A neural dynamic process model of combined bottom-up and top-down guidance in triple conjunction visual search

Raul Grieben (raul.grieben@ini.rub.de)
Ruhr-Universität Bochum, Institut für Neuroinformatik
Universitätsstraße 150, 44801 Bochum, Germany

Gregor Schöner (gregor.schoener@ini.rub.de)
Ruhr-Universität Bochum, Institut für Neuroinformatik
Universitätsstraße 150, 44801 Bochum, Germany

Abstract

The surprising efficiency of triple conjunction search has created a puzzle for modelers who link visual feature binding to selective attention, igniting an ongoing debate on whether features are bound with or without attention. Nordfang and Wolfe (2014) identified feature sharing and grouping as important factors in solving the puzzle and thereby established new constraints for models of visual search. Here we extend our neural dynamic model of scene perception and visual search (Grieben et al., 2020) to account for these constraints without the need for preattentive binding. By demonstrating that visual search is not only guided top-down, but that its efficiency is affected by bottom-up salience, we address a major theoretical weakness of models of conjunctive visual search (Proulx, 2007). We show how these complex interactions emerge naturally from the underlying neural dynamics.

Keywords: neural dynamic process model; dynamic field theory; visual search; binding; feature sharing; grouping; triple conjunctions; bottom-up salience; top-down guidance;

Introduction

Finding an object in a natural visual scene is something we do all the time without thinking about it. Sometimes it can be a little harder, for instance, when you scan the shelves of a supermarket you are unfamiliar with. Visual search is the basis of most actions directed at objects in the world, including talking about them. That is one reason why visual search has been the center of a vast research effort. Many laboratory studies have focused on the question of how object features guide visual search (Wolfe & Horowitz, 2017) and how different features are combined or “bound” (Treisman, 1998).

Treisman’s original “feature integration theory” (FIT) (Treisman & Gelade, 1980) postulated that visual features are processed in parallel, but attention is deployed serially to each object to bind the features. Treisman differentiated between inefficient *serial* conjunctive searches and efficient *parallel* feature searches. Wolfe proposed an alternative account called “guided search” (GS) (Wolfe, 2007) that postulated a continuum of search efficiencies, while retaining the core idea of binding through selective attention. In GS, the slope of the RT × set size function is the standard measure of search efficiency.

Evidence provided by Found (1998) that a third feature, that was correlated but irrelevant, could improve the efficiency of conjunctive search was not explained by either FIT or GS. The question of whether this increased efficiency was the result of preattentive binding remained open.

Nordfang and Wolfe (2014) revisited triple conjunction searches and found evidence that both *grouping*, the number of different distractor groups in a search display, and *feature sharing*, the number of features shared between a distractor and the target, had a substantial effect on search efficiency. In the five experiments relevant for evaluation of our model (1a, 1b, 3, 4, and 6) they tested seven[1] conditions with distractor groups sharing zero features (3D(0)), one feature (3D(1), 12D(1)), two features (3D(2)) or with distractor groups composed of items with zero, one, and two shared feature values (3D(012), 12D(012), 26D). Three (3D conditions), 12 (12D conditions), or 26 (condition 26D) different distractor groups were pseudo-randomly pulled from 26 distinct triple conjunctions (*color* (red, green, blue) × *orientation* (0°, 45°, 90°) × *shape* (rectangular, oval, jagged)). The distribution of features in each condition was constrained: In condition 3D(0), no distractor could share any feature with the target. In conditions 3D(1), 3D(012), 12D(1), 12D(012), and 26D, each feature type was present in 1/3 of the items (1/3 were red, 1/3 green, and 1/3 blue, and so on). In condition 3D(2), the target was always the red, vertical rectangle, 2/3 of the distractors were red, 2/3 vertical, and 2/3 rectangular. In addition, they probed two set sizes, 27 and 54.[2] In three further experiments they looked into questions not directly relevant for our model.[3] The experimental results showed that, despite a constant distribution of features, search efficiency decreased with increasing number of distractor groups in the display. This was not the result of a subset search strategy. Furthermore, they found evidence for a nonlinear effect of shared features: distractors sharing two features with the target had the highest negative impact on search efficiency, distractors sharing

[1] The eighth condition 5D was dropped by the authors after the first experiment (1a). As this condition offers no additional benefit over the conditions kept across experiments, we left it out of the model evaluation.

