An integrative, experience-based theory of attentional control

Matthew H. Wilder
Department of Computer Science, University of Colorado, Boulder, CO, USA, &
Institute of Cognitive Science, University of Colorado, Boulder, CO, USA

Michael C. Mozer
Department of Computer Science, University of Colorado, Boulder, CO, USA, &
Institute of Cognitive Science, University of Colorado, Boulder, CO, USA

Christopher D. Wickens
University of Illinois, USA,
University of Colorado, USA, &
MA & D Operations, Alionscience, USA

Although diverse, theories of visual attention generally share the notion that attention is controlled by some combination of three distinct strategies: (1) exogenous cuing from locally contrasting primitive visual features, such as abrupt onsets or color singletons (e.g., L. Itti, C. Koch, & E. Neiber, 1998), (2) endogenous gain modulation of exogenous activations, used to guide attention to task-relevant features (e.g., V. Navalpakkam & L. Itti, 2007; J. Wolfe, 1994, 2007), and (3) endogenous prediction of likely locations of interest, based on task and scene gist (e.g., A. Torralba, A. Oliva, M. Castelhano, & J. Henderson, 2006). However, little work has been done to synthesize these disparate theories. In this work, we propose a unifying conceptualization in which attention is controlled along two dimensions: the degree of task focus and the contextual scale of operation. Previously proposed strategies—and their combinations—can be viewed as instances of this one mechanism. Thus, this theory serves not as a replacement for existing models but as a means of bringing them into a coherent framework. We present an implementation of this theory and demonstrate its applicability to a wide range of attentional phenomena. The model accounts for key results in visual search with synthetic images and makes reasonable predictions for human eye movements in search tasks involving real-world images. In addition, the theory offers an unusual perspective on attention that places a fundamental emphasis on the role of experience and task-related knowledge.

Keywords: attention, computational modeling, eye movements

Citation: Wilder, M. H., Mozer, M. C., & Wickens, C. D. (2011). An integrative, experience-based theory of attentional control. *Journal of Vision, 11*(2):8, 1–30, http://www.journalofvision.org/content/11/2/8, doi:10.1167/11.2.8.

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; Treisman, 1982) and theoretical perspectives (e.g., Itti & Koch, 2000; Julesz, 1984; Koch & Ullman, 1985; Neisser, 1967). 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. In general, the saliency of a location expresses how much that location stands out from its spatial or temporal neighborhood in terms of its primitive visual features.

Attention need not be deployed in a purely exogenous manner but can be influenced by task demands (e.g., Bacon & Egeth, 1994; Folk, Remington, & Johnston, 1992; Wolfe, Cave, & Franzel, 1989). The ability of individuals to attend based on features such as color or orientation has led to theories proposing *feature-based endogenous* control (e.g., Mozer, 1991; Mozer & Baldwin, 2008; Navalpakkam & Itti, 2007; 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.

Experimental studies support a third attentional control strategy in which attention is guided to visual field regions

doi: 10.1167/11.2.8 Received March 23, 2010; published February 9, 2011 ISSN 1534-7362

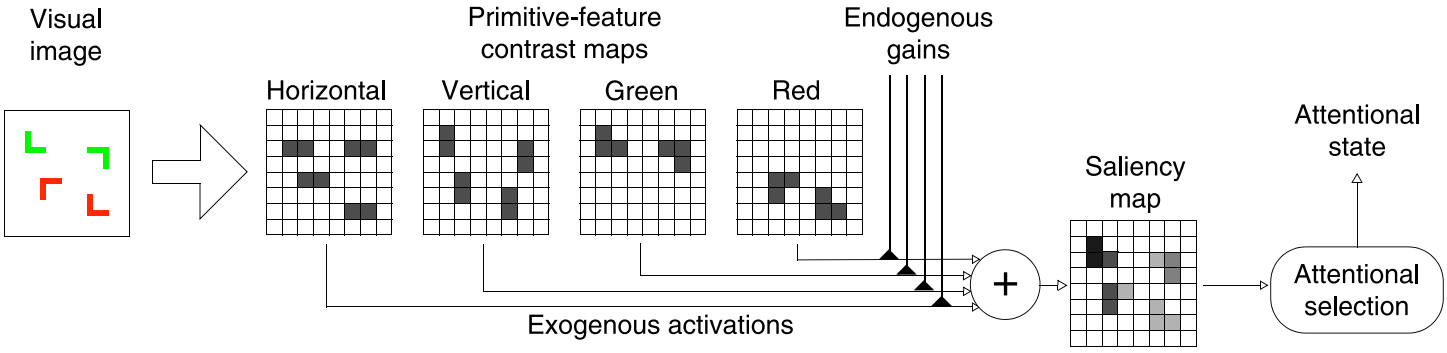


Figure 1. A depiction of feature-based endogenous control. The final saliency map is obtained by taking a weighted combination of the primitive feature maps. The weights are used to highlight relevant feature channels.

likely to be of interest based on the current task and global properties of the scene (e.g., Biederman, 1972; Neider & Zelinsky, 2006; Siagian & Itti, 2007; Torralba, Oliva, Castelhano, & Henderson, 2006; Wolfe, 1998a). 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, one can infer its gist (Oliva & Torralba, 2001) and this gist can guide attention.

A unifying framework

Instead of conceiving the three control strategies as distinct and unrelated mechanisms, a key contribution of this work is to characterize the strategies as components of a broader *control space*. This control space represents a continuous range of potential strategies that all share the same underlying mechanism but are tuned along two dimensions to yield different behavior. As depicted in Figure 2, the two dimensions of the control space are *task specificity* and *contextual scale*.

Task specificity refers to the degree to which control exploits the current tasks and goals. High task specificity refers to situations in which the attentional system is strongly constrained by the nature or properties of the task. Low task specificity occurs in situations where the individual has no particular goal or when attention operates in a task-independent manner. In this paper, we focus on visual search, and consequently, tasks can be defined in the currency of objects—i.e., a search for a specific object or a class of objects. Within the context of visual search, high task specificity refers to search for a particular object and low task specificity refers to exploration in the absence of a particular goal.

In the control space, the contextual scale dimension is analogous to the granularity of image processing. At the feature-level contextual scale, small regions in the image are processed separately and saliency predictions for one region are roughly independent of the predictions for a different region. Conversely, processing at a scene-level contextual scale is more holistic—general scene properties extracted from the entire image are used to make saliency predictions across the whole field of view. In the middle of the continuum, saliency predictions are derived from all features within a region roughly large enough to contain a whole object but not the whole scene.

The two dimensions of the control space are helpful in explicating the similarities and differences between the three control strategies we have described. Exogenous control appears in the lower left corner of the space in Figure 2 because it operates independently of current goals and uses a feature-level contextual scale. Feature-based and scene-based endogenous controls are placed in the upper region of the space because both operate with a high degree of task specificity. However, they reside at different points

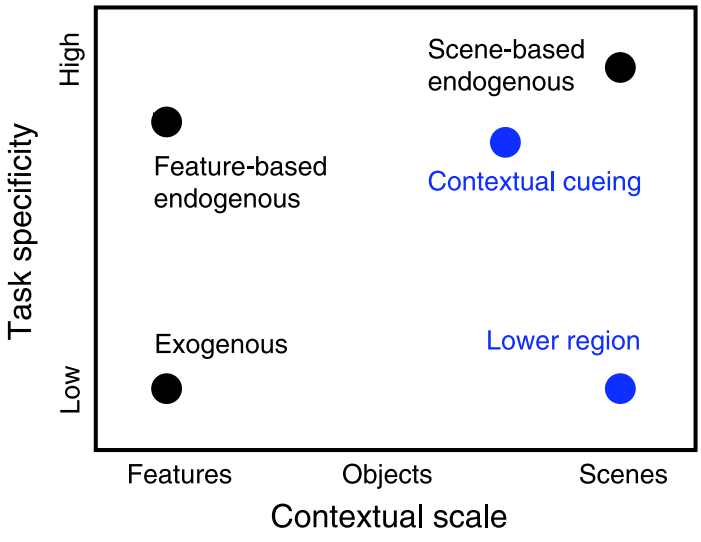


Figure 2. A two-dimensional control space that characterizes exogenous, feature-based endogenous, and scene-based endogenous control of visual attention (black points). The blue points represent other attentional phenomena that have a natural interpretation within this framework and are discussed later.

along the contextual scale continuum. Feature-based endogenous control classifies a small region as salient if specific features are present in that location. Scene-based endogenous control uses the whole image to determine the gist of the scene. When searching for a specific object or group of objects, the gist is then used to differentiate regions of the scene by relevance.

What benefit is there to laying out the three strategies in this control space? The control space offers us the means to reconceptualize attentional control by illuminating the relationships among the various control strategies—highlighting the dimensions along which one strategy is similar to and different from another. In doing so, the control space suggests additional strategies that might be employed by attention. We return to this issue later, when we recharacterize other attentional phenomena in terms of the control space.

Combining control strategies

Given the existence of at least three distinct control strategies, one might ask how the strategies are intermixed. A simple hypothesis is that control processes operate at a single point in the control space, and that point is selected so as to optimize performance in a specific environment. If the selected point was in between two of the primary strategies depicted in Figure 2, it might appear that multiple strategies were operating simultaneously. Alternatively, multiple strategies might be employed in parallel, and attention would be governed by some combination of or arbitration among the strategies.

Recently, several hybrid models have been proposed that incorporate multiple strategies operating in parallel. Torralba et al. (2006) present a model—which we refer to as *TOCH*—that integrates object-specific scene-based endogenous control with standard exogenous control. More recently, TOCH was expanded to include target-specific saliency by Ehinger, Hidalgo-Sotelo, Torralba, and Oliva (2009). Navalpakkam and Itti (2007) combine exogenous with feature-based endogenous control in a model we refer to as *NI*. In the upgrade of Guided Search (GS) to the current 4.0 version, Wolfe (2007) added bottom-up exogenous activation and scene-gist processing to the existing feature-based endogenous control. Siagian and Itti (2007) propose an architecture—which we call *SI*—that computes gist and exogenous saliency in parallel using the same biologically plausible visual features. The exogenous component drives the saliency map but can be aided by the gist information.

Most of these hybrid models take an engineering approach to attention—stacking different components together to improve performance. In contrast, we present a theoretical approach that wraps most of the functionality of these hybrid models into one consistent framework that encompasses any combination of strategies in the control space. From the perspective of the control space, these hybrid models combine task-specific processing at a scene-level contextual scale with some form of processing at a feature-level contextual scale. These models are therefore focused on several points in the control space; in contrast, our framework is capable of exploiting the whole space of control strategies. Thus, we view this framework as a generalization of hybrid models such as TOCH, NI, GS, and SI.

In order to implement combinations of control strategies, we must be able to specify a control strategy at any point in the space. Previous models offer some guidance as to how to implement the primary control strategies, which lie near corners of the space in Figure 2. An immediate challenge, however, is to characterize the interior of the control space—what it means for a control strategy to operate at an intermediate contextual scale and with an intermediate degree of task specificity. Embedded in this challenge is the difficulty of relating the primary control strategies such that they share the same processing mechanisms. Parameterizing a model over a continuum of contextual scales seems feasible by varying the granularity of image processing, but how would varying degrees of task specificity be implemented? To answer this question, it is useful to digress and consider the role of experience in attention. By illuminating the relationship between attentional strategies and experience, we formulate a novel hypothesis that strongly intertwines attention and experience. This hypothesis offers a natural way to represent varying degrees of task specificity.

Experience-based attention

Consider the influence of experience on the three primary control strategies. Exogenous control is typically envisioned as a hardwired bottom-up process that is independent of an individual's experience (see Zhang, Tong, Marks, Shan, & Cottrell, 2008, for an alternative view). Experience plays a larger role in feature-based endogenous control: task instructions or knowledge can modulate bottom-up processes to amplify task-relevant features. Nonetheless, representations in the attentional system are still considered fixed and independent of experience. Scene-based endogenous control places a greater emphasis on experience. Performance in a particular environment—e.g., city streets, kitchens—depends on associations between coarse image properties and scene types learned through experience.

These descriptions suggest that the influence of experience on attentional control varies depending on the specific strategy employed. In contrast to this classical view, we suggest here that all attentional control strategies can be formally characterized as experience dependent.

We motivate this perspective with the observation that experience appears to have an effect on attentional processes that are typically thought of as exogenous. For example, in figure–ground assignment, Peterson and Skow-Grant (2003) find that the choice of figure is biased by past

experience. In the domain of real-world images, Cerf, Frady, and Koch (2008) and Cerf, Harel, Einhauser, and Koch (2008) offer further support for the role of experience in attention by comparing a standard exogenous model of attention with an extended model that incorporates a face detection component and other object recognition models. The authors find that the addition of these components significantly improves the correspondence between model predictions and free-viewing human eye movements even for images that do not contain any of the relevant objects.

