Chapter 4
Toward Autonomous Intelligence: From Active 3D Vision to Invariant Object and Scene Learning, Recognition, and Search

Stephen Grossberg

Abstract How do we learn what a visually seen object is? How do our brains learn without supervision to link multiple views of the same object into an invariant object category while our eyes scan a scene, even before we have a concept of the object? Indeed, why do we not link together views of different objects when there is no teacher to correct us? Why do not our eyes move around randomly? How do they explore salient features of novel objects and thereby enable us to learn view-, size-, and positionally invariant object categories? How do representations of a scene remain binocularly fused as our eyes explore it? How do we solve the Where's Waldo problem and thereby efficiently search for desired objects in a scene? This article summarizes the ARTSCAN and ARTSCENE families of neural models, culminating in the 3D ARTSCAN Search model that clarifies how the brain solves these problems in a unified way by coordinating processes of 3D vision and figure-ground separation, spatial and object attention, object and scene category learning, predictive remapping, and eye movement search. ARTSCAN illustrates revolutionary new computational paradigms whereby the brain computes: Complementary Computing clarifies the nature of brain specialization, and Laminar Computing clarifies why all neocortical circuits exhibit a layered architecture. ARTSCAN also provides unified explanations and simulations of brain and behavioral data, and computer simulation benchmarks that support the model,

S. Grossberg (✉)
Center for Adaptive Systems, Boston University, Boston, USA
e-mail: steve@bu.edu
URL: http://cns.bu.edu/~steve

S. Grossberg
Graduate Program in Cognitive and Neural Systems, Boston University, Boston, USA

S. Grossberg
Center for Computational Neuroscience and Neural Technology, Boston University, Boston, USA

S. Grossberg
Departments of Mathematics, Psychology and Biomedical Engineering, Boston University, Boston, USA

R. Wang and X. Pan (eds.), *Advances in Cognitive Neurodynamics (V)*,
Advances in Cognitive Neurodynamics, DOI 10.1007/978-981-10-0207-6_4

which provides a blueprint for developing a new type of system for active vision and autonomous learning, recognition, search, and robotics.

Keywords Adaptive resonance theory · Object recognition · Spatial attention · Object attention · Category learning · View-invariance · Visual search · Predictive remapping · Attentional shroud · Consciousness · Visual cortex · V3A · Inferotemporal cortex · Parietal cortex · Prefrontal cortex · Perirhinal cortex · Parahippocampal cortex

4.1 Invariant Object Category Learning, Recognition, and Search

The 3D ARTSCAN search model predicts how valued objects are learned, recognized, and searched with freely moving eyes in a three-dimensional (3D) scene [2, 8, 11, 12, 15, 16, 18, 19]. Accomplishing this requires a synthesis of mechanisms for spatial and object attention, invariant object category learning, predictive remapping, reinforcement learning and motivation, and attentive visual search. This synthesis provides functional explanations and predictions of interactions between brain regions such as cortical areas V1, V2, V3a, V4, PPC, LIP, ITp, ITa, and PFC. Such a competence is needed in a wide range of behaviors, including the recognition of objects and scenes, visual-based navigation toward a goal object, classical conditioning in response to recognized objects, and episodic learning of events that include objects. The current article provides a conceptual overview of some of the major new concepts and mechanisms that have been needed to achieve this competence. Model equations, simulations, and data references can be found in the archival articles. A few particularly salient data references are also included here.

4.2 ARTSCAN

One of the several basic problems for which solutions were offered in 3D ARTSCAN search concerns how the brain is able to learn a view-invariant object category. When the eyes freely scan a scene, they can foveate views of many objects. How does the brain know how to associate only views that belong to the same object with an emerging view-invariant object category, before a concept of the object is known, and without any external supervision? In particular, suppose that the eyes foveate a particular view of a teacup, leading to rapid learning of a view-specific category, say in the posterior inferotemporal cortex (ITp). When these ITp cells are activated, they also activate cells in the anterior inferotemporal cortex (ITa) that will learn to represent the object from multiple views and will thus become a view-invariant category. As the object view that is being inspected changes enough, a new view-specific category is learned, and the previous one is inhibited to enable this to happen. Inhibiting the first view-specific category eliminates the input to ITa that activated the cells there. Why do not these cells also shut off? They must not shut off because all of the view-specific categories that are learned while the object surface is scanned should be able to be associated with them, thereby creating a view-invariant category.