[2] We note that Nordfang and Wolfe (2014) reported set sizes 27 and 54. However, the experimental description and Figure 1 of the paper, suggest that the “3 Distractor Types” displays may have consisted of set sizes 28 and 55. We followed that description and used set sizes 28 and 55 for the 3D and 12D conditions, and set sizes 27 and 53 for the 26D conditions in simulation.

[3] They examined if a subset search strategy could explain the differences in efficiency (Exp. 2), if evidence for sharing and grouping could be found in brief presentations (Exp. 5), or extended the feature dimension to six to analyze how the sharing effect applied to more complex conjunctions (Exp. 6).

zero had no impact, and sharing one feature had only a small impact. The extension to higher dimensions showed that the relation between shared features and search efficiency stayed nonlinear. Exposure time had no effect on *grouping* or *sharing*. *Sharing* affected efficiency more strongly than *grouping*. Nordfang and Wolfe (2014) concluded that their findings could be explained by preattentive binding, but that very efficient top-down guidance based on a nonlinear *sharing* effect and/or nonlinear *grouping* effects in bottom-up salience may also account for the observations. As they expected these to be not trivial to model, the verification of their proposal remained open.

In this paper, we extend our neural dynamic process model for scene perception and top-down guided visual search (Grieben et al., 2020) to account for both the *sharing* and the *grouping* effect without a need for preattentive binding. We do so by incorporating a nonlinear bottom-up salience and extending the existing top-down guidance to integrate the necessary nonlinearity. To our knowledge, this is the first model that validates the theoretical hypothesis of Nordfang and Wolfe (2014) by not only incorporating their proposed mechanisms into a model but also fitting their experimental data. The model makes an important step toward a theoretical understanding of the interplay between bottom-up processing and top-down guidance in visual search, an issue in need of theoretical resolution (Proulx, 2007). In the model, parallel neural processes evolve in continuous time from which selection events emerge sequentially through dynamic instabilities.

Methods

The neural process model is based on Dynamic Field Theory (DFT; Schöner, Spencer, and DFT Research Group (2016)), a mathematical framework for using graded patterns of activation in neural populations evolving in continuous time to account for perception, action, and embodied cognition. Functional states are stable patterns of population activation. Their dynamic instabilities are the basis for the emergence of sequences of processing steps in which activation patterns transition between stable states.

Neural Dynamic Fields

By virtue of their forward connections from sensory surfaces or to motor surfaces, neural populations can be described as neural activation fields defined over feature or movement parameter dimension, $\boldsymbol{x}$ (Figure 1).

The continuous evolution, on the time scale τ, of fields emerges from the neural dynamics

$$\begin{aligned}\tau\dot{u}(\boldsymbol{x},t) = &-u(\boldsymbol{x},t)+h+s(\boldsymbol{x},t)+\xi(\boldsymbol{x},t)\\ &+\int \omega(\boldsymbol{x}-\boldsymbol{x}')\sigma(u(\boldsymbol{x}',t))d\boldsymbol{x}'\end{aligned} \qquad (1)$$

in which the negative resting level, h, and external input, $s(\boldsymbol{x},t)$, define a sub-threshold stable state, $u(\boldsymbol{x},t) = h+s(\boldsymbol{x},t) < 0$, as long as input is small and slowly varying. When localized input pushes sub-threshold activation