Another argument for the fundamental role of experience in attention comes from adaptation effects. Senders (1964) shows that fixation frequency to components in an instrument panel corresponds to the bandwidth (events per second) of that component. More recently, Geng and Behrmann (2005) find that locations with high spatial probability—i.e., regions where more activity occurs—are more likely to be selected by attention. In their *contextual cuing* paradigm, Chun and Jiang (1998) show that individuals can learn to predict target location contingent on the spatial configuration of elements in a display. In an interesting study combining chess and contextual cuing, Brockmole, Hambrick, Windisch, and Henderson (2008) show that experience with chess significantly affects attentional performance on the search task. Although the fact that attention *can* be adaptive does not imply that attention *always* adapts, the results summarized in this section suggest the parsimonious view that attentional processes constantly adapt to the ongoing stream of experience, and therefore, that attention should be viewed as fundamentally knowledge or experience based.

Reconceptualizing the control space

Adopting the perspective that all varieties of attentional control are fundamentally experience dependent, we return to consider our control space (Figure 2) and, in particular, the task specificity dimension. A control strategy that has a high degree of task specificity necessarily requires experience with the particular task. However, control strategies that have low task dependence (e.g., exogenous control) have traditionally been viewed in terms of hardwired bottom-up and experience-independent mechanisms. Here we propose instead that exogenous control has a dependence on a broad range of past experience, i.e., it depends on *every* task. A pure exogenous control strategy would thus be cast as "identify as salient locations that are interesting based on the combination of all past experience on a wide variety of learned tasks." This novel perspective is an important component of this work and distinguishes our model from existing models.

With our focus on visual search, each specific task corresponds to a target object, e.g., car keys, wallet, car, traffic light. The task specificity dimension can thus be recast in terms of the size of the subset of objects that guide attention—with decreasing specificity as the set size increases. A high degree of task specificity might correspond to a single target object, e.g., a car or a person. An intermediate degree of task specificity might be thought of as search for a more general class of objects, e.g., the set of all objects that are wheeled vehicles or living things. At the low task specificity end of the spectrum, attention is guided using all objects with which one has had experience. Defining exogenous attention in this manner is an extension of the concept presented in Cerf, Frady et al. (2008) and Cerf, Harel et al. (2008). This generalization implies that pure bottom-up attention does not exist as a separate mechanism but is rather a special case of what has traditionally been viewed as top-down attentional control. In line with this claim, Folk et al. (1992) have found that involuntary attention capture is contingent on attentional control settings.

With our reconceptualization of task specificity, we can describe any control strategy via a set of modules that specialize in particular targets. The modules must also be characterized in terms of the other dimension of the control space—the contextual scale at which they operate. This set of modules can be cast in terms of a dual to the control space, which we will call the *module grid*, depicted in Figure 3. Like the control space, the module grid has two dimensions, but the dimensions focus on implementation. The rows of the module grid enumerate specific targets of attention, and the columns specify an analog of contextual scale, which we call *range of influence* and will explain shortly. Each cell in the grid is a particular processing

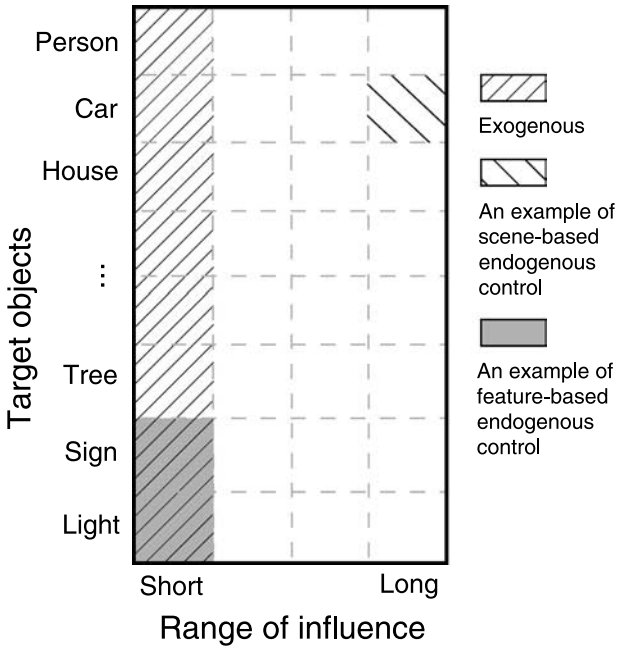


Figure 3. A depiction of the *module grid*, which is a transformation of the control space that emphasizes specific implementable attentional mechanisms. Each control strategy—which corresponds to a point in control space—can be understood as a collection of grid squares (modules) in the module grid.

mechanism that produces a saliency map given a visual image.

Control strategies, which occupy a single point in control space (Figure 2), can occupy regions in module grid. Figure 3 displays how the primary strategies in the control space map onto the module grid. Exogenous attention corresponds to a short range of influence and a combination across all target objects. Feature-based endogenous attention also operates with a short range of influence, but only those targets that have the defining features are employed. Scene-based endogenous attentional guidance utilizes a long range of influence and incorporates a small set of targets.

The second dimension of the module grid, range of influence, elucidates how a continuum of contextual scales can be implemented. We describe this dimension in terms of the spatial range of influence that a visual feature in the image has on the saliency map. A short range of influence, which corresponds to a feature-level contextual scale, implies that saliency map activation at a given location is determined only by nearby visual features. Conversely, a long range of influence implies that activation anywhere in the saliency map can be influenced by features anywhere in the visual field, i.e., the whole scene. Emphasizing the role of experience, the actual range of influence a visual feature will have depends on statistics of the task environment and past experience: even if the potential connectivity is present for long-range influences, the realization of these influences will depend on the particular task environment, determined through experience.

Through experience, each module becomes specialized for a particular target. We have shown that each primary strategy can be implemented as a subset of modules in one column of the module grid. Similarly, the module grid can represent any combination of strategies in the control space if multiple modules operate in parallel and their saliency maps are merged. We call this framework *TASC*, an acronym on *TAsk-Specific and Contextual-scale control*. Rather than viewing TASC as an alternative theory to existing models such as TOCH, GS, and SI, we view TASC as a generalization of these earlier theories, and each of these theories can be seen as a specific instantiation of TASC.

An illustration of saliency over the module grid

To give a concrete intuition about the operation of TASC, we present an example illustrating the model's behavior. The input to TASC is an image—natural or synthetic—and the output from TASC is a saliency map. Each module in the TASC module grid yields a saliency map associated with a specific target object and range of influence as shown in Figure 4. This example shows a street scene and the saliency maps for three different objects: people, trees, and sidewalks. In its full implementation, TASC would maintain representations for a large collection of objects. At the short range of influence, saliency maps generally make fine-grained predictions. These maps are similar to those obtained by feature-based endogenous models like GS 2.0. In contrast, the saliency maps from modules with a long range of influence show coarse regions where the object is likely to appear.

The saliency maps associated with the module grid are only an intermediate representation in TASC. The final output of TASC—a single saliency map—is obtained by selectively combining module saliency maps. As mentioned above, this selective combination allows for more advanced attentional strategies. Figure 5a presents a simple example of a combined saliency map that could result from TASC. This map is a combination across 11 objects and all ranges of influence. Figure 5b shows separate saliency maps for each of the 4 ranges of influence where a combination is performed across all 11 object modules. The map at the short range of influence corresponds to the exogenous control strategy.

TASC implementation

As with most models of attention, TASC 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 would not be much use for the saliency map because—according to the traditional early selection view of attention—the purpose of the saliency map is to determine how to focus the limited processing resources available for performing object recognition.

In TASC, computation of the saliency map is parallelized across the components of the module grid as depicted in Figure 6. Each TASC module is configured to perform an association from the retinotopic image to the saliency map. The modules all have the same architecture, though they are parameterized to implement varying ranges of influence and to be specialized for specific target objects. To obtain the single final TASC saliency map (bottom of Figure 6), the individual module maps are selectively combined.

Module implementation

A detailed description of the implementation of a module is presented in Appendix A, and we summarize here.

Patch processing

A specific range of influence is implemented in TASC by dividing the image into overlapping patches of a specific size. Each image patch is processed independently and

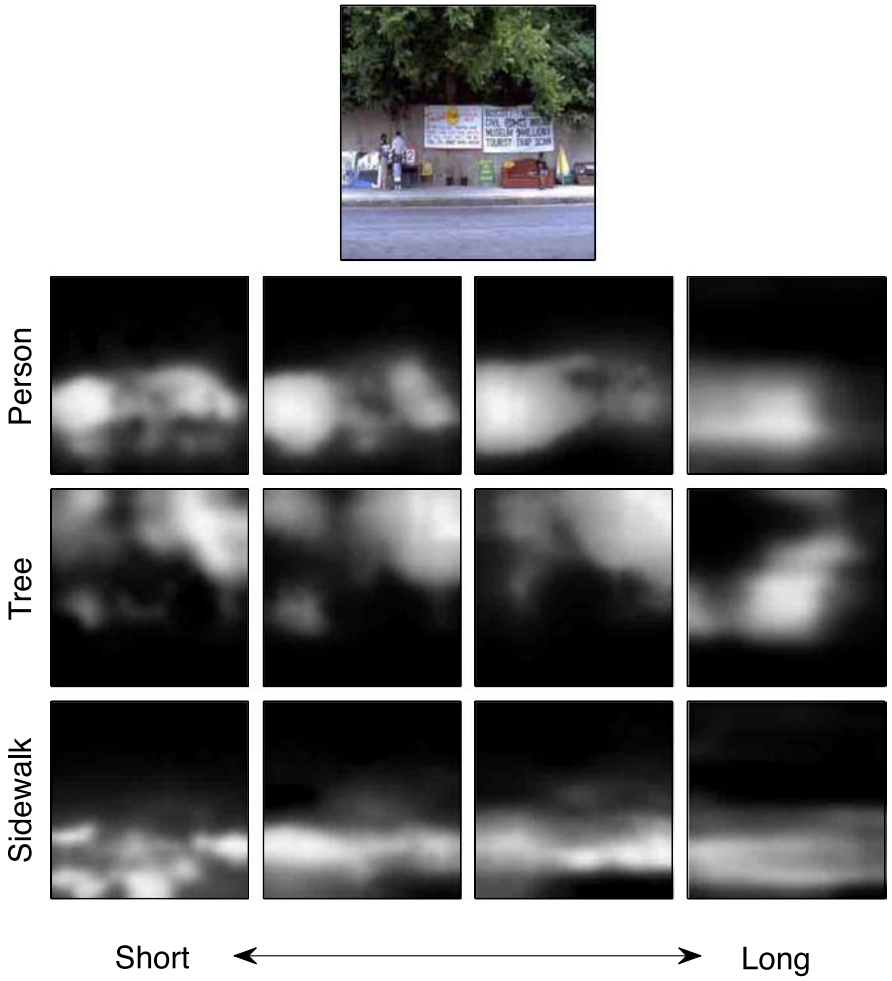


Figure 4. A street scene and saliency maps produced by TASC in response to this input for three different target objects and four ranges of influence.

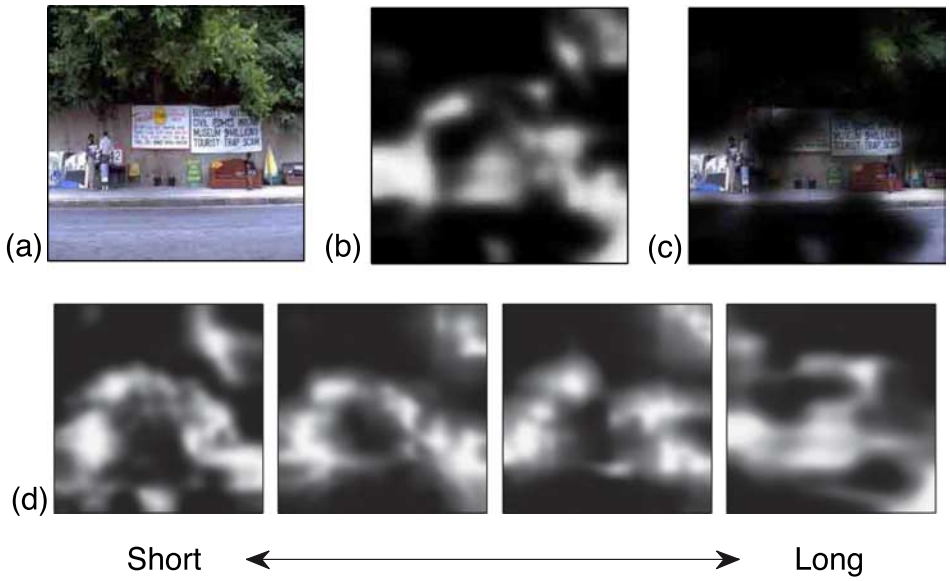


Figure 5. A saliency map (b) for the image in (a) using the naive control strategy in TASC—combine all maps in the module grid. (c) The saliency map overlaid on the original image. (d) Separate saliency maps for the 4 ranges of influence where all object modules are combined. These results come from a simulation with 11 different target object representations.

