes, but retains the information processing core.

An alternative is attentional engagement theory which recognizes that metric differences among distractors and between targets and distractors matter (Duncan.J & Humphrey, 1989). This account has been implemented in a connectionist architecture (Humphreys & Müller, 1993), in which inhibitory and excitatory coupling among feature encoding units leads to grouping effects that explain how search for feature conjunctions can occur pre-attentively (Humphreys, 2016). The model contains, however, elements of information processing not grounded in neural terms. Neurally mechanistic accounts for visual search (Deco & Rolls, 2004) are far from capturing the behavioral features of conjunctive search and are thus difficult to compare. **A probabilistic graphical model of visual attention (Chikkerur, Serre, Tan, & Poggio, 2010) provides an integrated formal account for feature binding in terms of probabilistic inference. The deployment of spatial attention to specific locations remains outside the Baysian framework, however.**

Our goal is to provide a neural processing account for feature binding through space that autonomously organizes visual search as a sequence of neural operations. We build on earlier work (Schneegans, Spencer, & Schöner, 2015) within Dynamic Field Theory (Schöner, Spencer, & DFT Research Group, 2015), a theoretical framework for understanding cognition grounded in neural population activity. In the model, neural activation patterns evolve continuously in time. Decisions emerge from dynamic instabilities, in which peaks of activations arise. Sequences of such events emerge from the interactions within the neural architecture. Thus, the neural dynamics fundamentally evolves in parallel across the entire architecture, but sequential processing steps emerge under appropriate conditions. Items are not a concept in this framework (Hulleman & Olivers, 2015), but dependencies on the size of activated regions may reflect similar dependencies.

Earlier we showed that this model can account for classical signatures of binding through space in change detection paradigms (Schneegans et al., 2015). Here we show that the theory can generate conjunctive visual searches. In this brief paper we only demonstrate the emergence of the processes of searching for a target object in the presence of distractor items. The particular scenario involves looking at a visual scene to which a target object is added at some point. The task is to find a visually matching object (Malcolm & Henderson, 2009). All processing steps emerge from the time-continuous dynamics of the neural architecture.

Dynamic Field Theory

Dynamic Field Theory (DFT) (Schöner et al., 2015) is a theoretical framework for understanding perception, motor be-

havior, and cognition based on neural principles. The activity in neural populations is modeled by activation fields, $u(x,t)$, whose metric dimensions, x,are defined by the input or output connectivity of the fields to sensory or motor surfaces. The neural dynamics of the activation fields,

$$\tau\dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int \omega(x-x')\sigma(u(x',t))dx$$

generates the time-continuous evolution of neural activation patterns on the time scale τ. Two classes of stable solutions exist (Amari, 1977). Activation patterns that remain below the threshold of a sigmoidal nonlinearity, $\sigma(u) = 1/(1+\exp[-\beta u])$, are stable as long as localized inputs, $s(x,t)$, remain weak relative to the resting level, $h < 0$, so that intra-field interaction is not significantly engaged. Such sub-threshold activation patterns may still be structured along the field's dimension, x. Supra-threshold peaks of activation are self-stabilized by the neural interaction, whose kernel, $\omega(x-x')$ is locally excitatory and inhibitory over longer distances, $x-x'$ (see figure 1). Self-stabilized activation peaks

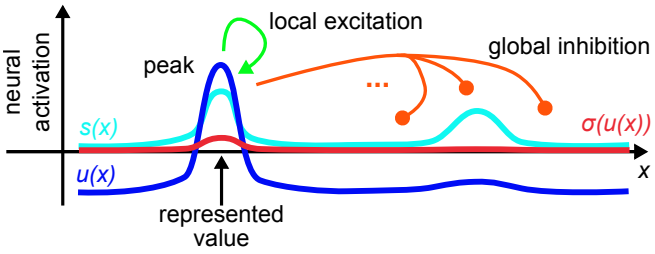

Figure 1: Dynamic neural field

are the units of representation in DFT. Their stability enables continuous online coupling to time-varying sensory and motor signals.