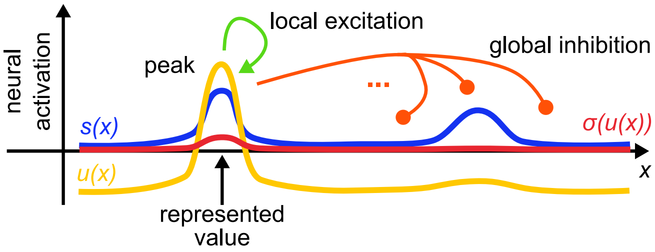

Figure 1: Dynamic neural field.

above the threshold of the sigmoidal nonlinearity, $\sigma(u) = 1/(1+\exp[-\beta u])$, the sub-threshold state becomes unstable in the *detection instability* and the system transitions to a supra-threshold peak of activation. Supra-threshold peaks of activation become unstable in the *reverse detection instability* when input is lowered sufficiently, with a bistable regime separating the two instabilities. The supra-threshold peak is shaped by intra-field neural interaction characterized by the interaction kernel, $\omega(\boldsymbol{x}-\boldsymbol{x}')$, that is excitatory over small, and inhibitory over large distances, $\boldsymbol{x}-\boldsymbol{x}'$, within the field. Stochastic switches between stable states may occur near instabilities. These are modeled by additive neural noise, $\xi(\boldsymbol{x},t)$.

Fields may operate in different dynamic regimes. In the *self-stabilized* regime, peaks are stabilized against decay and changes in input. In the *selective* regime, only a single peak is stable at a time. In the regime of *sustained activation*, peaks may persist when the localized input that induced them is removed. Stable activation peaks are the units of representation in DFT that encode perceptual estimates or movement parameters through their location within the space, $\boldsymbol{x}$, spanned by the field.

The dynamics of zero-dimensional fields or *nodes*

$$\tau\dot{u}(t) = -u(t)+h+s(t)+c\sigma(u(t))+\xi(t), \qquad (2)$$

has analogous stable states, “on” and “off”, and instabilities, detection (“activating the on state”) and reverse detection (“deactivating the on state”).

Networks of fields

Cognitive processes and motor behavior emerge from networks of fields defined by directional coupling among fields or nodes or, ultimately, to sensory-motor systems. Directional coupling or projection means that supra-threshold activation of one field provides either excitatory or inhibitory input to another field. The dependence of the projection strength on the dimensions of the fields is described by a connection kernel. Projections from higher-dimensional to lower-dimensional fields perform *dimensionality contraction* through marginalizing by integration. The reverse type of projection performs *dimensionality expansion* by providing input to sub-spaces (ridge or slice input). *Peak detectors* are neural nodes that receive the marginalized activation of a neural field as input. They switch to the on-state in a *detection* *instability*, if at least one supra-threshold peak exists in the input field. They remain in the off-state otherwise.

Match and Mismatch detection

For each feature dimension, three fields exist. The *expected* and *attended feature* fields represent, through a single peak of activation, feature values. They receive input from two different paths of the network. The *mismatch detection* field receives excitatory input from the *attended* and inhibitory input from the *expected feature* field. It generates a peak if *expected* and *attended feature* fields have peaks at different locations along the feature dimension.

For a given attended object location, the *feature matching* sub-network (Figure 2) compares (in parallel across feature dimensions) search cue (expected feature) and attended feature. A peak in all three fields (*attended feature*, *expected feature*, and *mismatch detection*) signals a no match, activating the *no-match response* node and inhibiting the *match response* node. Absence of a peak in the *mismatch detection* field, with peaks in the two other fields, signals a match and activates the *match response* node.

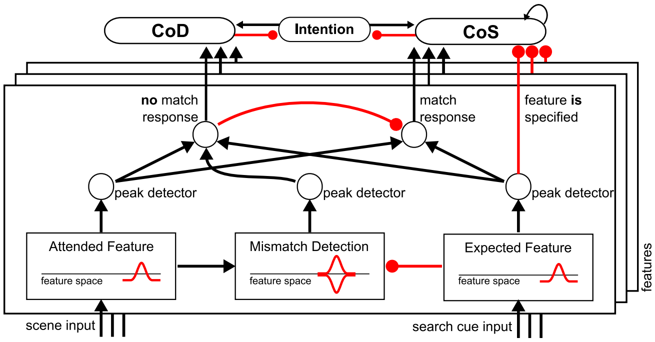

Figure 2: The *feature matching* sub-network. See the text for an explanation.