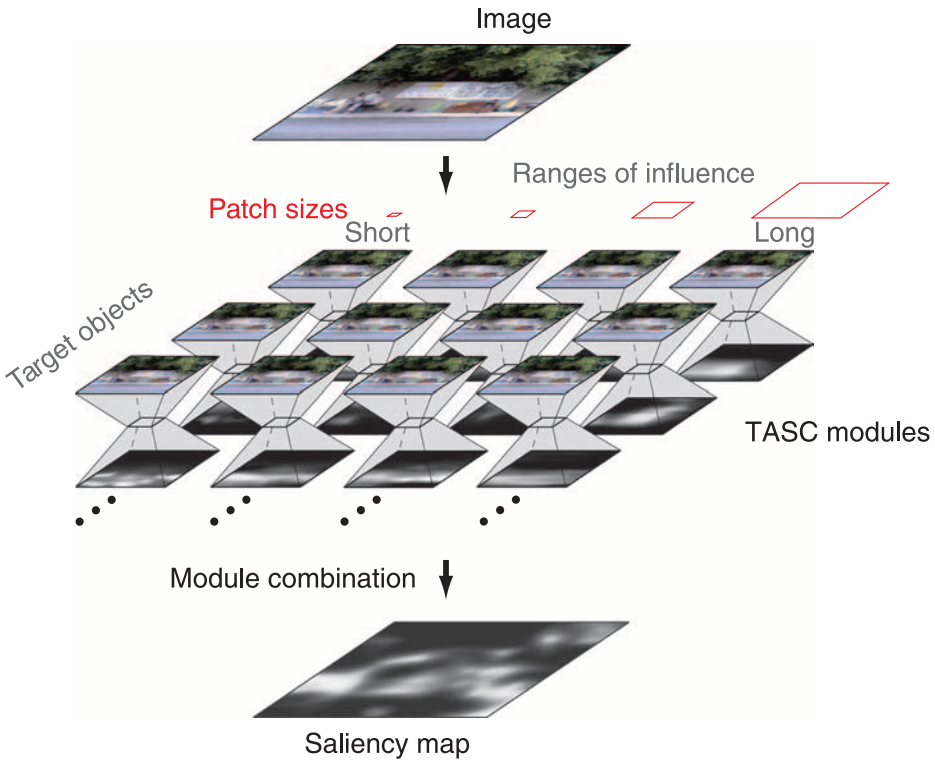


Figure 6. A schematic depiction of TASC. The primary components of TASC are the modules that correspond to specific target objects and ranges of influence. Each module processes the entire image independently and obtains a saliency map unique to that module. The final saliency map is computed by combining across all module saliency maps. This depiction shows 4 ranges of influence and 3 sample target objects. However, there could be more ranges of influence and certainly would be many more object modules.

contributes only to the patch in the saliency map at the same location. For modules with a short range of influence, small patches are used so that saliency predictions are based only on local features—i.e., those within the small patch. At the longest range of influence, the entire image is treated as a single patch that enables image features to have an effect on all saliency locations no matter how distant. The short range-of-influence modules correspond well to the NI model, whereas the long range-of-influence modules are more similar to the global gist processing component of TOCH. The red boxes in Figure 6 display the patch sizes for the 4 ranges of influence, which cover 1.56%, 6.25%, 25%, and 100% of the image. For all ranges of influence, the patches overlap by 50%.

Image processing stages

Rather than creating yet another model of attention, our aim in developing TASC was to generalize existing models such that existing models can be viewed as instantiations of the more general TASC framework. The module grid (Figure 3) encompasses the primary control strategies, and consequently, it generalizes the models that implement these strategies. To complete this generalization, the image processing stages of TASC modules were chosen to overlap with those of many existing models, including NI, GS, TOCH, and SUN. Table 1 lays out the standard processing stages found in many models and offers a comparison between TASC and four previously mentioned models. In NI, GS, and TOCH, color and luminance features are derived from the raw pixel values and orientation features are obtained using Gabor filters. SUN extracts features qualitatively similar to contrast-enhanced color, orientation, and luminance features using a set of independent component analysis (ICA) filters. Dimensionality reduction is used in TOCH and SUN because both models process larger regions of the image.

In TASC, the processing of each patch begins with a feature extraction stage where red, green, blue, and yellow color opponency values and 4 Gabor orientation features are obtained for each pixel location in the patch. Next, the values in each feature channel are compared to neighboring values in the same channel to yield center–surround contrast enhancement. Each feature activation value is scaled by the proportion of regional activation for the relevant feature channel that can be attributed to nearby locations. This results in the suppression of feature activations in homogenous image regions. Contrast enhancement is followed by two dimensionality reduction stages that eliminate redundancy and shrink the feature data to a computationally manageable size. A preliminary reduction is obtained by subsampling the feature data, averaging neighboring values within each feature channel. Subsampling is followed by applying Principal Components

Stages		NI	GS	TOCH	SUN	TASC
Feature extraction	Color	X	X	X	X	X
	Orientation	X	X	X	X	X
	Luminance	X			X	
Contrast enhancement	Center–surround	X	X		X	X
Dimensionality reduction	Subsampling			X		X
	PCA			X	X	X

Table 1. Processing stages of TASC compared to four existing models of attentional control.

Analysis (PCA) to extract only the most important feature properties of the patch. PCA is performed separately for the eight feature channels with an increasing number of components retained as the range of influence becomes longer. Although more principal components are retained at the longer ranges of influence, the total amount of feature information preserved across the whole image is greater for the shorter ranges of influence because there are far more patches. The amount of information retained for the entire image is chosen to be roughly consistent with NI and TOCH, which reside approximately at the two ends of the range of influence spectrum. A different PCA projection is learned for each feature channel and each range of influence from a database containing roughly 2,500 real-world and synthetic images. For a given feature channel and range of influence, though, the same projection is used for all image patches thus yielding a location-independent transformation.

Task-specific associative memory

To specialize modules for particular tasks, each module includes a task-specific associative memory that maps from the visual representation described in the previous section to an activation map that indicates the presence of an object.

NI, GS, TOCH, and SUN all use some sort of task-specific association to obtain saliency maps. In NI, linear weights for the individual feature channels are learned from a set of test images related to the search task. Similarly, GS uses a set of predefined gains to modulate the weight of each feature channel. TOCH achieves a task-specific activation by associating global features with likely object locations through a weighted sum of linear regressors where the weights depend on the global features in a nonlinear way. SUN uses a support-vector machine to make a target present/absent classification based on the image features. Though these approaches differ in their method, they all share the property that saliency is assigned according to some primarily linear function of the image representation.

In TASC, the associative memory is implemented as a set of neural networks distributed across the patches in the image. For each patch, there is set of input units to the network representing the patch data after dimensionality reduction. Each set of input units is fully connected to a set of output units via a layer of hidden units. The hidden units, however, are linear, and their purpose therefore serves to limit the rank of the mapping from input to output. The final patch saliency map is obtained by passing the network outputs through a logistic squashing function. To obtain the complete module saliency map, outputs from the individual patches are averaged where they overlap. This module saliency map is then smoothed via convolution with a Gaussian kernel to reduce artifacts due to patch edges and to help assure consistency across neighboring saliency values.

Because the shorter ranges of influence have far more patches than the longer ranges of influence, we make the hidden-layer bottleneck smaller at the shorter ranges, just as we chose a smaller number of principal components at the shorter ranges of influence. Our choice of rank obtains a roughly equal number of descriptive features per pixel across the four ranges of influence.

The linearity of the association network is important because it restricts the complexity of the mappings that can be learned. As a result, the model can at best perform a quick-and-dirty sort of object detection. We contrast this type of processing with the more elaborate, detailed, and certainly nonlinear processing required for full-blown object recognition. This distinction is key from our perspective, because without it, the roles of attention and object recognition highly overlap: both have the goal of determining where target objects are in an image. From the perspective of an experience-based theory of attention, the role of attention is to do a rapid, roughly parallel analysis of the visual field in the service of a goal, and the role of object recognition is to do a more thorough but resource-limited analysis of locations likely to contain goal-relevant information.

For task-specific network training, we use supervised learning, where the training data come from the LabelMe image database (Russel, Torralba, & Murphy, 2008), a large collection of images labeled by pixel according to the object present at that corresponding location. During training, each image is presented to the model, the model computes a dimensionality reduced feature representation for each patch, and this representation serves as input to the patch-specific neural networks. Each network is trained with a target output containing values of 1 at all locations in the patch where a given target object appears and 0 at all other locations. Training is performed in batches using the

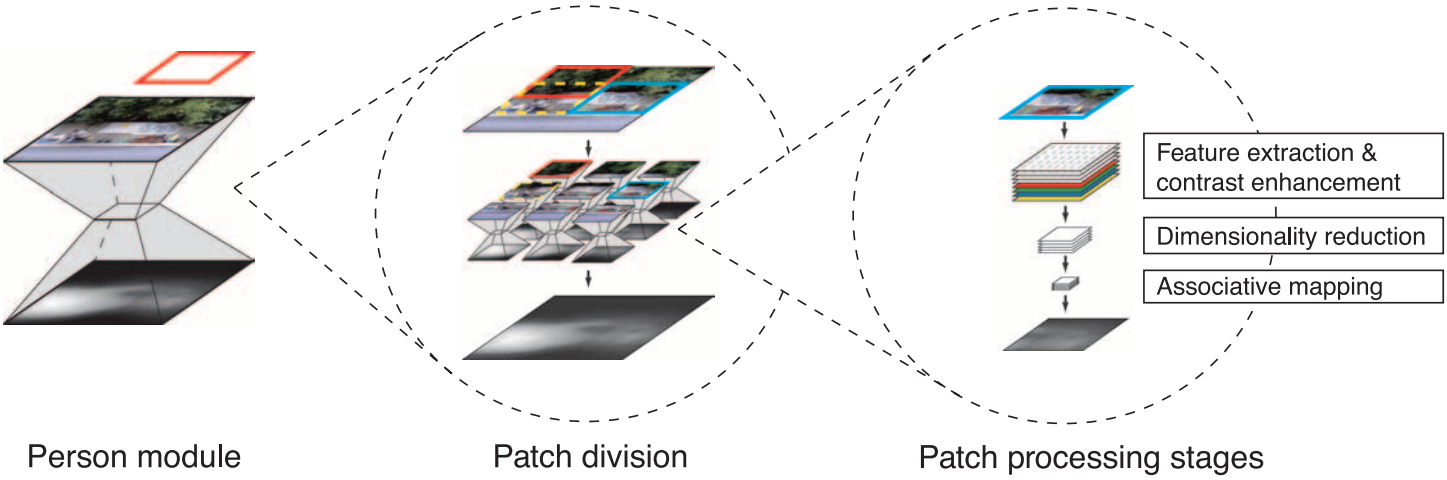


Figure 7. The inner workings of one of the TASC modules shown in Figure 6 (the person module at the second longest range of influence). As the middle enlargement shows, the image is first divided into overlapping patches with a patch size that depends on the range of influence. Each patch is processed independently to obtain a patch-specific saliency map. These overlapping patch maps are then averaged to form the module saliency map. The right side of this figure depicts the image processing stages and the associative mapping applied to each patch.

backpropagation algorithm to update weights and continues until a local optimum in error is obtained.

Summary of module implementation

The inner workings of a TASC module are summarized in Figure 7, which depicts the person module at the second longest range of influence. As shown in the center panel, the image is divided into overlapping patches that are each processed independently and contribute only to their corresponding region in the final saliency map. A saliency map for each patch is obtained by applying a learned associative mapping to the image representation derived from the initial pixel values through feature extraction, contrast enhancement, and dimensionality reduction (right panel of figure).

Combining multiple modules

So far we have discussed how each module in the module grid computes a saliency map for an input image. Each individual module could be used to guide attention. However, to implement exogenous control or even more complex strategies, such as hybrid models like TOCH, it is necessary to combine the saliency maps from a collection of modules. It is this combination of modules that gives the TASC framework its ability to model attentional guidance that varies in task specificity and utilizes multiple contextual scales.

We explored two rules for combining module saliency maps: averaging across modules and taking the maximum output across modules. In our implementation, we chose the max rule because it has a rich history in the literature on visual information processing (Gawne & Martin, 2002; Lampl, Ferster, Poggio, & Riesenhuber, 2004; Riesenhuber & Poggio, 1999; Yu, Giese, & Poggio, 2002; Zhaoping & May, 2007; Zhaoping & Snowden, 2006). Through simulations, we also found that the max rule produces saliency maps that are more consistent with experimental results.

Another important aspect of module combination is the decision of which modules to include. For free viewing simulations, i.e., pure exogenous control, in line with our claim that exogenous control draws on all past experiences, all modules are used in the combination. For simulating more constrained visual search experiments, TASC only uses only a subset of modules that are relevant to the search goals. In the current implementation, we avoid making specific assumptions about how modules are selected. It is certainly possible that the attentional system learns a more complex distribution over modules that is used to perform a weighted combination across all modules. This form of attentional learning is interesting and worthy of future investigation but is beyond the scope of the present paper.

