A Unified Theory of Exogenous and Endogenous Attentional Control

Michael C. Mozer and **Matthew H. Wilder**

{mozer,matthew.wilder}@colorado.edu

Department of Computer Science and Institute of Cognitive Science

University of Colorado, Boulder, CO 80309-0430

August 11, 2007

Abstract

Although diverse, theories of visual attention generally share the notion that attention is controlled by some combination of three distinct strategies: (1) exogenous cueing from locally-contrasting primitive visual features, such as abrupt onsets or color singletons (e.g., Itti & Koch, 2001); (2) endogenous gain modulation of exogenous activations, used to guide attention to task relevant features (e.g., Navalpakkam & Itti, 2005; Wolfe, 1994); and (3) endogenous prediction of likely locations of interest, based on task and scene gist (e.g., Torralba, Oliva, Catelhano, & Henderson, 2006). Because these strategies can make conflicting suggestions, theories posit arbitration and combination rules. We propose an alternative conceptualization consisting of a single unified mechanism that is controlled along two dimensions: the degree of task focus, and the spatial scale of operation. Previously proposed strategies—and their combinations—can be viewed as instances of this mechanism. Our theory offers a means of integrating data from a wide range of attentional phenomena. More importantly, the theory yields an unusual perspective on attention that places a fundamental emphasis on the role of experience and task-related knowledge.

1 Introduction

The human visual system can be configured to perform a remarkable variety of arbitrary tasks. For example, in a pile of coins, we can find the coin of a particular denomination, color, or shape, determine whether there are more heads than tails, locate a coin that is foreign, or find a combination of coins that yields a certain total. The flexibility of the visual system to task demands is achieved by control of visual attention.

Three distinct control strategies have been discussed in the literature, Earliest in chronological order, *exogenous* control was the focus of both experimental research (e.g., Averbach & Coriell, 1961; Posner & Cohen, 1984) and theoretical perspectives (e.g., Itti & Koch, 2000; Koch & Ullman, 1985). Exogenous control refers to the guidance of attention to distinctive, locally contrasting visual features such as color, luminance, texture, and abrupt onsets. Theories of exogenous control assume a *saliency map*, a spatiotopic map in which activation in a location indicates saliency or likely relevance of that location. Activity in the saliency map is computed in the following way. Primitive features are first extracted from the visual field along dimensions such as intensity, color, and orientation. For each dimension, broadly tuned, highly overlapping detectors are assumed that yield a coarse coding of the dimension. For example, on the color dimension, one might posit spatial feature maps tuned to red, yellow, blue, and green. Next, local contrast is computed for each feature—both in space and time—yielding an activity map that specifies distinctive locations containing that feature. These *feature-contrast* maps are summed together to yield a saliency map. Saliency thus corresponds to all locations that stand out from their spatial or temporal neighborhood in terms of their primitive visual features.

Subsequent research showed that attention need not be deployed in a purely exogenous manner but could be influenced by task demands (e.g., Bacon & Egeth, 1994; Folk, Remington, & Johnston, 1992; Wolfe, Cave, & Franzel, 1989). These results led to theories proposing *feature-based endogenous* control (e.g., Baldwin & Mozer, 2006; Mozer, 1991; Navalpakkam & Itti, 2005; Wolfe, 1994, 2007). In these theories, the contribution of feature-contrast maps to the saliency map is weighted by endogenous *gains* on the feature-contrast maps, as depicted in Figure 1. The result is that an image such as that in Figure 2a, which contains singletons in both color and orientation, might yield a saliency activation map like that in Figure 2b if task contingencies involve color or

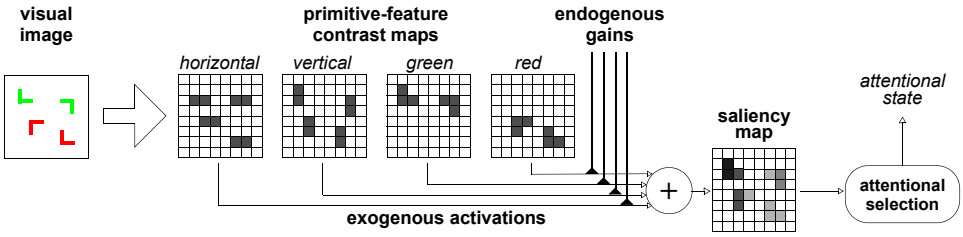

Figure 1: A depiction of feature-based endogenous control

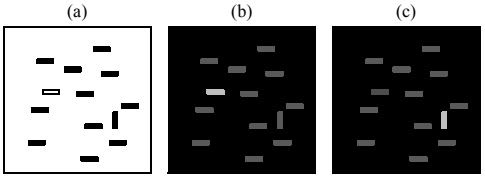

Figure 2: (a) a display containing two singletons, one in color and one in orientation; (b) a saliency map if color is a task-relevant feature; (c) a saliency map if orientation is a task-relevant feature

like that in Figure 2c if the task contingencies involve orientation.

Experimental studies support a third attentional control strategy, in which attention is guided to visual field regions likely to be of interest based on the current task and coarse properties of the scene (e.g., Biederman, 1972; Neider & Zelinsky, 2006; Torralba, Oliva, Catelhano, & Henderson, 2006; Wolfe, 1998). This type of *scene-based endogenous* control seems intuitive: If you are looking for your keys in the kitchen, they are likely to be on a counter. If you are waiting for a ride on a street, the car is likely to appear on the road not on a building. Even without a detailed analysis of the visual scene, its gist can be inferred—e.g., whether one is in a kitchen or on the street, and the perspective from which the scene is viewed—and this gist can guide attention.

1.1 Arbitrating Among Control Strategies

Given evidence that three distinct control strategies—exogenous, feature-based endogenous, and scene-based endogenous—can influence the allocation of attention, more than one strategy might be applied in any situation. Thus, theories must address the metacontrol issue of which strategy or combination of strategies to apply.

The saliency map can serve as a common output medium for the three strategies, allowing the strategies to operate in parallel and have their results combined in the saliency map. Alternatively, an arbitration mechanism may select which strategy to apply. For example, in Guided

Search (Wolfe, 1994), the saliency map sums the output of distinct exogenous and feature-based endogenous control processes. And the more recent framework of Torralba et al. (2006) proposes parallel pathways for determining exogenous and scene-based endogenous control, and activity in the saliency map is the product of activity in the two pathways.