Mismatch within a single feature dimension is sufficient to activate the *condition of dissatisfaction* (CoD). In contrast, the *condition of satisfaction* (CoS) node is only activated if all attended features match the search cue. Together with the *intention* node, these two nodes are used to autonomously generate sequences of neural processing steps (Sandamirskaya & Schöner, 2010).

The neural dynamic process model

To account for the effects of feature sharing and grouping on the search efficiency of triple conjunction searches (Nordfang & Wolfe, 2014), we reduced our previous neural dynamic process model (Grieben et al., 2020) to its visual search component (removing sub-networks related to scene memory and transient detection). The simplified outline of Figure 3 groups dynamic neural fields into sub-networks (boxes) and their connectivity (arrows). The model is, however, really just a system of coupled neural integro-differential equations of the type shown in Equation 1. All neural activation fields and

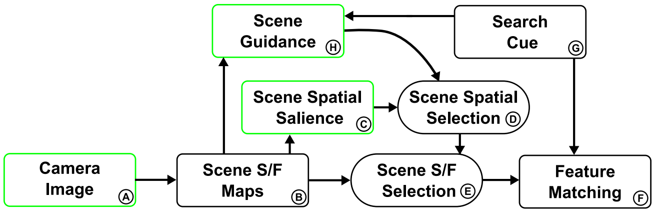

Figure 3: An overview of the neural dynamic process model. *Boxes* represent sub-networks of fields and *arrows* their couplings. Green outlines highlight sub-networks changed with respect to the previous model.

variables evolve continuously in time, dependent on online visual input. Instabilities create the impression of discrete events, but these simply emerge from the dynamics. The real-time numerical solution of the equations was achieved by implementing the model in *cedar*, a graphical programming interface for DFT models that also supports online visualization (Lomp, Richter, Zibner, & Schöner, 2016).

Feed-forward feature maps and salience map

The bottom-up pathway of the model (and of human perception) is a parallel preattentive process purely driven by input. In the model, visual input may come from a live camera image (*A*) or, in the current case, from randomly generated search displays (*A*1) (Figure 4).

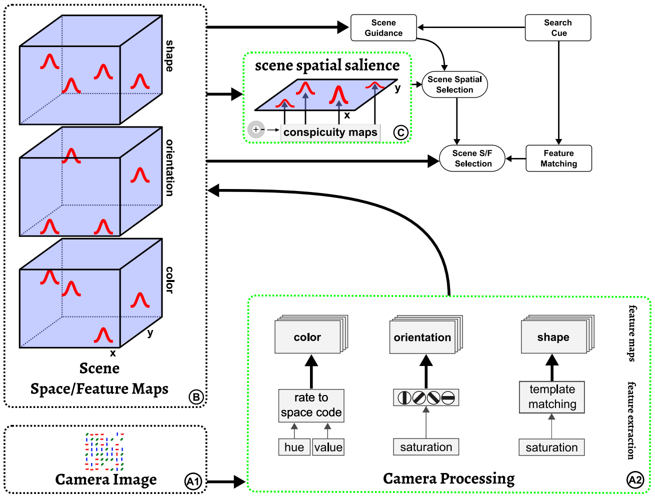

Figure 4: The bottom-up pathway of the model. See text for explanation. Green outlines highlight sub-networks changed with respect to the previous model.