Modules trained for simulations

For all the simulations presented in the Simulations of TASC section, we use one parameterization of the model that includes a collection of modules chosen to allow testing with a wide range of attentional tasks. This versatility is achieved by training a diverse set of object modules using the 4 ranges of influence presented earlier. The simplest object modules used are those trained for a single feature: red, green, blue, vertical, and horizontal. These modules are trained from a set of 500 images containing vertical and horizontal bars of different colors with labels appropriate to the target of interest. In addition

to these simple modules, we use real-world images from the LabelMe database to train modules for the following objects: car, person, bike, sidewalk, road, building, tree, window, light, head, sign, mug, and painting (with roughly 200–800 images per object). The model parameters were chosen a priori based on existing models and our intuitions about sensible information bottlenecks at each stage of processing. We wish to emphasize that these parameters remain fixed for all simulations. As a result, the only variability from one simulation to the next is the selection of relevant modules for the specific attentional task. Though in some cases this generality inhibits TASC's performance relative to other attentional models, we feel that this is balanced by the wide range of phenomena we explain with one fixed model.

Simulations of TASC

Having described our implementation of TASC, we present simulations to demonstrate TASC's flexibility in accounting for a diverse set of findings from the attentional literature. Our goal in presenting simulation results from TASC is to demonstrate the breadth of the model. TASC provides a coherent, integrative framework for interpreting a wide range of attentional phenomena, phenomena that have previously been addressed by disparate theoretical frameworks.

We begin with several simulations that confirm TASC's ability to use the primary control strategies that motivate the control space. That TASC successfully accounts for the human data in these tasks is not too surprising given that the implementation of TASC was chosen to be consistent with previous models that also correctly capture human behavior. Still, there is value in these simulations because they show that one single framework is capable of accounting for diverse types of human behavior—few models have been subjected to as wide a range of attentional tasks. From here, we explore other regions of the control space and ask how they relate to various types of attentional behavior. We find that TASC yields a natural explanation for phenomena such as contextual cuing and figure–ground assignment that do not naturally fall out of other models. Furthermore, TASC's dependence on experience conforms well with results in spatial probability cuing tasks and with sequential effects found in attention.

Because of the range of experimental paradigms and tasks we model, it is worth noting again that the same TASC parameters are used for all simulations; the only difference from one simulation to the next is the subset of modules that contribute to the saliency map. TASC can handle both real-world images and simple displays of the sort used in psychological experiments. In contrast to some attention models in the psychological literature, TASC starts not with an abstract data structure representing presegmented features but with raw pixel images. This same input representation is used for both artificial and real-world tasks.

Visual search

Visual search tasks require that an individual detect a target element in a display surrounded by distractor elements. Visual search is one of the most extensively studied tasks in the psychological literature. (See Wolfe (1998b) for a meta-analysis of a wide range of visual search experiments and Wolfe and Horowitz (2004) for a review of the attributes that guide attention.) The seminal work of Treisman and Gelade (1980) revealed what appeared at the time to be a fundamental dichotomy between feature search and conjunction search. In feature search, the target differs from all distractors along a single-feature dimension; for example, the target is red and the distractors are green (see the display at the left of Figure 8a). In conjunction search, the target is defined by a conjunction of features and shares one feature in common with each distractor; for example, the target is a red vertical bar among distractors that are green verticals or red horizontals (see Figure 8b). Feature search is efficient in that search time is not dependent on the number of distractors in the display; conjunction search is inefficient, due to the increased confusability between targets and distractors. Any account of visual search needs to start by explaining the distinction between feature and conjunction searches. This explanation is a necessary but not sufficient condition for the plausibility of a theory. The dichotomy between feature and conjunction searches has given way to many subtleties and quirks that populate the literature. Our goal in this work is not to explain visual search in all its detail but to demonstrate that TASC is a plausible candidate to tackle the visual search literature.

Both feature search and conjunction search require a feature-based endogenous control strategy because the characteristics of the target are known in advance. In contrast, the oddball detection or pop-out search paradigm requires the use of an exogenous control strategy. In an oddball-detection task, the target differs from distractors along one feature dimension, but the feature dimension and the target feature value are not specified in advance of a trial. For example, the participant might be required to find a red target among green elements, a green element among red, a vertical among horizontals, or a horizontal among verticals (see the display in Figure 8c). The target feature changes from trial to trial and thus part of the task involves determining the relevant discrimination on the current trial. In addition to capturing the dichotomy between feature and conjunction searches, an attentional model must also yield efficient search for oddball-detection tasks.

To model visual search, we construct displays that contain elements varying along two feature dimensions, color and orientation. To assess search performance, we

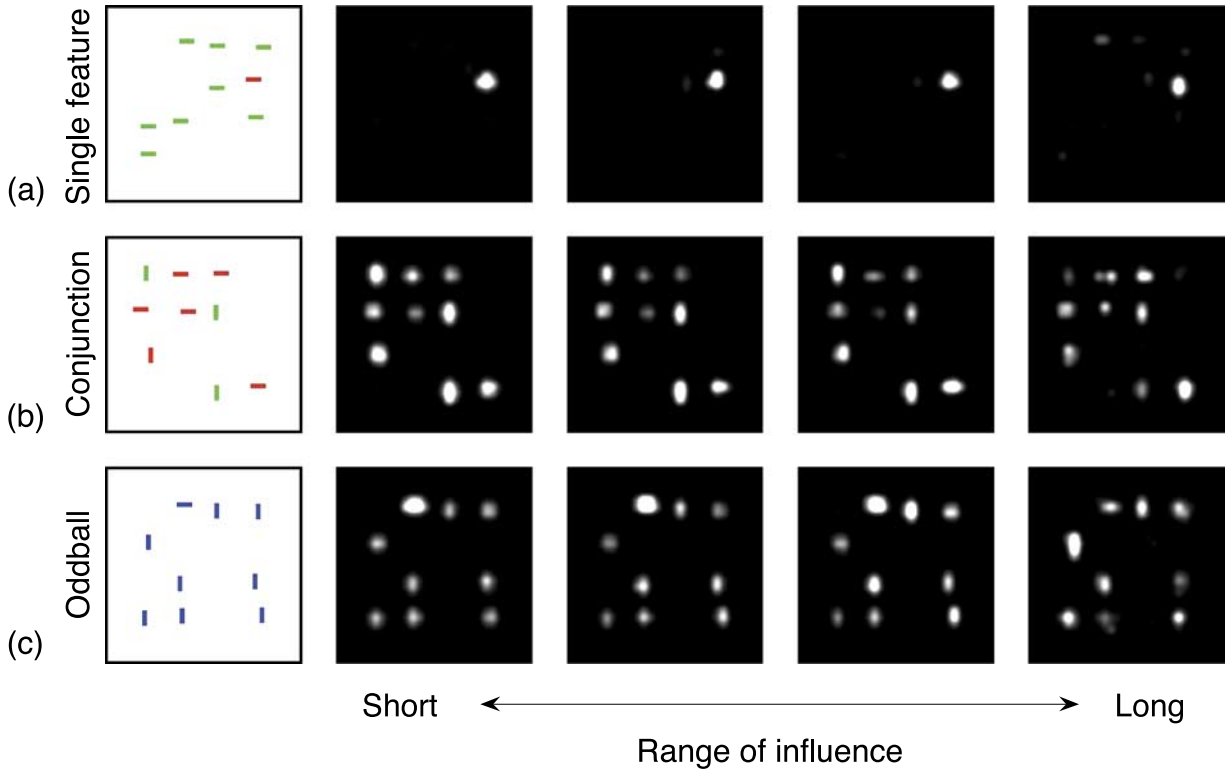


Figure 8. Saliency maps at 4 ranges of influence for three different visual search paradigms: (a) feature search—a single red target among green distractors, (b) conjunction search—a single red vertical among green verticals and red horizontals, and (c) oddball search—a singleton exists in each image, but the specific features of the target and the defining feature dimension are unknown before the trial.

require a measure of response latency to detect a target as a function of the number of distractors in the display. (The efficiency of search is reflected in the slope of this curve.) How are TASC's saliency maps translated into response latencies? We adopt an assumption from GS 2.0 (Wolfe, 1994) regarding how the saliency map is used in search. GS 2.0 supposes that display locations are examined in order from most salient to least, and that response latency increases linearly with the number of display locations examined. To compute the saliency rank of a display element, the maximum saliency value in the immediate region of the display element is determined, and elements are ranked by saliency. One could embellish TASC to use the response accumulation mechanism of GS 4.0 (Wolfe, 2007) to obtain more realistic response times. However, the response latencies on average would still be monotonic in saliency ranking.

Feature search

Feature search is the paragon of paradigms requiring a feature-based endogenous control strategy—subjects are searching for the presence of one particular feature in an image. In TASC, feature-based endogenous control is achieved by selecting the row of the module grid that corresponds to the feature of interest. As discussed in the TASC implementation section, the model used for all simulations includes modules for 5 basic feature objects: red, green, blue, horizontal, and vertical.

To simulate feature search, TASC is presented with a set of 300 novel images, each of which has exactly one target that has the same property for all images (e.g., the target is always red). The set of images is divided into an equal number of displays with 4, 8, and 12 distractors. TASC yields a saliency map for each image by combining across the four ranges of influence for the appropriate target object row. The left panel of Figure 8a shows a sample display for red-target feature search with 8 distractors and the activation of saliency maps across the four ranges of influence. The location containing the target is highly salient at each range of influence, and no other locations are salient, suggesting efficient search. Clearly, the target location in the combined saliency map will have more activity than any other location. Figure 9a shows the mean target ranking as a function of display size averaged across 5 simulations each with a different defining target feature including red, green, blue, vertical, and horizontal. Each data point corresponds to the average target ranking across 500 displays with the specified number of distractors. Regardless of the number of distractors in the display, TASC obtains a saliency ranking of 1 for the target element, suggesting a fast response time that is independent of display size, consistent with behavioral studies showing efficient feature search.

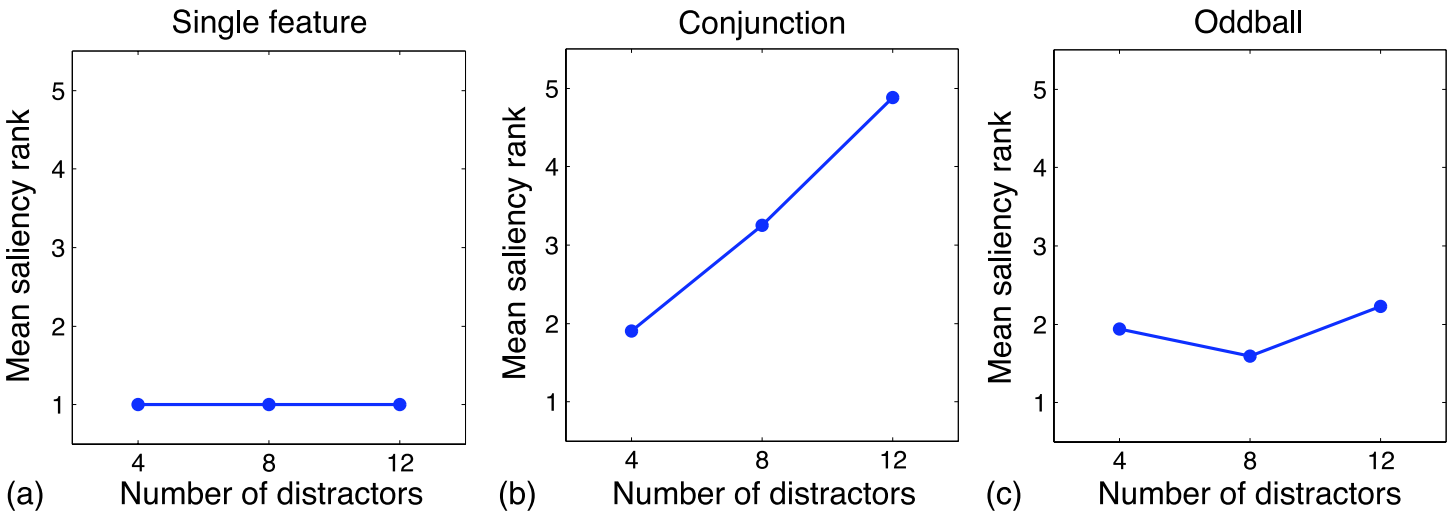


Figure 9. Mean target-saliency ranking for three visual search tasks in TASC as a function of the number of distractors in the display. Response time is assumed to be monotonic in target ranking. (a) Search for a target defined on one feature dimension (e.g., red among green distractors), (b) a classic conjunction search task (e.g., red vertical target among green vertical and red horizontal distractors), and (c) oddball-detection task in which the target is a singleton along one dimension, but the dimension is not specified in advance of a trial.

Feature search is efficient in TASC because the contrast enhancement stage strengthens the target feature value and the neural net only activates regions that contain the target feature. Although TASC is trained to perform feature search, the model is agnostic as to whether this training corresponds to experience within an individual's lifetime or experience on an evolutionary time scale. The main claim of the model is that feature search is not special; it relies on the same processing mechanisms as any other search task.