Although theories of attentional control suppose distinct control strategies operating in parallel, little experimental evidence exists to support this notion. In neuroimaging studies, large-scale neural systems for endogenous and exogenous control are indistinct (e.g., Rosen et al., 1999; Peelen, Heslenfeld, & Theeuwes, 2004). Of course, distinct mechanisms of control may operate in the same cortical area, but behavioral studies also provide evidence against multiple control strategies operating in parallel. Rather, control strategies appear to trade off. For example, increasing task difficulty via target-nontarget similarity decreases the impact of an irrelevant singleton in brightness (Proulx & Egeth, 2006; Theeuwes, 2004). That is, as the need for feature-based endogenous control increases, exogenous control decreases.

1.2 A Unifying Framework

Instead of conceiving of the three control strategies as distinct and unrelated mechanisms, the principal contribution of this work is to characterize the strategies as points in a *control space*. As depicted in Figure 3, the control space is two dimensional, one being the task dependence of control—the degree to which control is modulated by the current goals and tasks—and the other being the spatial scale—from local to global—of featural information used for control. Exogenous control uses information at a local spatial scale and operates independently of current goals. Feature-based and scene-based endogenous control both operate with a high degree of task dependence, but utilizing information at different spatial scales. (The figure shows two other points in the control space which we discuss later in the chapter.)

What does laying out the three strategies—which we refer to as the *primary* strategies—in a control space buy for us? The control space offers us the means to reconceptualize attentional control. Rather than viewing attentional control as combining or arbitrating among the primary strategies, one might view control as choosing where to operate in this continuous two-dimensional space. By this scheme, control would always involve selecting a single strategy—a single point in the space—but the resulting strategy could appear to combine aspects of the primary strategies.

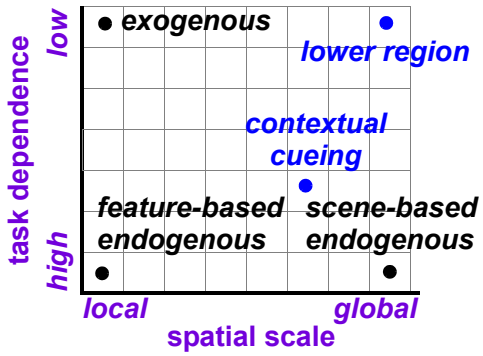

Figure 3: A two dimensional control space that characterizes exogenous, feature-based endogenous, and scene-based endogenous control of visual attention. Other cases of attentional control can also be characterized in terms of this space, including contextual cueing and lower-region figure-ground. Details in the text.

Ultimately, we would like to argue for this strong claim, but at the moment we do not have evidence to support such a claim. Instead, we focus in this chapter on showing that the control space offers a general perspective that encompasses existing notions of attentional control. We make this point by presenting a model that implements the notion of the control space and can account for key experimental results in the literature. The model therefore offers a unified perspective on attentional control and into the relationships among the primary strategies, which have heretofore been conceptualized as distinct and unrelated. We call the model ***S****patial-scale and* ***TA****sk-dependence* ***C****ontrol* ***S****pace*, or *STACS* for short.

1.3 An Illustration of Saliency Over the Control Space

To give a concrete intuition about the operation of STACS, we present several examples illustrating the model's behavior. The input to STACS is an image—natural or synthetic—and the output from STACS is a saliency map. To operate at different points in the control space, STACS must be given: (1) the spatial scale at which it should process the image, (2) the degree of task dependence, and, (3) when there is some task dependence, a specification of the task. We focus on visual search, and therefore define a task to be search for a particular object or class of objects.

Figures 4–6 show sample images and tasks, along with the saliency map produced by STACS at different points in the control space. Figure 4 shows a street scene, and the task is to search for the people. Examining the grid of saliency maps, the map in the upper left corner reflects local scale, task independent processing, or what we referred to as exogenous control—a saliency map

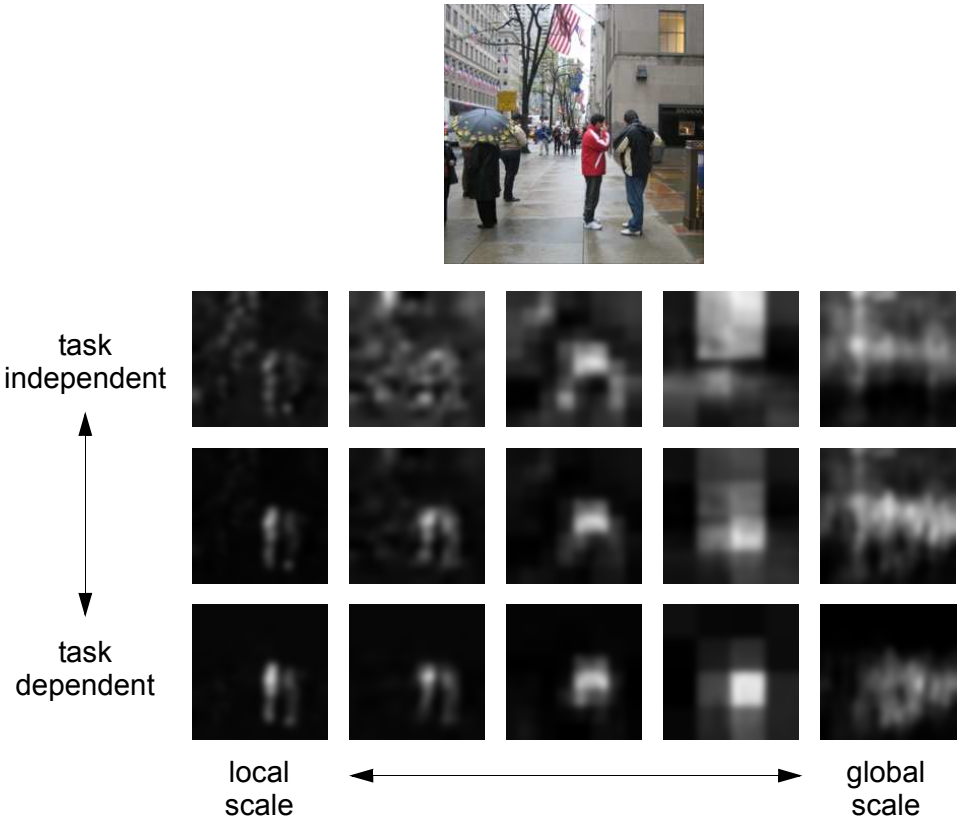

Figure 4: A street scene, and saliency maps produced by STACS in response to this input for five spatial scales (columns) and three levels of task dependence (rows). The task is to find the *people* in the image. In the text of this chapter, we interpret the saliency maps.