Three features are extracted in parallel: *color*, *orientation*, and *shape*. Color is extracted from hue-space. Orientation is obtained by filtering the thresholded saturation with four elongated center-surround filters. To align with the experiments of Nordfang and Wolfe (2014), we swapped the *size* feature of our previous model (Grieben et al., 2020) to *shape*. Shape was obtained by template matching (normalized cross-correlation), a simplified account for preattentive recognition of simple shapes (Huang, 2020). These feature filters generate inputs that model the responses of feature sensitive neurons characterized by tuning curves. The neural activation pattern across the entire neural population for each feature is represented in the respective *scene space/feature map* (B). These neural space/feature representations are defined over the two dimensions of visual space and over one feature dimension. Their activation is marginalized along the feature dimension, using a center-surround filter as the projection kernel, resulting in a conspicuity map (C) for each feature. The inhibitory part of the center-surround kernel makes that the relative bottom-up salience of an object decreases linearly with the number of features shared with its flankers and also depends linearly on the number of flankers that share at least one feature with it. The excitatory part of the center-surround kernel (which is less strong for the shape feature dimension) makes that objects that are surrounded by empty space or by flankers that share no features with them become more salient.

These conspicuity maps are integrated in a spatial salience map, *scene spatial salience* field C (Itti & Koch, 2000). The output of this field (Figure 5), its activation passed through a sigmoidal threshold function, is the nonlinear bottom-up salience map that is responsible for the *grouping* effect. In our previous model (Grieben et al., 2020) all objects had the same bottom-up salience. The bottom-up salience map is low-pass filtered with a Gaussian filter.

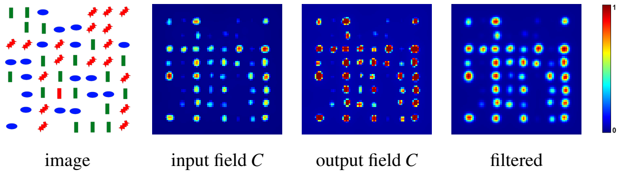

Figure 5: Bottom-up salience. See text for an explanation.

Attentional selection

The core cognitive processes of visual cognition require an attentional selection decision. The *scene spatial selection* field (D) plays, therefore, a central role in the model (see Figure 6). This field operates in the dynamic regime of selection, so that only one supra-threshold peak can be formed at any point in time. This provides the neural substrate for feature binding in the manner of Treisman's feature integration theory (Treisman & Gelade, 1980).

The *scene guidance* sub-network (H) consists of three *space/feature overlap* fields (H) that receive sub-threshold input from the *scene space/feature maps* (B) and feature input from the target *search cue* (G). At locations at which the cued features and the scene maps overlap, supra-threshold peaks form. The activation patterns of these fields are marginalized along the feature dimension to provide spatial input to the *feature guidance* field ($H1$). The resting level of the *feature guidance* field ($H1$) is down-regulated dynamically via inhibitory connections from the search cue sub-network (G) so that it decreases linearly with the number of cued features. This dynamical down-regulation is required to compensate for the linear dependence of the peak amplitude of the inputs to the field on the number of cued features. The output of this guidance field (Figure 7), its activation passed through a sigmoidal threshold function, provides nonlinear top-down bias for the *scene spatial selection* field (D), and is responsible for the *sharing* effect.