Conjunction search

To perform single-feature search, TASC employs modules that respond to the primitive feature that defines a target. To perform conjunction search, TASC simply combines the outputs from the two module rows that represent the component features. For example, Figure 8b shows the saliency maps for a red vertical target, obtained by combining the outputs of red and vertical modules (shown here before combining across ranges of influence).

In contrast to the saliency maps for single-feature targets, which show a strong separation between targets and distractors (Figure 8a), the saliency maps for conjunction targets have significant spurious activation. The exact pattern of activity depends on the configuration of elements, because the contrast and dimensionality reduction stages of the model are influenced by local configurations. Figure 9b shows that the saliency rank of the target increases steeply with the number of distractors in the display, suggesting that TASC is unable to reliably detect a conjunction target. This result is consistent with the significant positive search slope found in classic conjunction search experiments.

A key feature of TASC responsible for its inefficiency on conjunction search is the max rule used for combining saliency maps of different targets (here, red and vertical). The max operator yields a roughly 1:1 saliency ratio for targets versus distractors, whereas summing activations from the two saliency maps yields a 2:1 saliency ratio on expectation.

Note that the saliency map activations are not random. If they were random, then the mean rank would, on expectation, be half the number of items in the display. The observed rankings are better, indicating that some information about the conjunction is available to the model. Specifically, the saliency of a red vertical element is, on expectation, higher than the saliency of an element that is just red or just vertical. This is a consequence of the presence of noise in the activations due to configuration and compression effects. For example, if the activations of red and vertical features have a Gaussian distribution with mean 1.0 and standard deviation 0.10, then the max operator should produce, on expectation, a target-to-distractor activation ratio of 1.06. This result occurs because the expected value of the max of two independent and identically distributed random variables is greater than the expected value of a single random variable with the same distribution.

As we explained earlier, the max operator was motivated from the fact that max essentially implements a disjunction, and disjunctions are needed to vary the task specificity as we have defined it. According to the TASC framework, a fundamental dimension of control is this task specificity, and we conjecture that controlling task specificity is evolutionarily more useful than having the capability to perform conjunction search. The architecture is thus optimized for its typical use—searching for objects with

varying degrees of specificity—not the artificial laboratory task of conjunction search.

Thus, we did not simply design inefficient conjunction search into TASC in order to be consistent with behavioral data. Although the model's design could be altered to improve conjunction search, the alterations would have negative consequences to the model's other abilities. Inefficient conjunction search might therefore be a reflection of design trade-offs in the cognitive architecture.

If the same conjunction task was repeatedly practiced over a long period of time, one might expect that a new row of modules would be formed specifically for the conjunction. Just as an individual who regularly searches for people will likely have modules tuned to detecting people, an individual who is always looking for red–vertical bars might have a row of modules customized to objects that are red and vertical. TASC predicts that search would become more efficient as a module row becomes customized to the search object because the max combination across two module rows would be replaced by the more precise response of one row. Consistent with this prediction in TASC, experiments have shown that many search tasks that are inefficient at first become more efficient with extended experience (e.g., Caerwinski, Lightfoot, & Shiffrin, 1992; Leonards, Rettenbach, Nase, & Sireteanu, 2002; Mruczek & Sheinberg, 2005; Sireteanu & Rettenbach, 2000; Steinman, 1987).

One exception, however, occurs with synthetic displays that contain conjunctions of color and orientation (Leonards et al., 2002; Sireteanu & Rettenbach, 2000). Regardless of the amount of experience with conjunction targets of this type, search remains inefficient. Why can a *car* saliency module, for example, be learned to facilitate efficient search, but a *red–vertical* saliency module cannot? The answer is likely rooted in the difference between conjunction-search targets in synthetic displays and objects in real-world scenes. Localization of a synthetic conjunction target requires precise alignment of primitive features. In contrast, features in real-world objects are generally more redundant and their exact alignment is usually not needed for identification. It is likely that the quick and dirty process embodied by TASC for object detection will lose this precise alignment of which color feature goes with which shape feature in a dense neighborhood of many colored shapes. When the alignment of primitive features is lost, the target can no longer be discriminated from the distractors and search becomes inefficient. Though the current implementation of TASC yields some ambiguity in the alignment of features due to dimensionality reduction, efficient search was obtained for conjunction displays when a single module row was trained for the conjunction target. We attribute this performance to the arbitrary selection of patch sizes, dimensionality reduction parameters, and display sizes. A more accurate selection of these settings would likely yield inefficient conjunction search even when a single module row is trained on the conjunction.

Oddball-detection search

In an oddball-detection or pop-out search task, the target is defined only by its contrast with the distractors. On one trial, the target may be the red item among green, and on the next trial, the target may be the vertical among horizontals. The task precludes knowing the target's featural identity in advance of a trial. Consequently, oddball detection is facilitated not by endogenous control but by detecting local contrast for any feature on any dimension. The notion of exogenous attention in TASC is ideally suited for oddball detection. In TASC, exogenous attention is defined as the combination (disjunction) of many object-specialized modules. To simulate the oddball-detection task, TASC was shown a set of novel images, each containing an array of 4, 8, or 12 homogeneous distractors, and a single target having a different value on one feature dimension (color or form). The critical dimension and feature values vary from trial to trial. On each trial, an overall saliency map is obtained by combining across the TASC modules tuned to the 5 primitive feature objects: red, green, blue, horizontal, and vertical.

Figure 8c presents an example of one trial of the oddball-detection task and the resulting saliency maps for each range of influence. The three shortest ranges of influence obtain greater saliency for the target than other locations, as would the combination across the four maps. However, the target–distractor saliency ratio is lower for oddball detection than for feature search. Figure 9c shows the saliency ranking for the oddball-detection simulation. Most importantly, the graph does not show a systematic search slope: the target does pop out regardless of display size. However, as one would expect by examination of the sample saliency maps in Figure 8, the mean saliency ranking of the target is somewhat higher in oddball detection than in feature search, predicting that oddball detection should be slower than feature search. Finding direct evidence of this prediction in the literature is a challenge, but there is indirect evidence (e.g., Lamy, Carmel, Egeth, & Leber, 2006), and it would seem intuitive considering that feature search is more narrowly delineated. Typical theories of exogenous control would not necessarily predict a distinction between feature and oddball searches.

TASC is efficient in oddball detection primarily because when there is an oddball on one dimension (i.e., color or orientation), the contrast enhancement stage will amplify the activity of the oddball target and will suppress activity of the uniform, homogenous surrounding distractors. Because each module's associative network is roughly linear, enhancement of the target early in processing will propagate to enhancement of the saliency activation at the output stage. Because oddball detection is defined in TASC as the combination of all primitive feature modules, the combination of module outputs may have a slight dampening effect on the target's relative salience. This dampening effect is evident in Figure 8c, where all elements of the display have some saliency. Nevertheless, the target generally has the greatest saliency, leading to search times that are independent of the number of distractors.

Real-world images

The previous simulation demonstrated that the notion of exogenous control as defined by TASC yields sensible results for the oddball-detection task. It is natural to question whether TASC's form of exogenous control is appropriate not only for artificial displays used in experimental tasks but also for real-world vision and attention. Figure 5a shows a street scene and the corresponding saliency map, Figure 5b, obtained in TASC by combining across the shortest range of influence modules for all 13 target objects trained with real-world images. Although the result seems reasonable, it is difficult to judge. Consequently, methodologies have been developed to evaluate theories of exogenous control by comparing locations deemed to be salient by a model to locations where individuals tend to fixate during free viewing. Zhang et al. (2008) explore several methodologies and compare their model SUN to the models of Bruce and Tsotsos (2006), Gao and Vasconcelos (2007), and Itti, Koch, and Neiber (1998). They find that SUN and the model of Bruce and Tsotsos (2006) perform similarly and outperform the other two models.

We have chosen not to subject TASC to this formal evaluation, for the following reason. TASC defines exogenous control as the combination across all object models that have ever been useful. The current implementation of TASC—with 13 real-world objects chosen from a fairly arbitrary set based on the availability of labeled data—does not reflect the broad and diverse set of objects to which individuals are exposed. Nevertheless, we expect that a more complete implementation of TASC should perform comparably to SUN because TASC and SUN are based on similar theoretical assumptions about the nature of exogenous control. Like TASC, SUN claims that a lifetime of experience determines which features of the visual world should capture attention. However, the two models make different predictions about how experience is used. Specifically, SUN posits that exogenous attention should be directed to locations containing atypical features—features that are highly unusual based on past experience. In contrast, TASC posits that exogenous attention is based on past experience with specific tasks. It seems certain that experimental tests could be designed to distinguish between these two hypotheses. Although the current implementation of TASC—with only a handful of task-specific modules—precludes a formal evaluation of its hypothesis concerning exogenous attention in real-world scenes, we can certainly evaluate TASC on *specific* tasks in real-world scenes. We turn to this challenge next.

The simulations to this point have demonstrated that TASC is capable of utilizing two of the three primary control strategies discussed in the Introduction section: exogenous and feature-based endogenous. Torralba (2003) suggests that scene-based endogenous control is also a relevant attentional strategy for real-world images. In TASC, scene based-endogenous control corresponds to control that operates with a high degree of task specificity and at a scene-level contextual scale. Now we verify that TASC can integrate scene-based endogenous control by testing the model on a set of images and corresponding human eye-movement data from Torralba et al. (2006). By evaluating the overlap between TASC saliency predictions and human fixations, we demonstrate that TASC matches human eye-movement data more accurately than a simple exogenous model and that it performs roughly as well as TOCH, which combines scene-based endogenous control with exogenous control.

Torralba et al. (2006) perform a rigorous analysis of TOCH by comparing saliency predictions from the model to human eye movements recorded for the same set of images. The experiment in Torralba et al. involved three different search tasks over a set of images that contained indoor and outdoor scenes. For each target object—person, mug, and painting—an ordered sequence of fixation locations for each image was collected from eight participants. Saliency maps for each image were also obtained from a version of TOCH tuned to the relevant target object. The model's performance for each image was measured by computing the percent of participant fixations within the most active regions of the saliency map. These "most active regions" were defined to be the set of locations whose saliency is ranked in the top 20% of the saliency distribution for an image. The saliency maps of TOCH were compared to saliency maps representing cross-participant consistency, which provided a useful upper bound on possible model performance. For additional comparison, Torralba et al. (2006) also obtained saliency maps from a simple exogenous model.

To assess TASC's performance on real-world images, we tested it on the same set of images used by Torralba et al. (2006). Other state-of-the-art models (e.g., Ehinger et al., 2009; Kanan, Tong, Zhang, & Cottrell, 2009) have reported similar or better performance than TOCH, but for simplicity, we assess TASC's performance in the context of the results presented in Torralba et al. The set of images used for testing was excluded from TASC's training set. Saliency maps were generated from the object module row relevant to the specific task and performance was measured using the method described above. Figure 10 presents the performance of TASC, along with results reproduced from Torralba et al. The results are divided by target object and by whether the target was present or absent. Each bar graph is further divided along the abscissa by fixation number within a trial. Thus, the light blue bar in the upper left bar graph corresponding to fixation 1 expresses that for the person search task with target present images, on average about 65% of the participants' first fixation in an image fell within the 20% most active region of the TASC saliency map for that image. Across all search cases and fixation numbers, TASC's performance is, on average, better than the simple bottom-up model (BU) and slightly worse than TOCH.

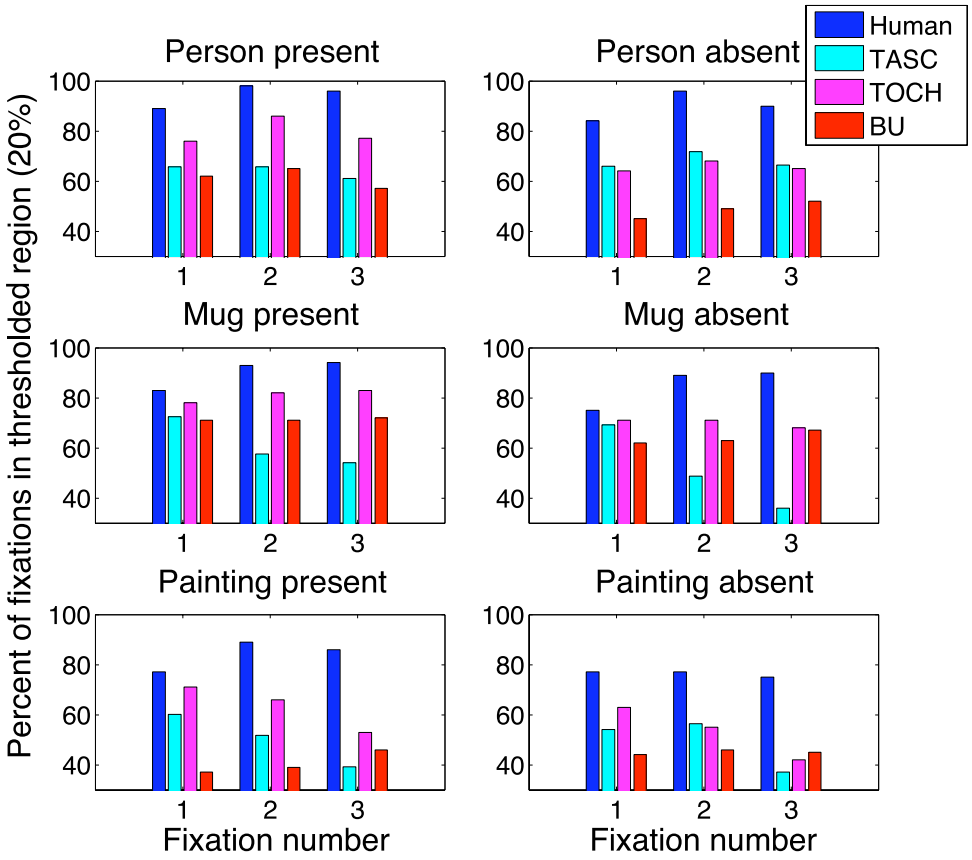


Figure 10. TASC performance on the eye-movement data set used in Torralba et al. (2006). TASC is compared to three alternatives described in Torralba et al.: TOCH, a pure bottom-up model (BU), and cross-participant consistency. The dependent measure is goodness of match to human eye movements, measured as the percent of human eye movements contained in the top 20% of the saliency map. The results are broken down by fixation number of the human participants within a trial.