Peaks emerge when the sub-threshold solution becomes unstable in the *detection instability*. Time-continuous, graded changes of input may thus induce events at discrete moments in time at which neural interaction engages. This is how sequences of neural activation patterns emerge from the underlying time-continuous neural dynamics in DFT. Peaks become unstable in the *reverse detection instability* which occurs at lower levels of input activation, leading to the hysteretic stabilization of detection decisions.

Fields may be *selective* when inhibitory interaction allows only a single peak to form within a field, or they may support multiple self-stabilized peaks. Peaks may be sustained once localized input is removed, forming the basis for working memory. Multi-dimensional fields may represent conjunctions of feature dimensions, for example, the conjunction of color and space. Zero-dimensional fields are dynamic neural nodes (DNN), that represent categorical states.

Field Architectures

The stability of the two classes of activation patterns makes it possible to couple fields while retaining their dynamic properties. The dynamics of the resulting neural architectures may thus still be understood in terms of the dynamic instabilities in each component field.

The coupling among activation field is may be structured by connection kernels that weigh the output of one field as it provides input to any location of the receiving field. Such projections may preserve the dimensionality of the fields, or may expand or contract the field dimensionality (Zibner & Faubel, 2016). *Dimensionality expansions* may take the form of ridges (or tubes, or slices), in which input along one or several of the receiving field's dimension is constant. *Dimensionality contractions* typically entail integrating along a contracted dimension or a relevant subspace.

Peak detectors are a limit case of contractions in which a dynamic neural node receives input from the integrated output of an entire field so that the node becomes activated only if at least one supra-threshold peak of activation is present in its source field. Dynamic neural nodes that project onto a field by expansion are called *boost nodes*. They may induce detection instabilities or their reverse in the target field. Within architectures, such boost nodes may effectively modulate the flow of activation by enabling or disabling particular branches of an architecture to create units of representation.

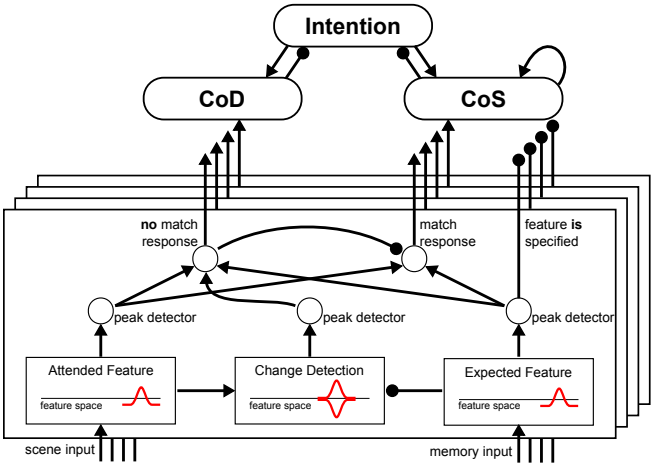

Figure 2: The Change Detection Module connected to an *Elementary Control Unit* (ECU). Feature values in the visual array and the scene memory are compared in parallel along each feature dimension. The change detection field and connected peak detector nodes signals a mismatch if the attended, the expected, and the change detector fields all carry a peak. A match is signaled if both the attended and the expected field carry peaks, but the change detector field does not. A single mismatch is sufficient to activate the *CoD*. The *CoS* is activated only when a match is detected along each of the specified dimensions.

Match and change detection The representational content of two fields may be compared by projecting onto a third field or node. In *match detection*, a peak in the third field is only formed when the peak locations in the two source fields overlap sufficiently. *Change detection* is the opposite relationship,

in which a peak is formed in the third field only when the peak locations in the source fields differ sufficiently (Johnson, Spencer, Luck, & Schöner, 2009). This is based on a combination of inhibitory and excitatory input as illustrated in the bottom three fields of Figure 2.