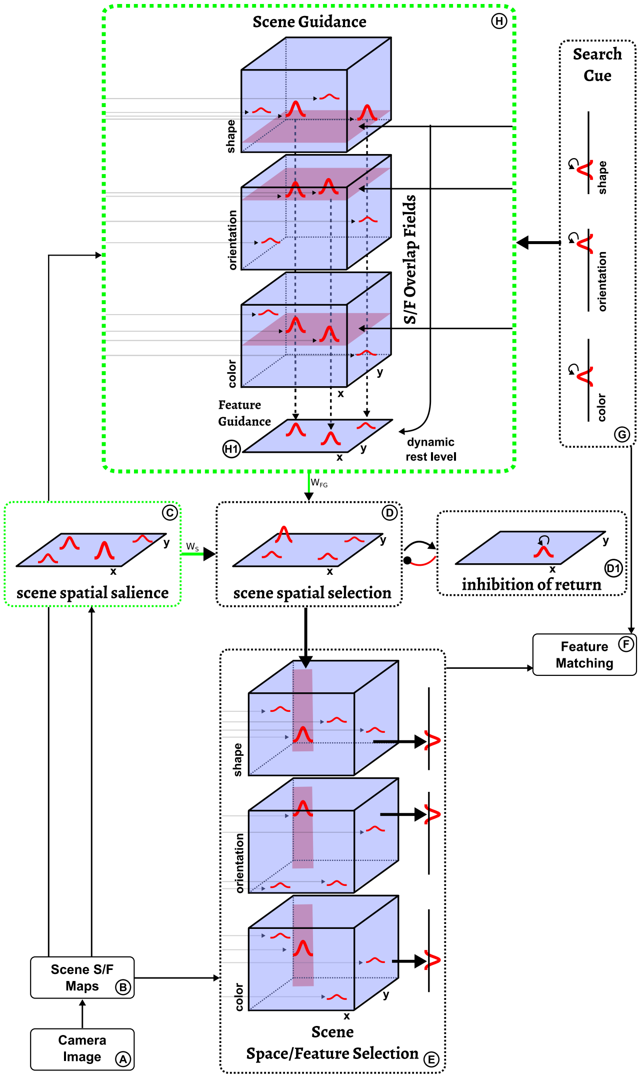

Figure 6: The sub-networks engaged in attentional selection and visual search. See the text for explanation. Green outlines highlight sub-networks changed with respect to the previous model.

The *scene spatial selection* field (D) receives weighted (W_S) bottom-up bias from the *scene spatial salience* field (C), and additional weighted (W_{FG}) top-down bias from the *scene guidance* sub-network (H) (Figure 8).

Visual search

Visual search is initiated automatically as soon as a peak is formed in the *scene spatial selection* field (D). It terminates

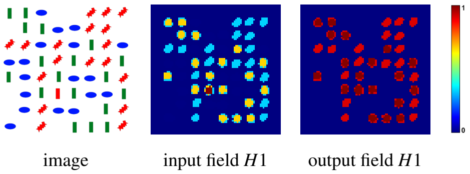

Figure 7: Top-down guidance. The input is a linear map of shared features, in which the target has the highest amplitude. The output is a non-linear transformation of that map, in which the target and distractors sharing two features have the same amplitudes, while distractors sharing one feature have a lower amplitude.

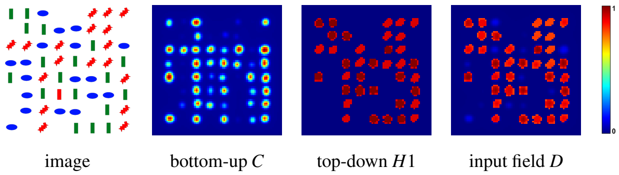

Figure 8: Combined bottom-up and top-down bias for the *scene spatial selection* field (D).

when all three features at the attended location match the features of the *search cue* (G), signaling successful completion of the visual search task. Responsible for this termination is the *feature matching* sub-network (F), whose *condition of satisfaction* (CoS) node is activated when this match occurs (see Figure 2). Otherwise, if at least one features mismatch is detected, the *condition of dissatisfaction* (CoD) node of the *feature matching* sub-network (F) is activated and inhibits the intention node (see Figure 2). This in turn destabilizes the *scene spatial selection* sub-network D, which deactivates the CoD itself. The intention node is released from inhibition (see Figure 2) and a new attentional selection takes place. That selection is biased away from previously attended locations through inhibitory input to the *scene spatial selection* field (D) from the *inhibition of return* field ($D1$) that contains self-sustained peaks at previously attended locations.

Results

We presented to our model five batches of stimuli, each consisting of 50 randomly generated search displays of the type shown in Figure 9, this for each combination of condition and set size (total of 700 displays per batch). Since we model a localization task, all displays contained a target. We used the same model parameters for all batches, conditions, and set sizes. The search displays were generated by an algorithm that followed the description of Nordfang and Wolfe (2014).