It is important to note that the TASC implementation used here was not optimized for these images and search tasks. Instead, the model parameters were the same as those used in all other simulations including those with synthetic images. Apart from the lack of parameter tuning, another major explanation for why TASC performs worse than TOCH in this simulation is that TASC does not include any exogenous control—i.e., only one module row is used. Thus, the model is completely insensitive to any exogenous factors that might capture human attention. Exogenous control could be added to TASC for this simulation by including a large set of target objects. However, this was not central to the goal of the experiment, which was to show that TASC can utilize scene-based endogenous control. Performance in this simulation was also limited by the naive assumption that for a given task, all ranges of influence are equally relevant. These two limitations are likely responsible for TASC's relatively poor performance in the mug search. This relationship can be understood by noting that the performance of the BU model is closest to TOCH in mug search. Because TOCH consists of a bottom-up model filtered by global contextual guidance, the similarity between TOCH and BU in this task suggests that mug locations are only weakly predicted by global scene properties. Hence, the bottom-up component of TOCH, which operates on local feature information and employs exogenous control, is responsible for most of the model's predictive power. In the naive version of TASC, however, only one row of object modules is used and all ranges of influence are combined equally thus causing a dilution of the shortest range of influence by the longer ranges. In fact, the saliency maps obtained from the shortest range of influence alone perform similarly to TOCH and BU in the mug search task. As discussed in the TASC implementation section, a more complex implementation of TASC would model the relevance of all modules to a particular goal. Under this approach, the final TASC saliency map could weakly incorporate other object modules to achieve partial exogenous control and could also weight the ranges of influence appropriately. Learning these relevance weights, however, is beyond the scope the current work. Despite these limiting factors, the present result demonstrates that TASC is able to integrate relevant scene-based saliency predictions to achieve a performance level that exceeds a simple exogenous strategy and is not terribly behind a state-of-the-art model designed to utilize exogenous control with scene-based

endogenous control and tuned specifically to match the human data for this experiment.

Relationship between successive fixations

In a separate analysis of the eye-movement data in the study above, we found a strong predictive relationship of successive fixation locations. Specifically a model that generates a saliency map with Gaussian activity centered around the most recent fixation location performs comparably to state-of-the-art models including TASC and TOCH. This finding is similar to the center bias modeled in Zhang, Tong, and Cottrell (2009). Most models of attention, however, ignore the location of the previous fixation. When a bias toward the past fixation location was integrated into TASC, the model's performance significantly improved. This result suggests that future attentional models should incorporate a fixation trajectory more directly into the saliency predictions. It also spreads doubt on the standard inhibition-of-return mechanism often posited to generate fixation sequences in attentional models (e.g., Itti & Koch, 2000).

Contextual cuing

The notion of scene-based endogenous control is, in part, motivated by findings suggesting that attention can be guided by properties of the image other than visual features of the target. In a seminal study showing that such guidance can be learned, Chun and Jiang (1998) found that repeated distractor configurations in a difficult visual search task lead to faster response times. In this study of *contextual cuing*, participants were shown displays like that in Figure 11a containing a single target—a rotated letter T of any color—among distractors—Ls at various orientations. The participants' task was to report whether the T is oriented to the left or to the right. Unbeknownst to participants, some display configurations (i.e., the spatial layout of the distractors) 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 blocks of trials, participants respond roughly 60 ms faster to predictive displays than to random or *nonpredictive* displays. Figure 11b displays the contextual cuing effect reported in Experiment 1 of Chun and Jiang (1998). A significant difference in response time between the nonpredictive and predictive cases is present after the first epoch, which consists of 5 blocks each with 24 trials—i.e., 5 presentations of each predictive display.

The contextual cuing effect implies that attention is guided in part by properties of the entire scene. Thus, models, such as TOCH, that incorporate scene-based endogenous control should be able to capture the contextual cuing effect. TOCH by itself would be incapable of simulating a contextual cuing task because it lacks a target-specific saliency component. The extension of TOCH, which includes target specific processing (Ehinger et al., 2009), would perhaps be able to model the contextual cuing effect but would still be limited by the fact that the scene-based component produces uniform predictions across the horizontal plane of an image. In the following simulation, we demonstrate that TASC is well suited to explain the contextual cuing effect. Furthermore, we find that the TASC framework offers a novel interpretation of contextual cuing—supported by empirical studies—that implicates use of a strategy slightly different than scene-based endogenous control.

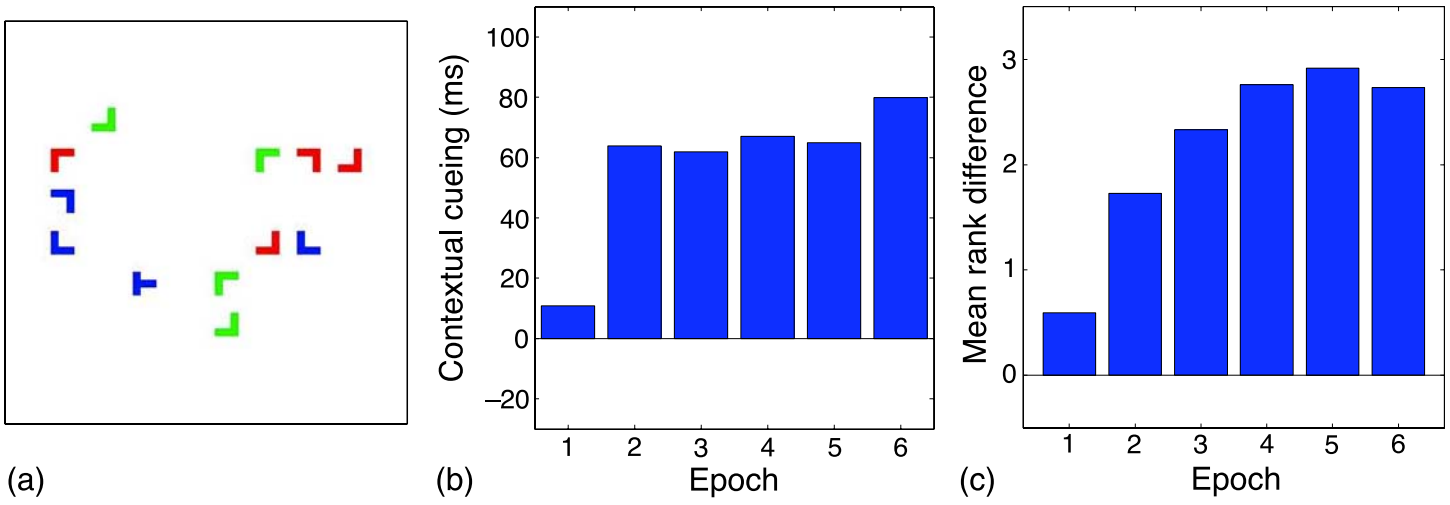


Figure 11. (a) A sample contextual cuing display. (b) A reproduction of the main contextual cuing result from Chun and Jiang (1998), Experiment 1), showing the difference in response times for nonpredictive and predictive displays. Responses are significantly faster for predictive displays after the first epoch. (c) TASC simulation of Chun and Jiang (Experiment 1), showing the difference in rank between the nonpredictive and predictive displays. Predictive ranks are significantly lower after the first epoch.

Our simulation of contextual cuing uses the same model parameters as in other simulations. However, the training procedure is modified to conform to the sequential nature of the contextual cuing task. Instead of training the model to asymptote as in other simulations, for the contextual cuing simulation, TASC is exposed to images in series and learning is incremental. We begin with a training phase in which a row of modules becomes accustomed to the target, in this case a sideways T. This training stage corresponds to the practice trials in a contextual cuing experiment. The images in the training phase are generated using the same methods as in Chun and Jiang except that none of the configurations are predictive. As in Chun and Jiang, the model is then presented with 30 blocks of 24 images, 12 with predictive displays that repeat from block to block and 12 with nonpredictive displays that always change. Target ranking in the final combined saliency map is again used as a measure of response latency. The simulation have one free parameter: the step size for gradient descent weight updates in the neural net though in practice the value of this parameter has little effect on the main results.

Figure 11c shows the difference in ranking between nonpredictive and predictive trials, averaged across 5 separate simulations, and exhibits a pattern very similar to Chun and Jiang's behavioral result in Figure 11b. Most importantly, both figures show no difference between the two cases at the onset of the experiment and a noticeable difference beginning in the second epoch. Figure 12a depicts the block-by-block learning pattern for predictive and nonpredictive displays in TASC. Learning occurs quickly for the predictive displays and performance is greatly facilitated compared to the minimal improvement for nonpredictive displays. Figure 11c can be obtained from Figure 12a by averaging across groups of 5 blocks and computing the difference between cases.

The results presented in Figure 12a are obtained from the final saliency map that—as in previous simulations—combines outputs across all ranges of influence. However, we would intuitively expect the longer ranges of influence to be more responsible for the contextual cuing effect because short ranges cannot detect configurations of distractors that extend beyond a local region. This is in fact the case as shown by Figure 12b, which presents the average predictive and nonpredictive ranks across the whole simulation computed using the precombined saliency maps from the 4 ranges of influence. The first 5 blocks are removed from each average because significant learning occurs during those trials. It is clear that the difference in rank increases as the range of influence grows and that the target rank for predictive displays decreases with longer ranges of influences. The two longest ranges of influence appear to be most responsible for the response time facilitation.

From the perspective of the TASC control space, contextual cuing, with its dependence on the long ranges of influence, corresponds to an attentional strategy that primarily exploits the region of the control space with high task specificity and a scene-level contextual scale. However, TASC predicts that the attentional guidance that produces the contextual cuing effect is not just an example of scene-based endogenous control. As suggested by Figure 12b, the second longest range of influence—which covers quadrants of the image—is also capable of producing the effect. This difference suggests that contextual cuing might be treated as an intermediate strategy, distinct from scene-based endogenous control, in the TASC control space. Figure 2 adds a point in the control space for contextual cuing with high task specificity and an intermediate contextual scale near the scene end of the spectrum. Notice also that contextual cuing is placed lower

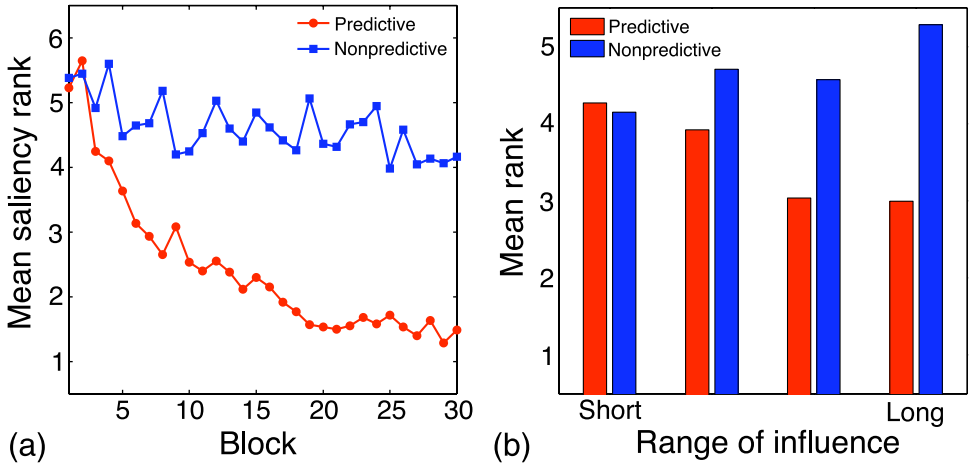


Figure 12. (a) Block-by-block simulation results from TASC on the Chun and Jiang (1998), Experiment 1) simulation. The ordinate shows the mean rank of the target for predictive and nonpredictive displays. (b) Mean ranks for the 4 ranges of influence in TASC from the contextual cuing simulation. The first 5 blocks are removed from the averages. The two longest ranges of influence demonstrate significant facilitation for displays with predictive contexts.

than scene-based endogenous control on the task specificity axis. Because the same configuration could provide facilitation for a variety of target objects, the relationship between a specific target and the contextual cuing configuration is weaker than the relationship between an object and the scene gist in a real-world task. In essence, contextual cuing could still have an effect when the task is more loosely defined. This novel prediction in TASC that contextual cuing draws from an intermediate contextual scale is supported by the contextual cuing experiment in Olson and Chun (2002), which observed a response time facilitation when predictive configurations were constrained to one quadrant of the image.