Process organization In dynamic field architectures, sequences of activation states are organized through substructures of dynamic fields and dynamic nodes illustrated in Figure 2. Each such *Elementary Control Unit* (ECU) (Richter, Sandamirskaya, & Schöner, 2012) represents a particular processing step by an *intention node* that activates a subset of the architecture. Its intended outcome is detected by a peak detector that represents the *Condition of Satisfaction* (CoS). Failure to achieve this outcome may lead to activating a node that represents the *Condition of Dissatisfaction* (CoD) (for example, via a change detector that picks up the mis-match between the intended and the achieved representational state). Typically, both CoS and CoD nodes inhibit the intention node. CoS nodes are often self-excitatory so that they remain activated and prevent reactivation of the intention node, while CoD nodes are not self-excitatory, allowing for renewed attempts at the same process. Different such ECUs may be coupled to organize more complex tasks.

Neural dynamic architecture

The dynamic neural architecture illustrated in Figure 3 is capable of autonomously *exploring* the visual array by attentionally selecting locations, *memorizing* feature values associated with those locations, and *visually searching* for cued feature conjunctions. The neural dynamics switches between these three functional modes as the neural nodes organized into three ECUs shown on the right become active. By boosting a portion of the architecture, each node enables relevant activation fields to form peaks (illustrated by matching colors in Figure 3). The switches among the three functional modes are mediated by the change detection module described in Figure 2 that signals when matching feature values are written in the scene memory (*memorize*,*explore*) or when locations have been found at which visual features match the cue (*visual search*). The architecture is implemented in *cedar*, a software framework for neural dynamic systems that enables real-time numerical simulation (Lomp, Richter, Zibner, & Schöner, 2016).

Visual input is obtained from a camera that points down onto a table surface, on which colored rectangles of varying orientation, width and length are placed. Figure 4 illustrates the four feature channels. Color is extracted by transforming RGB values into hue-space. Orientation is obtained from four elongated center-surround filters that are applied to the thresholded saturation of the camera stream. Width and length are extracted using a pyramid of center-surround filters of increasing size and implementing a one-way inhibition along this dimension.

Exploration is the process of sequentially attending to salient locations in the visual array. This process is driven by a "dorsal" channel, the *Spatial Salience* map in retinal coordinates that receives weighted input from the *Space/Feature (S/F) Bottom-Up* feature channels. The Spatial Salience map projects through a semi-linear threshold function into the *Spatial Salience Attention* field that selects a single location