The ARTSCAN model [11] is the first model to be developed in the ARTSCAN family. It proposed how view-invariant object categories can be learned and recognized as the eyes freely scans a 2D scene. In particular, ARTSCAN predicted that the ITa cells are not inhibited because a parietal reset mechanism that could have inhibited them is itself inhibited while the eyes scan the attended object surface. The reset mechanism is predicted to be inhibited by an *attentional shroud* [24], or form-fitting distribution of spatial attention, also in the parietal cortex. The shroud is maintained by a *surface-shroud resonance*, or positive feedback loop, between prestriate visual cortex (e.g., V4) and parietal cortex. This prediction implies that when spatial attention shifts to another object so that its shroud collapses, the parietal reset will briefly be disinhibited, leading to a transient reset burst that inhibits the view-invariant object category. Then, the brain is ready to attend a new object and to learn to recognize it. Experimental evidence for this predicted sequence of events was reported by Chiu and Yantis [9] using fast event-related fMRI in humans. These data provide an important experimental marker to further test this hypothesis.

The ARTSCAN circuit also clarifies how the eyes can scan multiple views of an object before shifting spatial attention to another object [23], thereby enabling such a view-invariant category to be learned. This explanation clarifies, in particular, why the eyes do not gaze randomly around a scene. This scanning process uses feedback between the object surface and its generative boundaries via *surface contour* feedback signals. The surface contour feedback signals arise in the thin stripes of cortical area V2. They strengthen boundaries that are consistent with them in the pale stripes of cortical area V2, while inhibiting spurious boundaries, and in so doing trigger of figure-ground separation, so that spatial attention *can* focus on one object surface at a time.

A parallel branch of these surface contour signals play several additional functional roles. These additional roles are made possible by the fact that surface contour signals are computed by a contrast-enhancing on-center off-surround network in response to successfully filled-in surface brightnesses and colors within surface regions that are surrounded by closed boundaries. Only such surfaces can enter conscious awareness. Because of the contrast-enhancing lateral inhibitory process, surface contours have larger activities at high curvature points, which are just the kinds of positions where salient features occur. Thus, the signals in this parallel branch, which is predicted to pass through cortical area V3A [3], can be used to command the eyes to look at the positions of salient features. This is accomplished by contrast-enhancing the surface contour further to pick the most active position at any time, and then iterating this process while spatial attention remains focused on that object surface. These positions become the target positions for eye movements on the object surface, and act as *attention pointers* [7] for where the eyes will look next.

In addition to activating saccadic eye movements via brain regions such as the frontal eye fields and superior colliculus, a parallel branch of these positional signals also acts to quickly update *gain fields* that keep the shroud, which is computed in head-centered coordinates, stable during scanning eye movements to different salient features on the object surface, so that the shroud *can* keep the reset mechanism inhibited during these movements. These positional signals hereby cause a *predictive remapping* [21] of the shroud in anticipation of where the eye movement will go. A great deal of data have been explained and predicted by these mechanisms, including data about the reaction time costs of moving the eyes to positions outside an object versus to the positions inside it [1].

The surface-shroud resonance process has additional functions. One important one concerns my prediction that all conscious percepts of visual qualia are surface percepts that are part of surface-shroud resonances. This prediction reconciles two earlier predictions; namely, that "all conscious states are resonant states" [13] and "all conscious percepts of visual qualia are surface percepts" [14]. Together these two predictions led to the question: What sort of resonance supports conscious surface percepts of visual qualia? My answer is: a surface-shroud resonance. This prediction enables the explanation of even more data, including clinical data about how parietal neglect occurs; see Grossberg [16] for a review.