that indicates high contrast locations in the image independent of the task. The map in the lower left corner reflects feature-based endogenous control—a saliency map that predicts the presence of people based on local features. And the map in the lower right corner reflects scene-based endogenous control—a saliency map that predicts the likely locations of people based on the scene gist. The middle row of maps in the Figure correspond to an intermediate degree of task specificity, and the middle columns in the Figure correspond to an intermediate spatial scale.

Figures 5 and 6 show a similar grid of saliency maps for the tasks of searching for cars and buildings, respectively. As in Figure 4, it is apparent that the task plays a significant role in modulating the saliency in the task-dependent regions of the control space. The point of these three examples is to show that a continuum of saliency maps can be achieved, and how this continuum encompasses the primary control strategies.

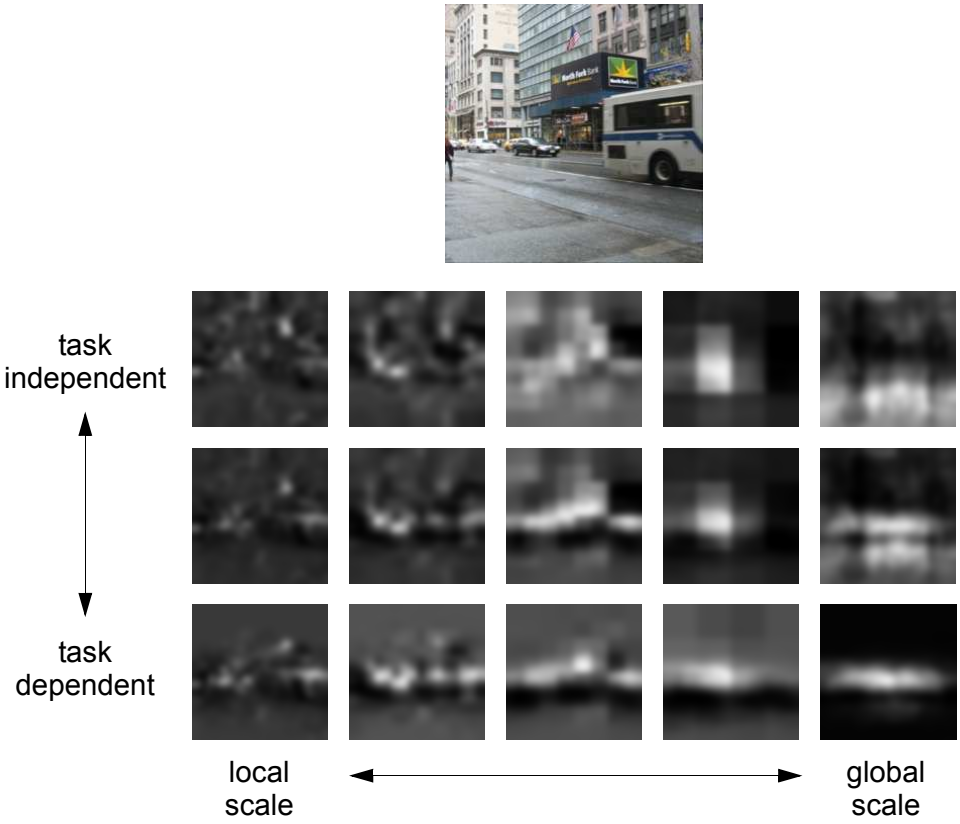

Figure 5: A street scene, and saliency maps produced by STACS in response to this input for five spatial scales (columns) and three levels of task dependence (rows). The task is to find the *cars* in the image.

2 An Implementation of the Unified Framework for Attentional Control

As with all models of attention, STACS assumes that computation of the saliency map can be performed in parallel across the visual field with relatively simple operations. If computing saliency was as computationally complex as recognizing objects, there wouldn't be much use for the saliency map, because the purpose of the saliency map is to provide rapid heuristic guidance to the visual system.

Given the control settings of spatial scale and a specific task, STACS is configured to perform the mapping from image to saliency map. Rather than creating yet another model of attention, our aim in developing STACS was to generalize existing models, such that existing models can be viewed as instantiations of STACS with different control settings. We focus on two classes