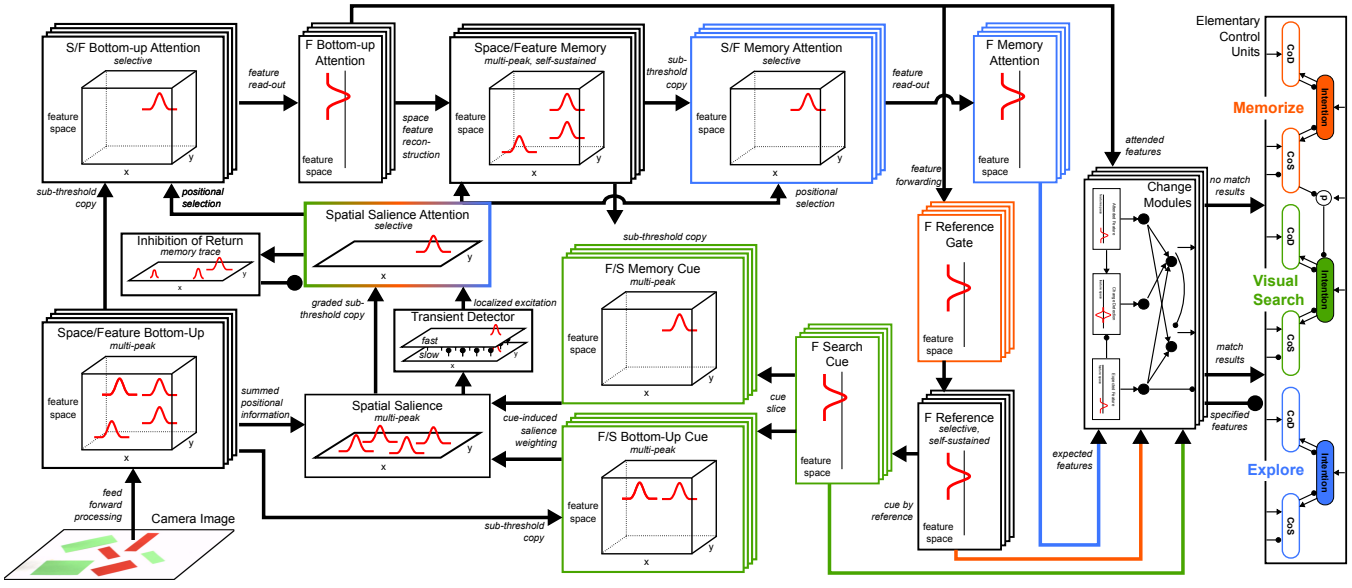

Figure 3: Scene representation architecture. Each field stack represents four fields, one for each feature dimension: color, orientation, width, and length. Connections between different field stacks couple the corresponding feature fields. Fields and nodes highlighted in color receive a boost when the intention-node highlighted in the same color is activated. They are below the activation threshold when that node is not activated.

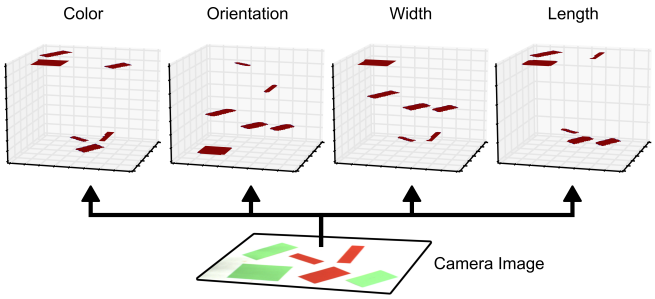

Figure 4: The camera image and the four extracted 3D maps (2D visual space vs. 1D feature dimension plotted along the vertical axis).

by forming a peak at the most salient position. That location is selected in the *S/F Bottom-up Attention* field and the feature values at that location are then extracted in the *F Bottom-up Attention field*.

This separation of feature and location is critical when the coordinate transform from the retinal to a fixed, scene-centered frame is taken into account. That coordinate transform needs to operate only on the output of the Spatial Salience Attention field, which is then recombined with the extracted Bottom-up Feature attention to form the *S/F Memory* defined over allo-centric space.

Match between the feature value in the bottom-up and the memory attention fields signals the condition of satisfaction of the explore node, which inhibits the associated intention node leading to a de-boosting of the Spatial Salience Attention and a destabilization of all peaks in all attention fields, while peaks in the *S/F Memory* field is sustained. The explore may be reactivated if it is not inhibited by the other modes, leading to renewed attentional selection of a salient location. The *inhibition of return* memory trace associated with the Spatial Salience Attention field steers attentional selection towards locations that have not recently been selected.

Attentional selection is also influenced by a two-layer transient detector that excites spatial locations, which are subject to rapid change in saliency, potentially overriding the current focus of attention ((see (Berger, Faubel, Norman, Hock, & Schöner, 2012) for details)).

Memorizing involves representing feature values in a reference field to use, in this architecture, as a cue to visual search. By boosting the *F Reference Gate*, the *memorize* intention node forwards the currently attended feature values to the *F Reference Memory* field, inducing formation of sustained activation peaks. The match between the feature values in the F Reference Memory with the feature values in the *F Memory Attention* field activates the condition of satisfaction of the memorize node. This releases the intention node of the visual search mode from inhibition.

Visual Search brings those locations into the attentional foreground at which feature values matching the cue are detected. Boosting the *F Search Cue* field, the *visual search*