4.3 Positional ARTSCAN

The positional ARTSCAN, or pARTSCAN, model [8], further developed ARTSCAN to explain how view-, position-, and size-invariant object categories can be learned and recognized during free scanning of a 2D scene. These invariances are not perfect if only because of the cortical magnification factor, and the model can quantitatively simulate the invariance properties that are exhibited in neurophysiological experiments on IT cells [18, 25].

pARTSCAN was able to learn these additional invariant properties by incorporating the fact that some IT cells exhibit persistent activities. It could then, in addition, explain quite a bit of additional neurobiological data, notably the target swapping data of Li and DiCarlo [22]. These data are conceptually important because they demonstrate conditions under which an invariant object category in IT can be readily recoded by swapping two objects during a saccadic eye movement to the position of the first object. These results raise the question: Why is not such "catastrophic forgetting" ubiquitous? pARTSCAN predicts that this recoding occurs because the reset mechanism does not get activated when the targets are rapidly swapped during an eye movement. This prediction can be tested by fusing the Chiu and Yantis [9] and Li and DiCarlo [22] paradigms: Increase the interstimulus interval between swapped objects and measure when a reset burst occurs. When the interstimulus interval between swapped targets is large enough to cause reset, recoding should not occur, or should at least be greatly attenuated.

4.4 Distributed ARTSCAN

Why does not a scene appear black outside the region that is selected by a surface-shroud resonance? The distributed ARTSCAN, or dARTSCAN, model was developed to clarify this sort of issue during scanning of 2D scenes [12]. dARTSCAN supplements the slow attention of a surface-shroud resonance, which has its source at surface representations in the What cortical stream, with the fast attention that is activated by transients due to object change or motion via the Where cortical stream. In addition, the spatial attentional representations of the parietal cortex are extended to spatial attentional representations of the prefrontal cortex that enables multimodal attention, notably simultaneous priming of multiple regions in a scene, thereby spreading attention beyond the focal attention studied in ARTSCAN and enabling faster switching of attention between objects. dARTSCAN has been used to simulate a variety of additional challenging data about spatial attention, including larger data sets about reaction time costs of shifting spatial attention to a position outside an object versus to delete "the" one inside it, attentional crowding, and useful-field-of-view tasks, including how video game players can train themselves to have broader attentional spans and greater situational awareness. Crowding is of particular interest due to the theoretical link of the ARTSCAN models between spatial attention in the Where stream and object recognition in the What stream. The model proposes how, when a given object cannot form its own shroud, and is rather part of a single shroud that envelops several nearby objects, then that object cannot be easily recognized.

4.5 ARTSCAN Search

The previous modeling variants are all consider issues related to object learning and recognition. Correspondingly, they propose how Where cortical stream mechanisms modulate What cortical stream mechanisms for this purpose. After learning to recognize an object, it is important to be able to search for it in a scene, and to thereby engage it through motor actions. Such a model needs What-to-Where stream interactions in addition to Where-to-What stream interactions. The next model clarifies how this happens by proposing a solution of the Where's Waldo problem. This ARTSCAN Search model suggests how either a cognitive prime in prefrontal cortex, or a motivational source such as the amygdala, can drive a search to determine the position of a valued object in a scene. At least two new design problems must be solved to do this.

One problem concerns the fact that invariant object categories are insensitive to the position of a target. Such invariance enables the brain to overcome the combinatorial explosion that would have occurred if every view, position, and size of an object on the retina needed to generate its own representation for purposes of recognition. In addition, it is much easier to motivationally amplify an invariant representation in the orbitofrontal cortex, using incentive motivational signals from the amygdala, than it would have been to deal with myriad non-invariant object representations. Once a valued invariant object representation is amplified, it can win the competition for attention and thereby drive further processing. However, because the invariant representation is insensitive to Waldo's position, its activation must somehow be able to activate representations that are sensitive to positions which can drive eye, arm, and other movements toward Waldo. This problem is solved in the model using the fact that ARTSCAN learns both view-specific object categories and invariant object categories. The view-specific categories, which are proposed to exist in ITp, are also sensitive to object position. Thus, somehow invariant object categories in ITa need to be able to activate appropriate view- and position-specific categories in ITp and, from there, positional representations of the object in the frontal eye fields and parietal cortex.

However, such top-down signals are typically, without further processing, priming signals that can sensitize or modulate the activity of target cells, but cannot, by themselves, fully activate them. Such top-down priming signals are said to obey the ART Matching Rule, and theorems have been proved showing how modulatory top-down expectations that focus attention using the ART Matching Rule can led to self-stabilized learning [4–6], thereby solving what I have called the *stability-plasticity dilemma* [13]. In order to fully fire primed cells, a volitional signal from the basal ganglia is also needed. Convergence of a top-down expectation with a volitional signal converts the subliminal priming signals into signals that can vigorously fire their target cells and thereby activate a top-down cascade of processing steps to locate Waldo.