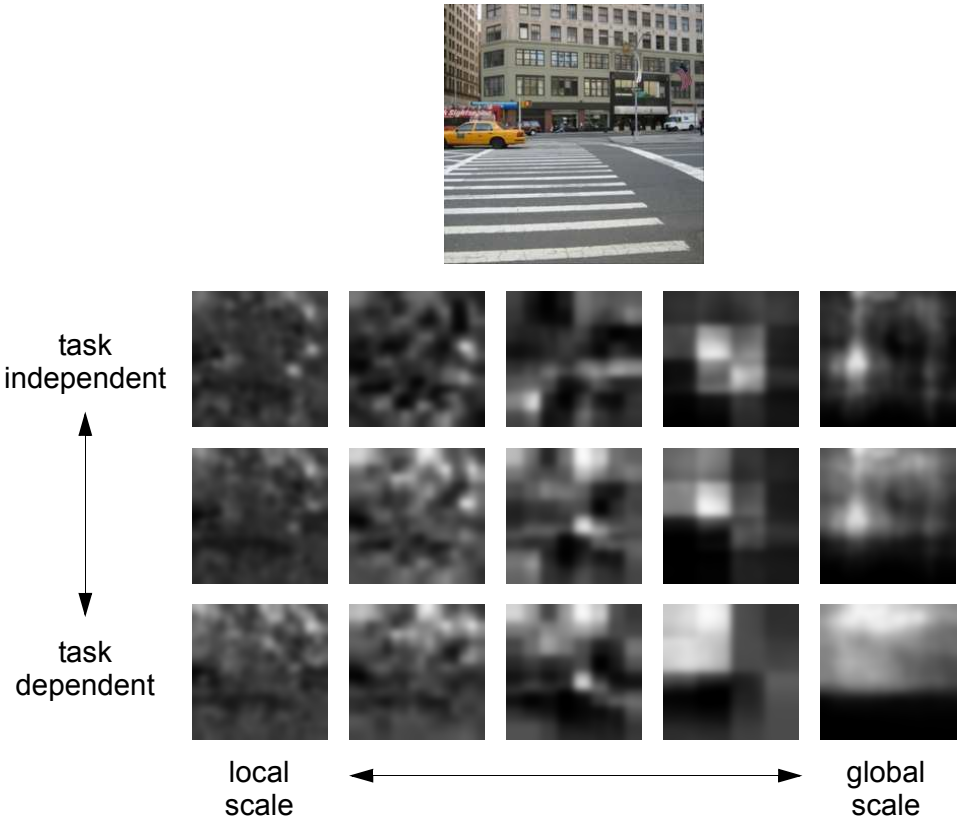

Figure 6: A street scene, and saliency maps produced by STACS in response to this input for five spatial scales (columns) and three levels of task dependence (rows). The task is to find the *buildings* in the image.

of models in particular. One class combines exogenous and feature-based endogenous control, as exemplified by Navalpakkam and Itti (2005)—hereafter *NI*—and the Wolfe (1994) Guided Search model which is quite similar to NI in its key claims. Another class combines exogenous and scene-based endogenous control, as most recently represented by Torralba, Oliva, Catelhano, and Henderson (2006)—hereafter, *TOCH*. Table 1 shows that the basic processing stages of NI and TOCH are quite similar, and the union of the two models yields a more general model, STACS. We now describe these stages in more detail, and explain how STACS can be configured to implement various control strategies. Figure 7 provides a schematic depiction of the STACS architecture.

2.1 Feature Extraction

All images presented to STACS are normalized to be 256×256 pixels. Color filters (red, green, blue, and yellow) are applied to each pixel in the image, as are Gabor filters centered on each

Table 1: Processing stages of two existing models of attentional control

stage	**Navalpakkam & Itti (2005); Wolfe (1994)**	**Torralba, Oliva, Castelhano, & Henderson (2006)**
parallel feature extraction with detectors having broad, overlapping tuning curves	color, orientation, luminance	color, orientation at multiple spatial scales
contrast enhancement	via center-surround differencing	via cross-dimensional normalization
dimensionality reduction	no	yes
associative network to compute saliency	linear	mostly linear with a Gaussian squashing function

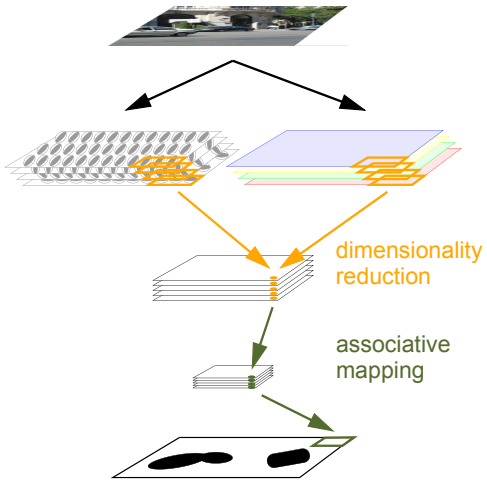

Figure 7: A schematic depiction of the STACS architecture, showing the patchwise analysis of an image to yield a saliency map.

pixel at four orientations (0°, 45°, 90°, and 135°) with a diameter of 9 pixels. Feature extraction is performed using the matlab Saliency Toolbox, which essentially implements the NI model with a few minor simplifications.[1] Figure 8a shows a sample image and Figure 8b shows the feature representation at the finest spatial scale.

2.2 Contrast Enhancement

All subsequent analysis is based on local *patches* of the image. The size of the patch depends on the spatial-scale control setting given to the model. We implemented 5 scales, with patches at the

[1]The saliency toolbox, www.saliencytoolbox.net, was developed by Dirk Walther.

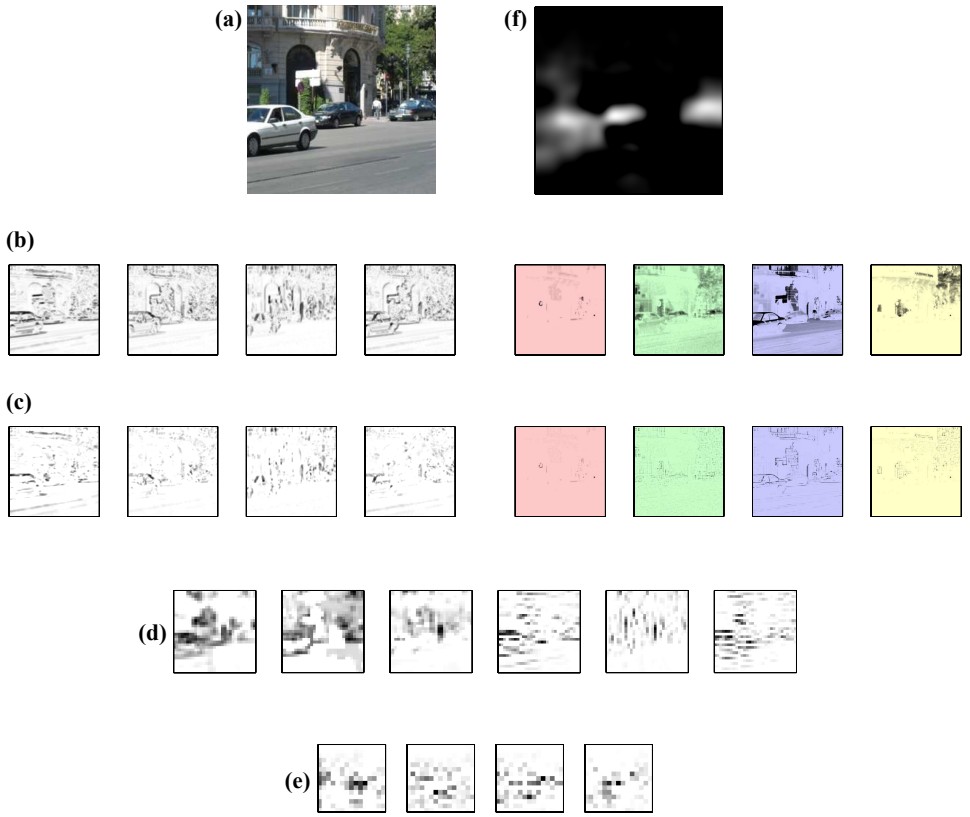

Figure 8: Activities in STACS at the finest spatial scale at various stages of processing: (a) an input image; (b) feature activities for four orientation maps and four color maps; (c) activities following contrast enhancement; (d) activities from 6 principal components following subsampling; (e) activities of four hidden units; and (f) the output saliency map.