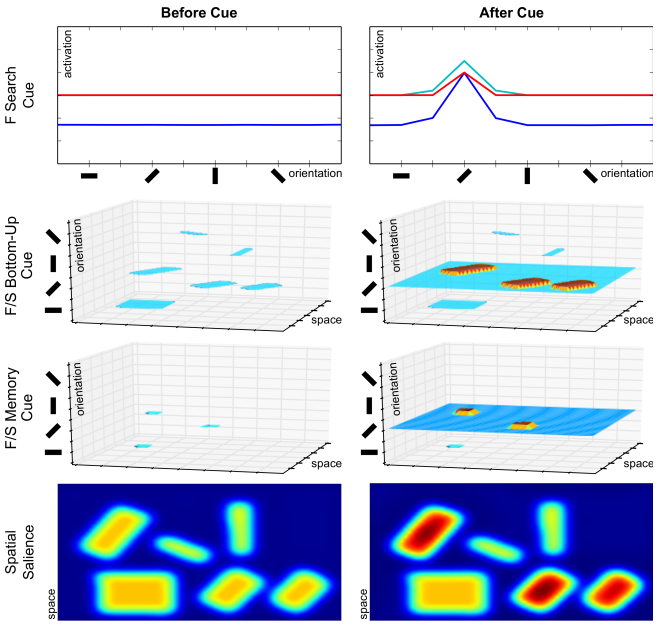

Figure 5: Influence of a cued feature value (top-right) on the *Spatial Salience* field (bottom). The three locations matching the orientation cue are boosted in salience. Two of these locations are already in visual memory (third row, right column). Their salience is slightly elevated.

intention node enables the induction of a peak representing the cued feature values based on input from the F Reference field.

The cued feature provides slice input to the *F/S Memory Cue* and the *Bottom-up Cue* fields. As a result, the *F/S Memory Cue* represents locations at which scene memory represents matching feature values. The *F/S Bottom-Up Cue* represents locations at which the current visual array signals matching feature values. The two fields provide top-down input to the Spatial Salience field that biases attentional selection toward locations matching the cue (figure 5).

Once a location has been selected, the associated feature values are extracted along the bottom-up path and matched to the search cue. A match activates the condition of satisfaction of visual search, a mismatch activates the condition of dissatisfaction. The threshold for match detection is tuned to signal match only when all specified feature dimensions contribute and it is that conjunction of match conditions that effectively "binds" the feature values in visual search.

Visual Search using a Reference Object

Figure 6 demonstrates how sequences of neural events emerge from the time-continuous neural dynamics as the two functional modes of memorize and visual search are combined to perform a conjunctive search based on feature values extracted from a cued visual location. In this demonstration we add a new object to a visual scene that has been previously explored and committed to scene memory. The changed location is detected in a field of transient detectors that provides