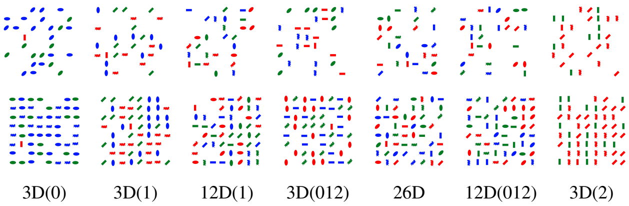

Figure 9: Randomly generated example displays for each combination of condition and set size.

Reaction time (RT) was determined in the model as time elapsed from the first activation of the visual search intention node to the activation of its CoS node. The slopes (range and mean) of the RT × set size functions for our model are shown in Table 1 together with the slopes from Nordfang and Wolfe (2014). We also tested our previous model (Grieben et al., 2020) on the same five batches to examine the effect of the bottom-up salience through the contrast to the present model. The slopes (range and mean) of the RT × set size functions for the previous model are also shown in Table 1. Time units in the models were fixed to align the model's with experimental time scales.

Discussion

We extended our neural dynamic process model for scene perception and top-down guided visual search (Grieben et al., 2020) to qualitatively fit the feature sharing and grouping effects found by Nordfang and Wolfe (2014) for triple conjunction searches. The proposed model accounts for the

Table 1: The slopes of the RT × set size functions from the experiments, the previous model, and our model.

	Experiments (Nordfang & Wolfe, 2014)							Model (Grieben et al., 2020)		Model (this paper)	
	1a	1b	3	4	6	Slopes	$\bar{x}$	Slopes	$\bar{x}$	Slopes	$\bar{x}$
3D(0)					-1.2	-1.2	-1.2	0.0	0.0	0.0	0.0
3D(1)	2.0	4.0	2.4	3.0	2.4	2.0 - 4.0	2.8	0.0	0.0	1.1 - 2.8	1.9
12(1)			2.8	4.8		2.8 - 4.8	3.8	0.0	0.0	2.1 - 3.1	2.5
3D(012)	2.3	4.3		5.8	3.7	2.3 - 5.8	4.0	2.4 - 4.4	3.5	2.0 - 5.7	4.0
26D	4.9	6.5	3.4	6.2		3.4 - 6.5	5.3*	2.0 - 4.4	2.5	3.7 - 6.3	4.8
12D(012)			3.7	6.7		3.7 - 6.7	5.2*	2.2 - 4.4	3.5	3.9 - 6.7	5.3
3D(2)					19.8	19.8	19.8	8.2-15.1	11.2	19.8 - 22.3	21.2

* The mean for the 12D(012) condition is possibly misleading and the result of too few data points, since, from the direct comparison on a per experiment level it seems clear that this condition is presumably less efficient than condition 26D.

differences between the conditions observed by Nordfang and Wolfe (2014) without resorting to preattentive binding.

These authors proposed that the number of shared features (*sharing*) may have a nonlinear effect on the efficiency of top-down guidance. In the model, such a nonlinear top-down effect results from a combination of the dynamic down-regulation via inhibitory connections from the search cue sub-network (G) and the sigmoidal nonlinearity of neural fields as shown in Equation 1. This proposed explanation for the *sharing* effect is particularly appealing since it emerges naturally from the underlying neural dynamics (as illustrated in simplified form in Figure 10).

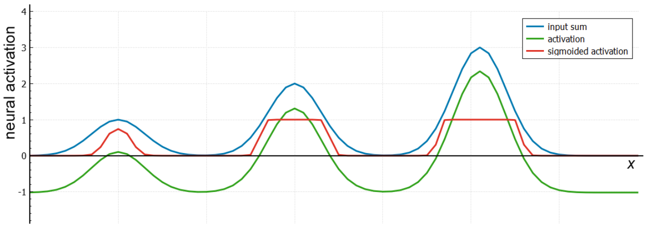

Figure 10: A simplified 1D version of the 2D feature guidance field. The target of a triple conjunction search (peak on the right) and a distractor sharing two features (center peak) have the same output amplitude (red line), while a distractor sharing one (peak on the left) has a lower output amplitude. The input amplitude (blue line), in contrast, encodes the number of shared features in a linear fashion.