scales being 16×16, 32×32, 64×64, 128×128, and 256×256. We also experimented with finer scales, but they were too fine for the images we used to provide much interesting information. The columns in Figures 4-6 correspond to the 5 spatial scales of our implementation.

To enhance distinctive regions in color or orientation, STACS—like nearly every other model of early visual processing—performs local contrast enhancement, increasing the activities of features that are distinct from their neighborhood. We explored several different methods, including that used by NI (see Itti & Koch, 2000). The method we found to yield the best feature pop-out involves multiplying the activity of each feature detector by the ratio of the Gaussian-weighted mean activity of that detector in the local neighborhood to the Gaussian-weighted activity of all features on the same dimension (orientation or color) in a broader neighborhood. The standard deviation of the 'center' and 'surround' Gaussians are 5% and 50% of the patch size, respectively, for all spatial scales. Figure 8c shows the contrast-enhanced image representation.

2.3 Dimensionality Reduction

From this stage of analysis and beyond, the image is analyzed in terms of local patches that overlap one another by 50%, with the patch size dependent on the spatial scale (as specified in the previous section). The representation of an image patch has $8d^2$ elements, where d is the diameter of the patch. For large patches, the dimensionality of this representation is quite large. Consequently, models processing at the coarser spatial scales, such as TOCH, include some type of dimensionality reduction of the representation. STACS incorporates the two dimensionality reduction techniques of TOCH: subsampling and principal components analysis (PCA). For the 5 spatial scales of STACS, the representation was subsampled along the x and y axes by factors f of 4, 4, 8, 8, and 16, and a resulting representation of size $8d^2/f^2$. The subsampled representation was fed into PCA the c components with highest eigenvalues were used to represent the patch, where, for the 5 scales, c is 32, 40, 48, 56, and 64. The principal components are extracted in a location invariant manner using training on random patches of a training image corpus of real-world, naturalistic images. Figure 8d shows the representation of the first six principal components across the sample image in Figure 8a.

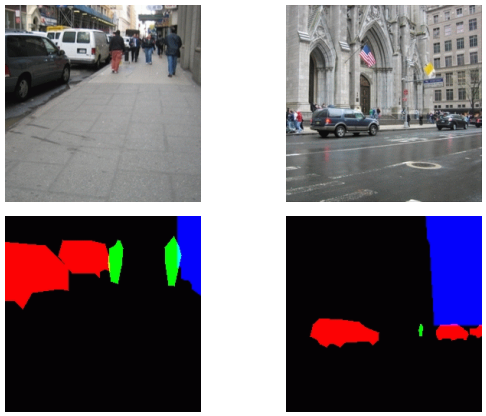

Figure 9: (above) Two images from the LabelMe data base, (below) pixel-by-pixel labeling of three objects—cars (red), people (green), and buildings (blue).

2.4 Associative Mapping

The last stage of STACS computes the saliency of each location in the visual field based on the dimensionality-reduced, contrast-enhanced representation of the image. In the spirit of using simple, minimal computation, the saliency computation is performed via an associative linear mapping. This mapping is performed in overlapping patches, and the patchwise outputs are summed to obtained the saliency representation. A supervised learning procedure obtains a set of linear coefficients for each patch (location) and each set of control settings (spatial scale and task).

Training is based on a labeled set of images: images in which each pixel is classified as to the presence or absence of a target object at that particular location. Figure 9 shows two sample images and labeling (via different colors) as to the locations of cars, people, and buildings. The images and target labels are from the LabelMe data base (`http://labelme.csail.mit.edu`). We trained models for the tasks of searching for cars, people, buildings, lamps, trees, roads, windows, and signs. Each of these objects was found in at least 80 images in the LabelMe data base.

STACS is trained to perform object localization: given a task of searching for a target object in an image, STACS should identify where in the image the object appears. This goal is a quick and rough cut at object recognition. STACS is quick because of the simplistic architecture, but STACS is also accuracy limited because of the simplistic architecture: The associative mapping is linear, and further, it is rank limited. We achieve the rank limitation by implementing the mapping as a neural network with a linear hidden layer (depicted as the next-to-last layer in the architecture of Figure 7). At all spatial scales, we use a bottleneck of 5 hidden units. Because they

are linear, the hidden units provide no additional computational power, but rather, they serve as a bottleneck on the flow of information used to compute saliency. Figure 8e shows the activity of hidden units across image patches, and Figure 8f shows the saliency activity pattern resulting from the associative mapping.

2.5 Implementing Task-Dependent Control

For each task and each spatial scale, the necessary knowledge to perform the associative mapping is contained in the connection strengths feeding to and from the hidden units. One can conceptualize the model as having a pool of hidden units for each task and spatial scale, and the appropriate pool is enabled depending on attentional control settings.

Because the associative mapping is linear, STACS produces sensible behavior when more than one pool of hidden units is enabled. For example, if the bicycle and car pools are both enabled, the saliency map will be a superimposition of the maps for bicycles and cars. Thus, linearity makes it possible to combine multiple tasks. One could imagine a hierarchy of tasks, from the very specific—corresponding to a single pool—or very general—corresponding to multiple pools, e.g., the pools for all wheeled vehicles. At an extreme, if the pools for all tasks are enabled, then STACS operates in a *task independent* manner. Thus, the control dimension of task specific to task independent is achieved by the enabling of different pools of hidden units. In Figures 4-6, the saliency maps with an intermediate degree of task dependence—in the second row of the Figures—are obtained by enabling the units for the specific task, and weakly enabling all other task units.