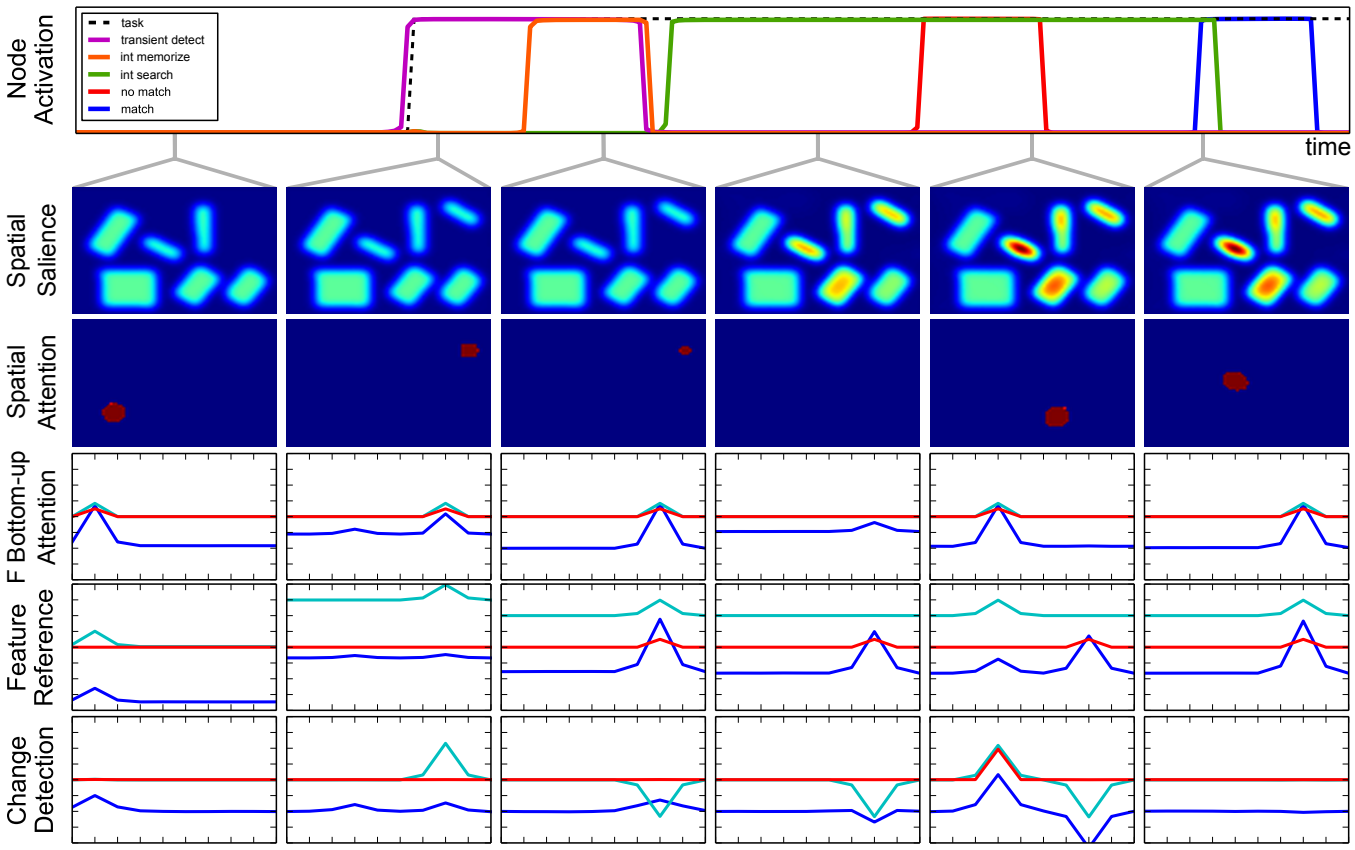

Figure 6: Time course of a visual search based on a reference object. The top row shows time courses of activation of relevant dynamic neural nodes. Activation snapshots at selected points in time (indicated by grey lines) are shown in the next two rows. The *Spatial Salience* and the *Spatial Salience Attention* field are shown across retinal space. The thresholded activation level is color coded (blue indicates low, red indicates high levels). In the three bottom rows, 1D-fields across orientation are illustrated (input in cyan, activation in blue, thresholded activation in red).

localized excitatory input to the Spatial Salience Attention field. This transient detector serves as an activation trigger for the overall task activating the *memorize* and *visual search* processing modes. The time courses of relevant dynamical neural node activation levels as well as activation snapshots of the relevant activation fields are depicted in Figure 6.

At the beginning of the demonstration (first column of Figure 6), the architecture is scanning the visual scene in explore mode with one object in the attentional foreground while other objects are already represented by activation peaks in the S/F Memory fields. Once the new object is added to the visual scene (second column), it is attentionally selected due to localized input from the array of transient detectors, overwriting the previously attended location. Feature values at this location are extracted analogously as in the explore behavior.

The active memorize mode helps forward this feature vector to the F reference field (third column). A matching peak in that field terminates the memorize mode by activating its CoS, which disinhibits the visual search mode and starts a cued visual search.

Top-down influence along the four feature dimensions biases the *Spatial Salience* field (analogously to figure 5). Note that the new object's position is inhibited by the inhibition of return memory trace (not depicted). Peak generation along different feature dimensions may take different amounts of time as a function of stimulus metrics, so that spatial selection in the *Spatial Attention* field may lead to a less than complete feature overlap with the cue. Here, the orientation *change detection* field forms a peak, sufficient to activate the *CoD* node (fifth column). That change response deboosts the *Spatial Salience Attention* field and enables the attentional selection of a second item, which, in this case, fits the cue along all feature dimensions. This activates the match mode and then the CoS node of the visual search mode. That concludes the visual search with the sought location and the associated feature values in the attentional foreground (sixth column).