The ARTSCAN Search model was shown, by computer simulations, to be competent to find Waldo in a scene composed of realistic CalTech 101 object images, even when the model computations use the cortical magnification factor. The model can also simulate all the data that other variants of ARTSCAN can. However, without further mechanisms, this model cannot quantitatively simulate more challenging data about object search, notably data in which iterative learning about scenic context can drive a more efficient search. Such a search is said to be *contextually cued* [10].

4.6 ARTSCENE and ARTSCENE Search

The ARTSCENE Search model [20] can do this. ARTSCENE Search builds upon the ARTSCENE model [17], which clarifies how the gist of a scene can be learned, and how gist may be refined by attention shifts that learn finer features of a scene. In ARTSCENE, gist is computed as a coarse texture category, and finer scenic features are finer texture categories. All these categories vote to predict scene types in a database of natural scenes.

The ARTSCENE Search model additionally proposes how sequential object and spatial contexts can be used to accumulate evidence about a scene that can be used to efficiently search for desired goal objects in it. Such a contextually cued search clarifies why, for example, after seeing several kitchen appliances, such as a stove, microwave, and sink, one is more likely to expect to see a refrigerator than a jungle, and also where to look to find that refrigerator in a familiar kitchen. Such a search requires that the brain uses object and spatial working memories and plans to determine where next to look. To achieve this competence, the ARTSCENE Search model simulates how temporal cortex, parietal cortex, perirhinal cortex, parahippocampal cortex, and prefrontal cortex all contribute to contextually cued memory and search. The model can quantitatively simulate many of the key properties of the rich psychophysical database about contextually cued search.

4.7 3D ARTSCAN

All of the above model variants consider learning, recognition, and/or search with freely moving eyes in a 2D scene. How does the brain accomplish this in a 3D scene? A key fact motivates how this is done by the 3D ARTSCAN model [19]. This fact concerns what happens when we fuse a Julesz binocular stereogram or Magic Eye autostereogram. It may take a few seconds before the images that are received by each of our eyes can binocularly fuse into a percept of a scene in depth. However, after fusion occurs, our eyes can move across the scene without breaking fusion, even though all the image features are received by different retinal positions after each eye movement. This property, which is dramatically illustrated by stereograms but which we take for granted during daily life, shows that the fused representations are not computed in retinal coordinates. Rather, they are computed in head-centered coordinates that remain invariant under eye movements.

In ARTSCAN, gain fields were needed to rapidly update the head-centered coordinates of a shroud so that it could maintain inhibition of the parietal reset mechanism during eye movements that scan salient features on an object. The 3D ARTSCAN model shows that additional gain fields are needed to rapidly update, and predictively remap, head-centered representations of binocularly fused perceptual boundaries so that they do not collapse every time the eyes move. The model also shows how these invariant boundaries can maintain the 3D surface percepts that we consciously see, even though these surface percepts are computed in retinotopic coordinates, as is obvious every time our eyes move and the conscious percept of each object in a scene shifts in the opposite direction. The 3D ARTSCAN model was also tested on CalTech 101 object images, and was shown capable of simulating various psychophysical data, notably the reaction time costs of shifting attention outside versus inside an object.

4.8 3D ARTSCAN Search and Autonomous Adaptive Mobile Robots

Taken together, these models embody a 3D ARTSCAN search model that clarifies how our brains can learn, invariantly recognize, and search for a valued goal object in a 3D scene. This model family has explained and predicted a wealth of psychological and neurobiological data about this topic, as can be reviewed in the archival articles. It can also be used as a blueprint for a future generation of increasingly autonomous adaptive mobile robots.