3 Simulation Results

Having described STACS, we now to turn to data and phenomena that lie within its scope. At this point in our development, our primary goal is to show that STACS serves as a unifying theory that can subsume existing models and the phenomena those models are designed to accommodate.

3.1 Exogenous Control

Exogenous control is achieved in STACS via control settings that specify local spatial scale and task-independent processing. Because task-independent saliency is obtained by superimposing the

outputs of task-specific saliency maps, exogenous control in STACS depends on what tasks it has been trained on. We have a fairly small corpus of eight tasks, one shouldn't expect STACS's exogenous control to be terribly robust in our preliminary implementation. Nonetheless, the basic approach is a novel claim of STACS: exogenous control deploys attention to locations that are of interest in *any* task. The upper left saliency map in Figures 4-6 are examples of exogenous control in STACS. As with all models of exogenous control, it's difficult to evaluate performance of STACS. Nonetheless, the sort of locations picked by STACS seem reasonable. Ultimately, we will validate STACS's performance via a collection of eye movement data in free viewing, such as that of Bruce and Tsotsos (2006). However, we suspect that all models of exogenous control will yield about the same performance and it will be impossible to discriminate models on the basis of their fit to free viewing eye movements alone. STACS has the additional virtue of explaining all varieties of attentional control.

3.2 Scene-Based Endogenous Control

STACS can perform in a mode like the TOCH model of scene-based endogenous control. TOCH computes regions of interest based on the task and a scene gist. STACS performs very similarly to TOCH when operating in a task-dependent manner at a coarse spatial scale. Indeed, the computations are nearly identical except that TOCH incorporates a Gaussian nonlinearity at the final stage of determining saliency. However, it's far from clear from the output of TOCH and STACS that this nonlinearity has a qualitative effect. The lower-right saliency map in Figures 4-6 are examples of scene-based endogenous control in STACS. STACS yields sensible behavior, qualitatively like that of TOCH: the upper region of the image is salient when searching for buildings, small bright discontinuities are salient when searching for people, and the middle region of the image along the side of the road is salient when searching for cars.

3.3 Feature-Based Endogenous Control

STACS can perform in a mode like the NI model of feature-based endogenous control. In STACS, feature-based control requires essentially the same training procedure as we described for scene-based control, except that rather than targets being complex objects such as people and buildings, the target is a simple feature, such as the presence of the color red or a vertical line. We train

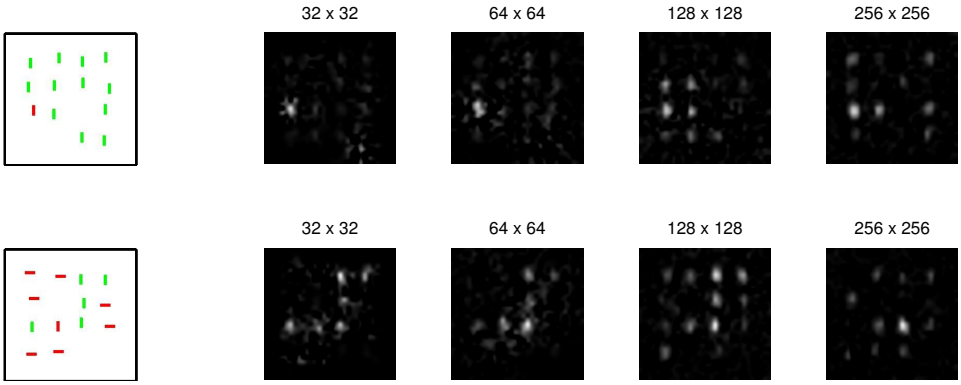

Figure 10: (top row) modeling simple feature search: on the left is an image containing a single red item among greens; on the right are saliency maps at different spatial scales for this image. (bottom row) modeling conjunction search: on the left is an image containing a single red vertical among green verticals and red horizontals; on the right are saliency maps at different spatial scales for this image.

a STACS hidden layer for each possible target feature. The training data consists of random synthetic images of oriented, colored lines of variable length. The target saliency output indicates the locations containing the target feature. (The learning that takes place in STACS is analogous to learning the meaning of a word such as 'red' or 'vertical'. Assigning saliency to a specific feature is computationally trivial. The challenge is in determining which task is associated with which feature.)

We now show how STACS performs the basic tasks in the visual search literature. First consider the simple-feature search task of finding a red item among green items. The top row of Figure 10 shows an instance of such a display. Beside the image are saliency maps for four spatial scales, all indicating high saliency for the target in this example. Now consider the conjunction-search task of finding a red vertical among green verticals and red horizontals, as shown in the bottom-row image of the Figure. To perform conjunction search, STACS can either be trained on conjunction targets (red verticals) or the task can be defined by conjoining the hidden-unit pools for the two component features (red and vertical). In both implementations, STACS is unable to clearly distinguish the target from the distractors, at any spatial scale. The reason for this difficulty is the linearity and bottleneck of the associative mapping: with limited resources, STACS is unable to accurately localize complex combinations of primitive features.

To perform a more formal analysis of visual search tasks, and to model human data, it is necessary to read response times from STACS, which in turn requires an additional assumption

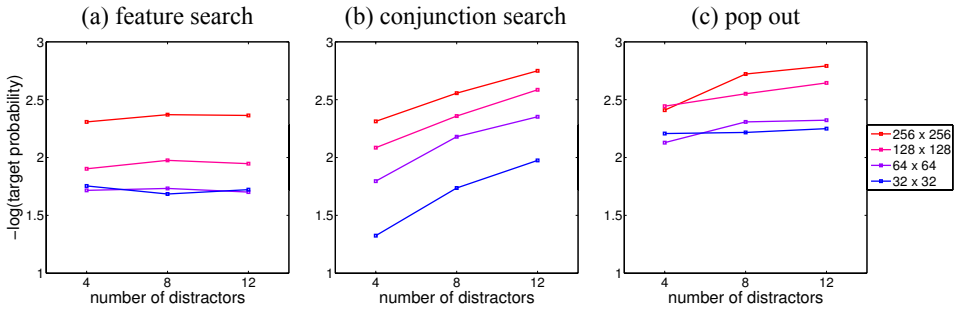

Figure 11: Simulation results from STACS for (a) feature, (b), conjunction, and (c) pop out search. Each graph contains one line for each of the four coarsest spatial scales. The x-axis of each graph indicates the number of distractor elements in the search displays. The y-axis represents response time read out from the model, assumed to be proportional to the negative log proportion of saliency in the neighborhood of the target.