Conclusion

We have shown how a sequence of activation patterns emerges in a neural dynamic architecture that represents four different feature-space conjunctions and is linked to on-line

visual input. The organization of that sequence enables the system to autonomously build a representation of the visual scene and to detect changes to the scene. Defining an added visual element as the target object, we showed how the architecture autonomously searches the scene for matching elements, both within its internal scene representation as well as based on the current visual input. That search is fundamentally based on parallel activation dynamics, but sequential examination of candidate regions of the visual array emerges. **Autonomously generating the entire sequence of processes required to find a cued object is the key innovation here, compared to related work within the same theoretical framework of neural dynamics (Fix, Rougier, & Alexandre, 2011) and within a Baysian probabilistic framework (Chikkerur et al., 2010).**

The present architecture does not address gaze shifts, although we have done so in earlier work (Schneegans, Spencer, Schöner, Hwang, & Hollingworth, 2014). The coordinate transforms involved in linking current retinal information to the scene representation are the route cause of the processing bottleneck that leads to the emergence of a sequential phase in the visual search demonstrated here.

Among tasks for future work is the need to link to the rich literature on how search times depend on the complexity of the search array (Wolfe, 2015). **Preliminary results show that when the correct item is selected first, search time increases linearly with the number of distractors due to global inhibition in the saliency field**.

References

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

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* (Vol. 7552 LNCS).

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A bayesian inference theory of attention. *Vision research*, *50*(22), 2233–2247.

Deco, G., & Rolls, E. T. (2004). A Neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, *44*, 621–642.

Duncan.J, & Humphrey, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, *96*(3), 433–458.

Fix, J., Rougier, N., & Alexandre, F. (2011). A Dynamic Neural Field Approach to the Covert and Overt Deployment of Spatial Attention. *Cognitive Computation*, *3*(1), 279–293.

Hulleman, J., & Olivers, C. N. L. (2015). The impending demise of the item in visual search. *Behavioral and Brain Sciences*(2017), 1–76.

Humphreys, G. W. (2016). Feature confirmation in object perception: Feature integration theory 26 years on from the Treisman Bartlett lecture. *Quarterly Journal of Experimental Psychology*, *69*(10), 1910–1940.

Humphreys, G. W., & Müller, H. J. (1993). Search via Recursive Rejection (SERR): A Connectionist Model of Visual Search. , *25*(1), 43–110.

Johnson, J. S., Spencer, J. P., Luck, S. J., & Schöner, G. (2009). A Dynamic Neural Field Model of Visual Working Memory and Change Detection. *Psychological Science*, *20*, 568–577.

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*, 14.

Malcolm, G. L., & Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision*, *9*(2009), 1–13.

Moran, R., Zehetleitner, M., Muller, H. J., & Usher, M. (2013). Competitive guided search: Meeting the challenge of benchmark RT distributions. *Journal of Vision*, *13*(8), 24–24.

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

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, & T. DFT Research Group (Eds.), *Dynamic thinking: A primer on dynamic field theory* (chap. 8). Oxford University Press.

Schneegans, S., Spencer, J. P., Schöner, G., Hwang, S., & Hollingworth, A. (2014). Dynamic interactions between visual working memory and saccade target selection. *Journal of vision*, *14*(11:9), 1–23.

Schöner, G., Spencer, J. P., & DFT Research Group, T. (2015). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.

Tatler, B. W., & Land, M. F. (2016). Everyday Visual Attention. In A. Kingstone, J. M. Fawcett, & E. F. Risko (Eds.), *The handbook of attention* (chap. 17). The MIT Press.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society (London) B Biological Sciences*, *353*, 1295–1306.

Wolfe, J. (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). Oxford University Press.

Wolfe, J. (2015). Visual Search. In A. Kingstone, J. M. Fawcett, & E. F. Risko (Eds.), *The handbook of attention* (pp. 27–56). The MIT Press.

Zibner, S. K. U., & Faubel, C. (2016). Dynamic scene representations and autonomous robotics. *Dynamic thinking: A primer on dynamic field theory*, 227–246.