References[1]

1. Brown, J.M., Denney, H.I.: Shifting attention into and out of objects: evaluating the processes underlying the object advantage. Percept. Psychophys. **69**, 606–618 (2007)
2. Cao, Y., Grossberg, S., Markowitz, J.: How does the brain rapidly learn and reorganize view- and positionally-invariant object representations in inferior temporal cortex? Neural Netw. **24**, 1050–1061 (2011)
3. Caplovitz, G.P., Tse, P.U.: Rotating dotted ellipses: motion perception driven by grouped figural rather than local dot motion signals. Vision. Res. **47**, 1979–1991 (2007)
4. Carpenter, G.A., Grossberg, S.: A massively parallel architecture for a self-organizing neural pattern recognition machine. Comput Vis Graph Image Process **37**, 54–115 (1987)
5. Carpenter, G.A., Grossberg, S.: Pattern recognition by self-organizing neural networks. MIT Press, Cambridge (1991)
6. Carpenter, G.A., Grossberg, S.: Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. Trends Neurosci. **16**, 131–137 (1993)
7. Cavanagh, P., Hunt, A.R., Alfraz, A., Rolfs, M.: Visual stability based on remapping of attention pointers. Trends Cogn. Sci. **14**, 147–153 (2010)
8. Chang, H.-C., Grossberg, S., Cao, Y.: Where's Waldo? How perceptual cognitive, and emotional brain processes cooperate during learning to categorize and find desired objects in a cluttered scene. Front. Integr. Neurosci. (2014). doi:10.3389/fnint.2014.0043
9. Chiu, Y.C., Yantis, S.: A domain-independent source of cognitive control for task sets: shifting spatial attention and switching categorization rules. J. Neurosci. **29**, 3930–3938 (2009)
10. Chun, M.M.: Contextual cueing of visual attention. Trends Cogn. Sci. **4**, 170–178 (2000)
11. Fazl, A., Grossberg, S., Mingolla, E.: View-invariant object category learning, recognition, and search: how spatial and object attention are coordinated using surface-based attentional shrouds. Cogn. Psychol. **58**, 1–48 (2009)
12. Foley, N.C., Grossberg, S., Mingolla, E.: Neural dynamics of object-based multifocal visual spatial attention and priming: object cueing, useful-field-of-view, and crowding. Cogn. Psychol. **65**, 77–117 (2012)
13. Grossberg, S.: How does a brain build a cognitive code? Psychol. Rev. **87**, 1–51 (1980)
14. Grossberg, S.: 3-D vision and figure-ground separation by visual cortex. Percept. Psychophys. **55**, 48–121 (1994)
15. Grossberg, S.: Cortical and subcortical predictive dynamics and learning during perception, cognition, emotion, and action. Philos. Trans. R. Soc. Lond. **364**, 1223–1234 (2009)

[1] Grossberg references downloadable from http://cns.bu.edu/~steve

16. Grossberg, S.: Adaptive resonance theory: how a brain learns to consciously attend, learn, and recognize a changing world. Neural Netw. **37**, 1–47 (2013)
17. Grossberg, S., Huang, T.-R.: ARTSCENE: a neural system for natural scene classification. J. Vis. **9**(6), 1–19 (2009)
18. Grossberg, S., Markowitz, J., Cao, Y.: On the road to invariant recognition: explaining tradeoff and morph properties of cells in inferotemporal cortex using multiple-scale task-sensitive attentive learning. Neural Netw. **24**, 1036–1049 (2011)
19. Grossberg, S., Srinivasan, K., Yazdanbakhsh, A.: Binocular fusion and invariant category learning due to predictive remapping during scanning of a depth scene with eye movements. Front. Psychol: Percept. Sci. (2014). doi:10.3389/fpsyg.2014.01457
20. Huang, T.-R., Grossberg, S.: Cortical dynamics of contextually cued attentive visual learning and search: spatial and object evidence accumulation. Psychol. Rev. **117**, 1080–1112 (2010)
21. Irwin, D.E.: Information integration across saccadic eye movements. Cogn. Psychol. **23**, 420–456 (1991)
22. Li, N., DiCarlo, J.J.: Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science **321**, 1502–1507 (2008)
23. Theeuwes, J., Mathôt, S., Kingstone, A.: Object-based eye movements: the eyes prefer to stay within the same object. Atten. Percept. Psychophys. **72**, 12–21 (2010)
24. Tyler, C.W., Kontsevich, L.L.: Mechanisms of stereoscopic processing: stereo attention and surface perception in depth reconstruction. Perception **24**, 127–153 (1995)
25. Zoccolan, D., Kouh, M., Poggio, T., DiCarlo, J.J.: Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J. Neurosci. **27**, 12292–12307 (2007)