concerning how the saliency map is used in search. Guided Search (Wolfe, 1994) makes the assumption that locations are searched in order from most salient to least salient. However, in Wolfe's model, there is a one-to-one correspondence between objects and locations. In our images, the notion of 'location' is ill defined: Should a location correspond to a pixel? If so, then are neighboring pixels two distinct locations? To avoid such ambiguities, we make an alternative assumption. We assume that the probability of attending to a given pixel is proportional to its saliency. Based on this assumption, we can compute the probability of attending to a region in the neighborhood of the target (i.e., the pixels containing the target and some immediately surrounding pixels). And we can make the standard assumption that response times are related to probability via a negative log transform, i.e., $RT = -log(P(n))$, where RT is the response time and n denotes the target neighborhood.

Figure 11a graphs feature-search performance as a function of the number of distractors in the display. The Figure contains four lines, one for each of the four coarsest spatial scales of operation. The response time (negative log probability) of the model is independent of the number of distractors, regardless of the spatial scale that the model is operating at. This simulation result is in accord with human experiments that find efficient search for features. Note that in absolute terms, the local spatial scales are most efficient for performing the search. One expects this result because the target can be identified based on local features alone (global context does not help), and because STACS has greatest spatial resolution at the local scale.

Figure 11b graphs conjunction-search performance as a function of the number of distractors.

Consistent with human experiments, search is inefficient in that response times increase linearly with the number of elements in the display. As with feature search, conjunction search is fastest if STACS operates at the local spatial scale.

Figure 11c graphs the outcome of a pop-out or oddball-detection task. The task is simply to find the element that differs from others in the display, e.g., the red one among green, or the vertical among horizontals. Although this task involves detecting a single feature, it is distinct from feature-search in that the feature dimension and value are not known in advance of a trial. Consequently, STACS must operate in a task-independent manner (i.e., without a specific target). We trained STACS on four feature search tasks—for verticals, horizontals, red, and green—and treated the pop out task as the union of these four specific tasks. As the Figure shows, STACS produces flat search slopes, at least for the local spatial scales. Note that although the slopes are flat, the absolute response time is slower than for feature search, not surprising considering that the feature-search task is more narrowly delineated. The simulation findings are consistent with human experiments.

3.4 Other Phenomena

We have shown that STACS provides an unified account of varied phenomena that were previously explained in terms of distinct control strategies. By viewing these control strategies as points in a control space (Figure 3), STACS leads one to question whether other points in the control space correspond to known attentional phenomena. In this section, we mention two examples.

3.4.1 Lower Region

Vecera, Vogel, and Woodman (2002) found a novel cue to figure-ground assignment. In displays like Figure 12a, viewers tended to perceive the lower region as the foreground figure (the gray region in this image). A rational explanation of this phenomenon is that because viewers are ordinarily upright, the lower region of their visual field contains objects of interest. We interpret the lower-region cue to figure-ground assignment as a sort of endogenous control that operates on a global spatial scale—i.e., it is based on the overall scene properties—and is task independent—i.e., it operates in the absence of a specific target of search. We have indicated this point in the control space in Figure 3. Lower region, occupying the fourth corner of the control space, is a natural

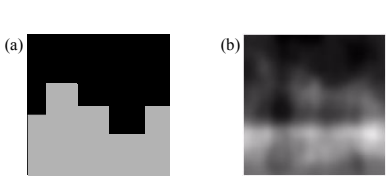

Figure 12: (a) a sample display from Vecera et al (2002); (b) sample output from STACS operating in a task-independent mode at the global spatial scale

complement to the three control strategies described previously.

When shown images like that in Figure 12a, STACS, configured for the global spatial scale and task-independent processing, yields a saliency map like that in Figure 12b. Because this saliency map is formed by combining task-specific maps, and we trained STACS on only eight tasks, one can justifiably be cautious in making strong claims about this result. Nonetheless, the result was an emergent consequence of training. STACS produces a similar result regardless of the colors of the two regions.

3.4.2 Contextual Cueing

Chung and Jiang (1998) found that repeating configurations in a visual search task led to a speed up of response times. Participants were shown displays like that in Figure 13a containing a single target—the letter T—among distractors–L's at various orientations. The participants' task was to report the orientation of the T (pointing to the left or the right). Unbeknownst to participants, some display configurations were repeated over the course of the experiment. In these *predictive* displays, the target and distractors appeared in the same locations, although their orientations and colors could change from trial to trial. After several exposures, participants are roughly 60 msec faster to predictive displays than to random or *nonpredictive* displays.

Figure 13b shows results from a STACS simulation of Chun and Jiang (1998). The simulation was trained for the task of detecting the specific targets (a T facing to either the left or the right). As in earlier simulations, the model's response time is assumed to be proportional to the negative log of the relative saliency (or probability) of the target. The four sets of bars are for different spatial scales. At all spatial scales, the model produces faster response times for predictive than nonpredictive displays. However, only at the most global scales is the effect large in magnitude. The small effect at the local scales is due to the fact that the model can learn to suppress locations

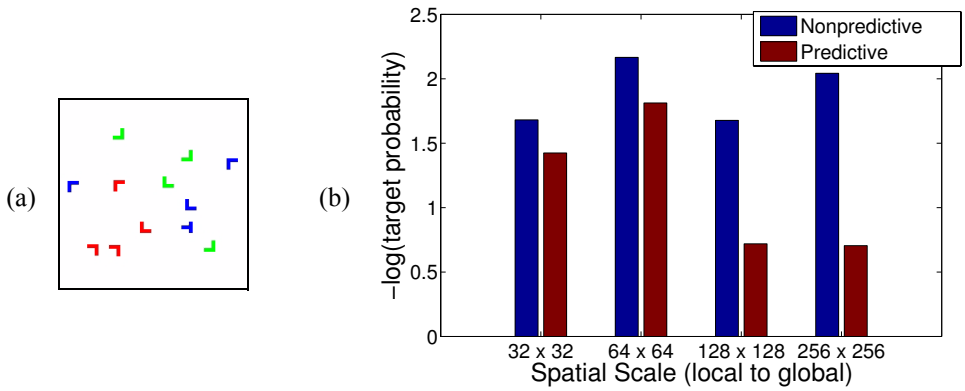

Figure 13: (a) a sample display from Chun and Jiang (1998); (b) performance of STACS operating at various spatial scales for predictive and nonpredictive displays

less likely to contain the target. The effect at the global scales is due to learning about display configurations. The model obtains a contextual cueing effect not only at the scale in which the entire display is processed as a whole (256×256), but also at the scale of half of the display (128×128), suggesting that contextual cueing might occur for displays in which only the distractors in the quadrant closest to the target are predictive.