Recent research on contextual cuing with real-world displays has led to a contradictory conclusion about how global and local contexts affect response times (e.g., Brockmole, Castelhano, & Henderson, 2006). For these pseudo real-world images, observers use global contextual cues for attentional guidance far more than local contextual cues. From the perspective of TASC, the conflict between these two results can be explained by recognizing that people have a greater wealth of experience viewing real-world scenes than artificial contextual cuing displays. Consequently, the attentional system is tuned to be directed by familiar global properties of real-world scenes but is only capable of learning more local configurations in contextual cuing images.

Spatial probability cues

TASC's emphasis on the role of experience provides a natural explanation for the contextual cuing effect in terms of rapid, online learning of the likely target location contingent on display configuration. To the extent that TASC incorporates learning of display-contingent targets, the TASC theory should necessarily also allow for learning of display-noncontingent targets. That is, if targets appear in certain locations with high probability, regardless of the display contents, then the learning mechanisms embedded in TASC will necessarily discover this probability distribution, and adaptation should be on the same brief time course as adaptation effects observed in contextual cuing.

Indeed, Experiment 1 of Geng and Behrmann (2005) shows that individuals respond more rapidly to targets at locations that frequently contain a target, relative to locations where targets rarely appear. Further, the time course of this learning is rapid—RT to high-probability locations is significantly lower than RT to low-probability locations when averaged over a block of 180 trials.

The architecture and design of TASC necessarily reproduces this result. Specifically, the neural networks of each module implicitly encode the prior probabilities of targets at each location via the bias weights of the neural network. That is, whenever the network experiences a target in some particular location, error-correction training will increase the bias on that location, thereby raising its default activation level. Over a sequence of trials, if a target appears at one location with high probability and another location with low probability, the network will yield higher activation for the target at the probable location, and consequently, the target–distractor saliency ratio will be greater for that target and the response time will be faster.

Geng and Behrmann also find that response times associated with targets in low-probability locations are slower when the high-probability location contains a distractor, versus when the location is empty. TASC has a ready explanation for this finding. The high-probability location will have an activation bias, regardless of the object appearing at that location. If any object appears at that location, target or otherwise, it will likely increase the activation at that location; call this *spurious activation*. The sum of the activation bias and the spurious activation may be large enough that in the final saliency map, the sum will outrank the activation of the actual target at the low-probability location, and if ranking determines the prioritization of search, response latencies will increase. However, when the high-probability location is empty, the absence of spurious activation at that location should make it less likely that the high-probability location obtains saliency that outranks that of the low-probability location.

Figure–ground assignment

The three primary control strategies presented in the Introduction section occupy three of the four corners of the TASC control space. We suggested that contextual cuing might populate a distinct point in the control space, but there is an obvious gap in the fourth corner of the space, corresponding to a control strategy involving a low degree of task specificity and scene-level contextual guidance. If attention truly has access to all points in the TASC control space, we should be able to identify a situation in which the strategy in the fourth corner is employed.

One candidate is the lower region effect observed in figure–ground assignment. Vecera, Vogel, and Woodman (2002) observed that for displays such as the first and third panels of Figure 13, viewers tend to perceive the lower region as the figure (the foreground) regardless of the arrangement of colors. This result can be interpreted as an attentional bias in favor of the lower region. Vecera et al. suggest a plausible basis for the effect: because of our body orientation with respect to the ground and the sky, objects of interest are more often found in the lower region of our field of view.

An interpretation of the lower region effect is not readily obtainable from existing attentional models. However, from the perspective of TASC, the effect can be naturally viewed in terms of an attentional control strategy that is based on overall scene properties (i.e., a scene-level contextual scale) and no specific target of search (i.e., low task specificity). We have placed this strategy, lower region, in the fourth corner of the control space (Figure 2). Lower region is a

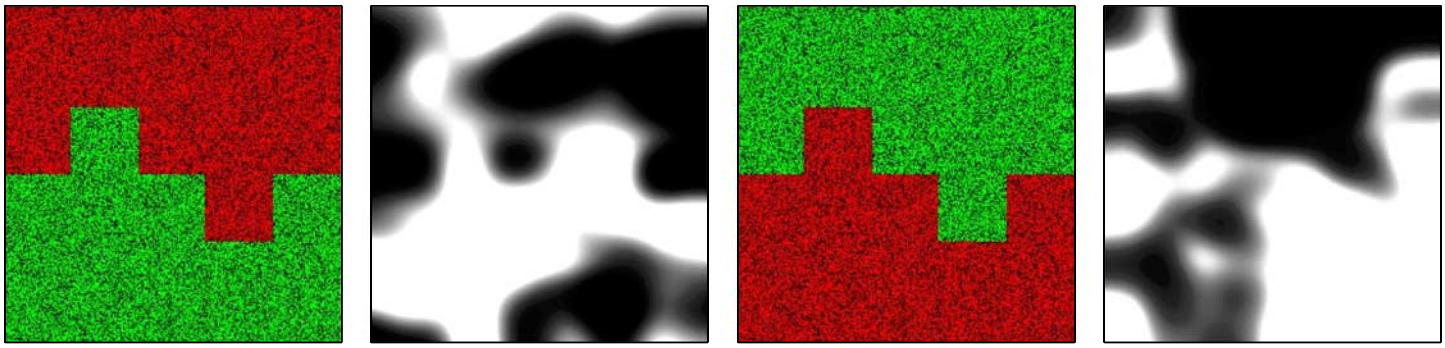


Figure 13. Sample displays similar to those used in the figure–ground assignment experiments in Vecera et al. (2002) and the associated outputs from TASC operating at the longest range of influence and with low task specificity. Saliency maps represent a combination across 13 object models trained on real-world images. A bias is shown for the lower region in both displays.

natural complement to the three primary control strategies described previously.

To demonstrate that TASC obtains the lower region effect, our implementation of TASC with 13 real-world object modules was presented with displays similar to those in Vecera et al. Because figure–ground assignment is performed quickly, and the limited processing time might yield a lower signal-to-noise ratio of early visual representations, we added noise to the feature detector responses in the model. The noise primarily served to obtain more overall partial activation of features. The saliency maps in the second and fourth panels of Figure 13 are based on combining all object modules and only the longest range of influence. Because the displays bear no resemblance to real-world images, the activity patterns in the saliency maps are difficult to interpret. However, TASC does show a clear bias for attending the lower region of the image regardless of the color layout—i.e., there is more saliency activity in the bottom half of both maps. These saliency maps are obtained from a limited version of TASC with only 13 objects and thus one can justifiably be cautious in making strong claims about this result. Some of the 13 objects seem likely to appear in the lower region of an image (e.g., car, person, bike, sidewalk, road); others seem likely to appear in the upper region (e.g., building, tree, window, painting); and others seem difficult to predict (light, head, sign, mug). Thus, the emergent result obtained from TASC seems nontrivial. Furthermore, though other models of attention may be able to partially achieve this result because they are trained on a set of similar images, none of the models can be specifically used to generate task-independent global attentional guidance of this sort.

Throughout this paper, the term *exogenous control* has been used to refer to bottom-up attentional processing at a feature-level contextual scale. This usage is consistent with most of the literature on attention. In the lower region effect, however, we recognize a different form of exogenous control that operates at a scene-level contextual scale. From the perspective of the TASC framework, exogenous attention—defined to be any attention that is driven primarily by factors outside the individual—should correspond to the entire bottom region in the control space—i.e., low task specificity across all contextual scales. This new conceptualization of exogenous attention may prove useful in evaluating attentional behavior in unstructured viewing situations. Although we favor this novel perspective, we continue to use the term *exogenous control* with its traditional meaning throughout the rest of this article.

Discussion

We have introduced a framework for attentional control that integrates many distinct control strategies. This framework, instantiated in the TASC model, makes contact with a wide range of attentional phenomena, including visual search in both naturalistic scenes and artificial displays consisting of simple elements varying in color and shape. Our goal in this work was to develop TASC to the point that it serves as an existence proof of the possibility that diverse forms of attentional control can be unified in a coherent theoretical framework.

TASC borrows from many other theories of attention, and in so doing, it serves as a synthesis of existing theories and as a means of characterizing the relationships among the theories. TASC follows a clear trend in the literature toward models that are comprehensive in scope and incorporate multiple attentional control strategies (e.g., Ehinger et al., 2009; Kanan et al., 2009; Siagian & Itti, 2007; Torralba et al., 2006; Zhang et al., 2008). TASC is distinguished by the fact that the distinct control strategies in TASC are implemented by a uniform, homogeneous architecture. Through simulations, we showed that exogenous control, feature-based endogenous control, and scene-based endogenous control could all emerge from the same underlying processing machinery.

Beyond accounting for phenomena that many other models have addressed, the TASC control space led us to explore the literature to identify phenomena that reflected control strategies not considered by other models and that—in contrast—are predicted to exist by TASC. First, in

the contextual cuing paradigm, our simulations of TASC predict that contextual cuing might result from spatial regularities at the subscene level—say, by repetition of configurations within a quadrant (Figure 12b). This prediction—a necessary consequence of the TASC framework—has been verified in experimental studies (Olson & Chun, 2002). Second, the TASC control space is populated by three primary strategies, and the space suggests a complementary fourth strategy—scene-based exogenous control. In a review of the literature, we identified a phenomenon that is well described by this control strategy: the lower region effect found in figure–ground assignment (Vecera et al., 2002). Individuals' bias to treat the lower region of an image as the figure is consistent with the notion of task-independent attentional control based on scene properties. An important aspect of TASC is that it provides a unified synthesis of existing models. However, these phenomena demonstrate that TASC also yields novel predictions about attention and is capable of explaining data that eluded previous models. Furthermore, the existence of these phenomena suggests that we may find tasks that require other attentional strategies occupying new locations in the space. For example, the categorical visual search observed in Schmidt and Zelinsky (2009) is consistent with an attentional strategy operating with a medium amount of task specificity (i.e., groups of objects are of primary interest).

Embedded in TASC is a key proposition: *attentional control is fundamentally experience based*. TASC consists of a collection of modules, each of which is specialized to the detection of a particular entity, and is assumed to arise from interacting with the world, and to continually adapt to the ongoing stream of experience. These entities can be what are traditionally referred to as features, such as the color red, or simple shapes, such as a square, or complex, hierarchical, articulated objects, such as a person. We have loosely referred to these entities as objects, although this usage stretches the canonical notion of objects in vision. The important point is that these objects are functionally defined—they are the goals of visual search, goals that have been learned and rewarded in the past.

According to TASC, attentional control varies in the degree of task (object) specificity via the combination of object modules. By this conception, what is typically considered top-down control occurs when one or a small number of object modules contribute to saliency; and what is typically considered bottom-up control occurs when all object modules contribute to saliency. Instead of considering exogenous control as a search with *no* target in mind, TASC conceptualizes it as a search for *all* possible targets. Consequently, TASC rejects the notion of pure bottom-up saliency, instead favoring the novel hypothesis that control *always* operates in a top-down manner, tuned through experience. If attention is always combining across a set of object modules, it is important to specify how this combination is performed. In our simulations, we found the use of a max or *disjunction* rule to be critical. This approach yields the sensible behavior that a region should be salient if it contains any of the potential target objects.

Failure of attentional control to exclude task-irrelevant signals can be attributed to limitations on the degree of task specificity that can be achieved. On grounds of survival, it may be adaptive for an organism not to become so focused on the task at hand that all extraneous warning signals from the environment are suppressed. The implication for a theory like TASC is that critical object modules contribute to saliency even when not pertinent to current goals. For example, the presence of a human face may be important to many different goals, including basic survival in a community. Thus, attentional capture can be considered not as a failure of attentional control to stay on task but as the inability of attentional control to narrow to a single task.

The role of experience

Several recent theories of attention have focused on the role of top-down knowledge in attentional control. TASC is motivated in part by TOCH, whose contextual component, presented in Torralba (2001, 2003), encodes and exploits correlations between scene gist and the locations of specific objects. Rao, Zelinsky, Hayhoe, and Ballard (2002) implement a model of visual search in which saliency is based on how well local visual features match a specific object template. Similarly, the SAIM model of Heinke and Humphreys (1997, 2003) incorporates object templates to determine a focus of attention via constraint satisfaction dynamics.