Nordfang and Wolfe (2014) also proposed that nonlinear *grouping* in the bottom-up salience map may have a effect on the efficiency of visual search. In the model, such a nonlinear bottom-up effect results from a combination of the aforementioned sigmoidal nonlinearity of neural fields, and a center-surround filter as the projection kernel between the feature (B) and the salience maps (C). Center-surround filters model visual receptive fields and have been used in established models of bottom-up salience (Itti & Koch, 2000).

To validate the model, we compared in Table 1 the range and mean of the slopes of the RT × set size function with the empirically observed slopes (Nordfang & Wolfe, 2014). Since bottom-up salience depends on the distribution of distractors in the generated random search displays, a direct comparison with the experimental results is difficult without using the original displays used in experiment. This dependency on the precise nature of the stimulus lies in the nature bottom-up salience and is, in that respect, not just an issue for modeling. In fact, the high variance of experimental results across replications for the same conditions (Table 1) may reflect that problem. We ran one batch of random displays per experiment and compared the range of the slopes and their means rather than making only a one-to-one comparison for that reason. Nordfang and Wolfe (2014) did not report the ranges and means. Varying experimental setups may limit how directly comparable the different conditions were. We think, however, that our approach to comparison is conservative enough to conclude that the model qualitatively fits the empirical data well. The direct comparison with our previous model (Grieben et al., 2020), in Table 1, shows that the latter model is not able to explain the differences between the 3D and the 12D, or the 26D conditions. Since that model only uses efficient top-down guidance, we conclude that this differences reflects the result of the bottom-up *grouping* effect. Our model predicts that bottom-up salience deteriorates the overall search efficiency, in conflict with efficient top-down guidance. This is in line with the conclusions of Nordfang and Wolfe (2014) that the bottom-up salience should be thought of as noise that does not help to find the target. If the bottom-up salience favors the target, higher conjunctive visual search can be surprisingly efficient seemingly invoking the need of binding without attention. Our model shows, however, that these complex and surprising effects can be explained within the framework of binding through attentional selection in the spirit of FIT (Treisman & Gelade, 1980) and GS (Wolfe, 2007).

Even though bottom-up salience may disturb the efficiency of top-down guided visual search, it is crucial for the visual exploration of a crowded scene in the absence of a task. Through the incorporation of bottom-up salience our model is now able to autonomously explore the scene by bringing objects into the attentional foreground through selective competition, even in the absence of a task-induced top-down bias. In this sense, our model could be seen as including a neural implementation of the biased competition theory of attention (Desimone & Duncan, 1995).

In conclusion, it is important to keep in mind that the neural process model presented here actually generates individual selection decisions, as neural noise is amplified by neural interaction into a macroscopic activation peak. This is how parallel neural processes give rise to the discrete events at which features of selected objects are matched, non-matches are rejected, and the end of visual search is autonomously detected.

References

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*(1), 193–222.

Found, A. (1998). Parallel coding of conjunctions in visual search. *Perception & psychophysics*, *60*(7), 1117–1127.

Grieben, R., Tekülve, J., Zibner, S. K., Lins, J., Schneegans, S., & Schöner, G. (2020). Scene memory and spatial inhibition in visual search: A neural dynamic process model and new experimental evidence. *Attention, Perception, & Psychophysics*. doi: 10.3758/s13414-019-01898-y

Huang, L. (2020). Space of preattentive shape features. *Journal of vision*, *20*(4), 10–10.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*(10-12), 1489–1506.

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.

Nordfang, M., & Wolfe, J. M. (2014). Guided search for triple conjunctions. *Attention, Perception, & Psychophysics*, *76*(6), 1535–1559.

Proulx, M. J. (2007). Bottom-up guidance in visual search for conjunctions. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(1), 48.

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

Schöner, G., Spencer, J. P., & DFT Research Group, T. (2016). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.

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

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

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). Oxford University Press.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 0058.