We therefore characterize contextual cueing as a phenomenon in the control space (Figure 3) that occurs with an intermediate-to-global spatial scale and that has at least some degree of task dependence.

4 Neurobiological Implications of the Model

We introduced a perspective on attentional control, STACS, that attempts to integrate theoretical ideas from existing models and to provide a unified framework for considering a wide range of attentional phenomena. STACS can be seen as a generalization and unification of existing models of attentional control, in particular TOCH and NI.

Although designed to identify salient locations in the visual field, STACS effectively performs a crude sort of object detection. For each location in the visual field, STACS estimates the probability that a target is present at that location given the visual features in the neighborhood. What distinguishes STACS from a full-blown model of object recognition is the fact that STACS is computationally limited: dimensionality reduction bottlenecks and the linearity of the model restrict what can be computed. The virtue of these limitations is that STACS is computationally bounded.

Computational simplicity should go hand-in-hand with speed in the brain, and a quick response is essential if saliency is to provide useful guidance for more detailed, computationally intensive processing of regions of the visual field most likely to be relevant.

Beyond the virtue of speed, linearity in the model has two important additional benefits. First, training STACS with gradient descent works well because the linear connections have no local optima. As a result, gradient descent learning procedures can be incremental and ongoing. In contrast, nonlinear neural networks tend to get stuck in a region of weight space from which they can not escape, even if the training data changes over time. Because training can be incremental and ongoing, it is easy to model the effects of recent experience on performance, as we did with the contextual cueing simulation.

A second benefit of linearity of the associative mapping is that it allows STACS, trained on single tasks, to perform—in principal—arbitrary combinations of tasks dynamically simply by enabling pools of task units. For example, search for a red vertical can be specified by the enabling the combination of red and vertical tasks, each of which has a dedicated pool of processing units. STACS can search for wheeled vehicles by enabling car, bus, bike, and train tasks. And, STACS can operate in an exogenous control mode simply by enabling *all* tasks in parallel. In practice, the ability of STACS to perform combinations of tasks is limited both by its ability to perform individual tasks and by the fact that the linear combination can produce locations with high saliency that are not the most salient in any one task.

This view of control suggests a specific role of cortical feedback. A neural implementation of STACS would require feedback connections that convey the control settings—the spatial scale at which the model is operating and the set of tasks that are enabled. These control settings modulate ongoing processing, specifically they enable or disable different neural pools. In contrast, a model such as SAIM (Heinke & Humphries, 2003) suggests a very different view of cortical feedback. SAIM uses top-down connectivity to obtain interactive constraint-satisfaction dynamics; as a result, the feedback connections are integral to object recognition. STACS's notion of cortical feedback as modulating basically bottom-up processing seems better in accord with current conceptions of cortical dynamics. Nonetheless, the substantial differences between the models should lead to interesting experimental tests.

5 Acknowledgments

This research was supported by NSF BCS 0339103 and NSF CSE-SMA 0509521.

6 References

Averbach, E., & Coriell, A.S. (1961). Short-term memory in vision. *Bell Systems Technical Journal*, *40*, 309–328.

Baldwin, D., & Mozer, M. C. (2006). Controlling attention with noise: The cue-combination model of visual search. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty Eighth Annual Conference of the Cognitive Science Society* (pp. 42-47). Hillsdale, NJ: Erlbaum Associates.

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & Psychophysics, 55*, 485–496.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*, 77–80.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (p. 155–162). Cambridge, MA: MIT Press.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28–71.

Folk C.L., Remington R.W., Johnston J.C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception & Performance, 18*, 1030–1044.

Heinke, D., & Humphreys, G. W. (2003). Attention, spatial representation and visual neglect: Simulating emergent attention and spatial memory in the Selective Attention for Identification Model (SAIM). *Psychological Review*, *110*, 29–87.

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

Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neuronal circuitry. *Human Neurobiology*, *4*, 219–227.

Mozer, M. C. (1991). The perception of multiple objects: A connectionist approach. Cambridge, MA: MIT Press.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*, 205–231.

Neider, M. B., Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*, 614–621.

Peelen, M. V., Heslenfeld, D. J., & Theeuwes, J. (2004). Endogenous and exogenous attention shifts are mediated by the same large scale neural network. *NeuroImage, 22*, 822–830.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance X* (pp. 531–556). Hillsdale, NJ: Erlbaum.

Proulx, M. J. & Egeth, H. E. (2006). Target-nontarget similarity modulates stimulus-driven control in visual search. *Psychonomic Bulletin & Review*, *13*, 524–529.

Rosen, A.C., Rao, S.M., Cafarra, P., Scaglioni, A., Bobholz, J.A., Woodley, S.J., Hammeke, T.A., Cunningham, J.M., Prieto, T.E., Binder, J.R. (1999). Neural basis of endogenous and exogenous spatial orienting: a functional MRI study. *Journal of Cognitive Neuroscience*, *33*, 135–152.

Theeuwes, J. (2004). Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review*, *11*, 65–70.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, *113*, 766–786.

Vecera, S. P., Vogel, E. K., & Woodman, G. F. (2002). Lower region: A new cue for figure-ground assignment. *Journal of Experimental Psychology: General*, *131*, 194–205.

Wolfe, J.M. (1998) Visual memory: What do you know about what you saw? *Current Biology*, *8*, R303–R304.

Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, *1*, 202–238.

Wolfe, J. M. (2007). Guided Search 4.0: Current progress with a model of visual search. In. W. Gray (Ed.), *Integrated Models of Cognitive Systems*. NY: Oxford.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 419–433.