In contrast to these theories that assume that experience leads to the formation of internal models of specific objects and their likely locations, other theories assume that what is learned is not about objects per se, but rather statistical properties of the environment, which might steer attention to locations containing surprising visual information (Itti & Baldi, 2009; Zhang et al., 2008) or might optimize attention to situations that are likely to occur in the future (e.g., Mozer & Baldwin, 2008; Mozer, Shettel, & Vecera, 2006). The various statistical theories agree in claiming that attention is modulated by statistics of the environment, but they differ in terms of the time period over which statistics are collected. Some theories rely on life long history (e.g., Kanan et al., 2009; Torralba, 2003; Zhang et al., 2008), some theories rely on the recent history of experience and suggest trial-to-trial modulations of attention (e.g., Itti & Baldi, 2009, temporal surprise; Mozer & Baldwin, 2008; Mozer et al., 2006; Yu, Dayan, & Cohen, 2009), and finally, some theories assume that experience does not extend beyond the statistics of the current image (e.g., Bruce & Tsotsos, 2009; Itti & Baldi, 2009, spatial surprise; Torralba et al., 2006, bottom-up component). Like all of these theories, TASC can be cast in a probabilistic framework and the saliency values can be readily interpreted as probabilities. TASC has the virtue of spanning a range of time scales of adaptation. Some simulations we

presented rely on experience on the time scale of years (e.g., attention in real-world scenes, lower region bias in figure–ground assignment); other simulations rely on experience on the time scale of minutes (e.g., contextual cuing). Adaptation to experience is seen on an even shorter time scale in research focusing on sequential effects in attention (e.g., Geng & Behrmann, 2005; Kristjansson, 2006; Maljkovic & Nakayama, 1994). Just as TASC adapts to the block-by-block statistics in contextual cuing experiments, the neural networks that learn specific object associations exhibit trial-by-trial sequential dependencies due to online learning. Regardless of how TASC plays out compared to other probabilistic theories, we view TASC as the culmination of a shift in the theoretical literature, from the assumption of hardwired, fixed mechanisms of attention to the view that every aspect of attentional control relies on adaptation to the ongoing stream of experience.

Adaptation of attentional control

Up to this point, we have focused on one aspect of adaptation: how experience tunes a module to perform a specific task. This tuning occurs via a gradient-descent update of the weights in the module's associative network following each trial in which the task is performed. This online updating provides a basis for explaining phenomena such as contextual cuing and probabilistic spatial cuing.

Beyond this notion of adaptation, TASC suggests a complementary type of adaptation that we have not yet discussed: how experience shapes the set of modules deemed relevant to a particular search goal. The TASC architecture assumes that for a specific goal, the saliency map is obtained by taking a disjunction over a subset of modules deemed relevant to the goal of search. In simulations we have presented, the relevant subset of modules was assumed to be known and fixed. Ultimately, however, the determination of relevance must be learned as well—a type of meta-attentional learning. This form of learning has been ignored up to this point in the attentional literature, though it seems to be an important part of the complete picture.

The relation of attention and object recognition

In TASC, the goal of attention is to identify locations in the visual field that contain objects of interest. Other recent theories of attention also relate attentional saliency to the probability that an object is present (Bruce & Tsotsos, 2009; Gao, Mahadevan, & Vasconcelos, 2008; Kanan et al., 2009; Mozer & Baldwin, 2008; Mozer et al., 2006; Navalpakkam & Itti, 2007; Torralba et al., 2006; Zhang et al., 2008). If attention is conceptualized as being fundamentally a means of object detection, then what distinguishes the role of attention from the role of full-blown object recognition?

Our answer, in essence, is that attention performs a quick-but-dirty sort of recognition: attention operates rapidly to select locations that are likely to contain a target. However, speed is obtained by performing only a coarse, rudimentary analysis of the visual field. Three aspects of TASC are responsible for achieving a quick-but-dirty response:

1. *Information bottlenecks*. TASC has bottlenecks at various stages of processing that restrict the visual information used to produce a response. Early in processing, principal components analysis (PCA) and subsampling operate to reduce the dimensionality of the visual representation. Late in processing, a bottleneck is imposed on the neural network by the small number of hidden units through which activation must flow to reach the saliency map. The hidden unit bottleneck effectively imposes a second PCA compression (Baldi & Hornik, 1989); this PCA is object-specific, whereas the PCA early in processing is always the same. Although some form of dimensionality reduction is widely used in neurobiological, psychological, and artificial intelligence models of object recognition as a means of filtering noise and redundancy in the visual representation, the degree of dimensionality reduction in TASC is far more severe, preserving only the strongest, coarsest regularities in the visual representation.
2. *No feedback loops*. Processing in TASC is one pass and purely feedforward. In contrast, models of object recognition typically have some form of recurrence, whether at the fine grain via feedback projections in neurobiological models of the ventral visual stream (e.g., Bullier, 2001), or at an intermediate grain via sequences of fixations to foveate on critical visual information in models of scene analysis that incorporate an anisotropic retina (e.g., Henderson & Hollingworth, 1998), or at a coarse grain via iterative hypothesis testing and confirmation in computer vision (e.g., Lowe, 1991). Of course feedback loops exist in TASC to accomplish the different forms of adaptation and learning. However, these loops do not change the response for the current image; they only have an impact on subsequent images as in the contextual cuing simulation.
3. *Linearity of associative neural network*. The linearity of the hidden unit response in the associative neural network restricts the complexity of the mapping that can be learned to a rank-limited linear transform from the input to the output. Linearity introduces strong restrictions on the accuracy of object recognition (Mozer, 1991).

The TASC modules compute an estimate of how likely each location in the visual field is to contain a specific object. The accuracy of this estimate is limited by the three properties of TASC we just described: information bottlenecks, no internal feedback, and linearity of the

associative net. These three properties also play a large role in accounting for various experimental data, e.g., the inefficiency of conjunction search. Although one could, in principle, design a model that more reliably determines the presence or absence of an object in the visual field, it would effectively become a recognition model, not an attentional model. In addition, to perform object recognition in parallel across the visual field would require an overwhelming amount of hardware.

Early selection theories, including TASC, suppose that the role of attention is to provide rapid, efficient guidance to resource-limited processes involved in object recognition and scene interpretation. Classic early selection theories (e.g., Broadbent, 1954; Treisman, 1960; Treisman & Gelade, 1980) posited that guidance is based on hardwired mechanisms that utilize primitive visual features such as motion and contrast in intensity or color or texture. This view persisted through the 1990s, even in theories of feature-based exogenous control (Wolfe, 1994), and even in modern computational models (e.g., Itti & Koch, 2000; Itti et al., 1998; Mozer, 1991). Only recently have early selection theories begun to consider the alternative view: through experience with one's environment, an individual learns to guide attention in a goal-relevant manner. TASC represents a culmination of this view by adopting the strong hypothesis that all forms of attentional control are fundamentally experience based. Whether TASC or some less extreme variant proves to be correct, what is perhaps the key contribution of our work is an appreciation of the fundamental shift in theoretical views of spatial attention.

Appendix A

In this section, we give an in-depth presentation of our implementation of TASC. Though not necessary for understanding the main thrust of the TASC framework, the equations and parameters described here are necessary for recreating an implementation. This is meant as a supplement to the TASC implementation section in the body of the paper that presents the overall framework and coarse processing stages of the model. The image processing steps described below are all performed on a patchwise basis and correspond to the stages illustrated on the right side of Figure 7. The model uses 4 different patch sizes for the 4 ranges of influence, $\phi_1 = 1.56$, $\phi_2 = 6.25$, $\phi_3 = 25$, and $\phi_4 = 100$ ordered from short to long range of influence, where ϕ represents the percent of the image the patch spans. The patch sizes are depicted in Figure A1a.

Feature extraction

TASC begins with an image such as in Figure A1a. From this image, the model records feature activations along 8 different channels: 4 color opponencies and 4 orientations as shown in Figure A2a. The value of each color channel is the difference between the RGB value for that color and its opposing color (red opposes green and blue opposes yellow). All negative color feature values are set to 0. If R, G, and B are the pixel values and $Y = (R + G)/2$, then the red, green, blue, and yellow feature values are given by $f_r = \max(0, R - G)$, $f_g = \max(0, G - R)$, $f_b = \max(0, B - Y)$, and $f_y = \max(0, Y - B)$. Technically, the feature values should be denoted $R(x, y)$ and $f_r(x, y)$ for each pixel location, (x, y), but we drop the location specification when it is not necessary. For the orientation channels, the intensity image is convolved with Gabor filters tuned to 4 different orientations: 0, 45, 90, 135 degrees. All feature extraction parameters are the same across the 4 ranges of influence.

The TASC feature extraction stage is very similar to feature extraction performed in the NI and TOCH models. All models use the same process for extracting orientation values. NI uses color opponency values similar to those used in TASC. In Itti and Koch (2000), which forms the basis for the early stages of NI, two color channels are used: red–green and blue–yellow (negative values allowed). TASC uses these same values but divides them into 4 channels to avoid negative feature values.

The primary difference between feature extraction in TASC and the NI and TOCH models is that TASC does not extract features at multiple spatial frequencies. Feature extraction utilizing multiple spatial frequencies, obtained, for example, through the steerable pyramid (Simoncelli & Freeman, 1995), is typically employed to improve the model's ability to handle image data sets with a diverse range of object and scene scales. To reduce the computational load, we chose to perform feature extraction at a single spatial frequency in TASC and found the results satisfactory for our simulations. Nevertheless, we envision a complete implementation of TASC that includes feature extraction at multiple spatial frequencies and expect that this addition would improve the overall performance.

Contrast enhancement

Feature extraction is followed by a contrast enhancement stage that mimics neural processing mechanisms in visual cortex. This stage strengthens the response to regions that differ significantly from their surround. Contrast enhancement is implemented in TASC by weighting each feature value by the ratio of center activity to surround activity. Figure A2b shows the result of contrast enhancement of the activations in Figure A2a. The model computes center and surround activity by convolving the feature data for each channel by two Gaussian kernels. Both kernels have a peak of 1, but they differ in their standard deviation, with σ_c corresponding the center kernel and σ_s to the surround kernel. For each pixel and feature channel, the model obtains a center value, *cent*, which measures the amount of

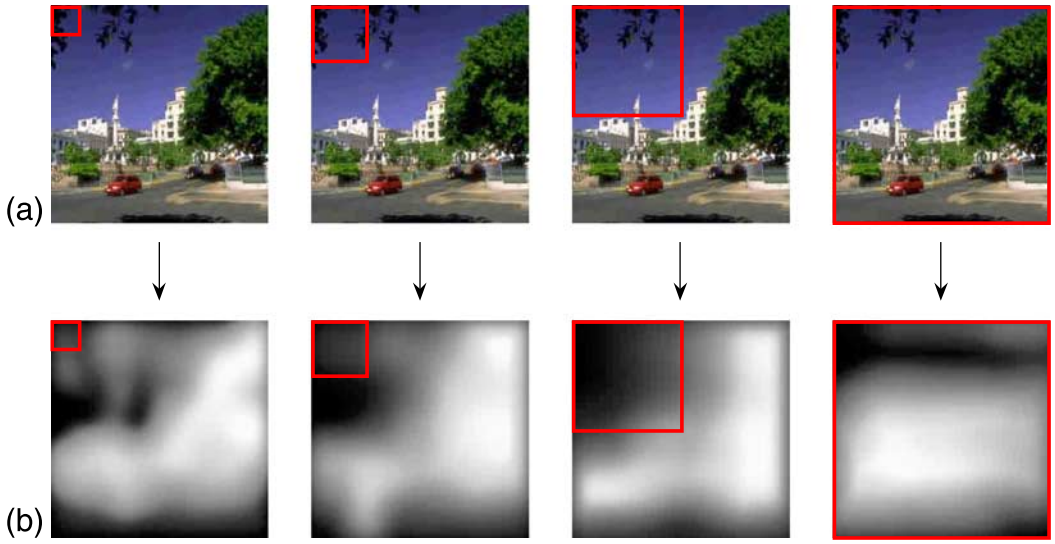


Figure A1. Mapping from patch in image to patch in saliency map for the 4 TASC modules that vary from short range of influence (left) to long range of influence (right). The saliency maps (b) are obtained by averaging overlapping patch maps obtained for the image in (a).

activity close to that pixel location and a surround value, *surr*, that measures the activity over the wider region surrounding the location. The contrast value, $c(x, y)$, is computed for each of the 8 feature channels as follows using the red feature channel as an example:

$$c_{\mathrm{r}}(x,y) = \frac{cent_{\mathrm{r}}(x,y)}{surr_{\mathrm{r}}(x,y)} f_{\mathrm{r}}(x,y), \tag{A1}$$

where

$$cent_{\mathrm{r}}(x,y) = \sum_{i,j \in I} f_{\mathrm{r}}(i,j) e^{\frac{-(x-i)^2}{2\sigma_{\mathrm{c}}^2}} e^{\frac{-(y-j)^2}{2\sigma_{\mathrm{c}}^2}} \text{ and}$$

$$surr_{\mathrm{r}}(x,y) = \sum_{i,j \in I} f_{\mathrm{r}}(i,j) e^{\frac{-(x-i)^2}{2\sigma_{\mathrm{s}}^2}} e^{\frac{-(y-j)^2}{2\sigma_{\mathrm{s}}^2}}, \tag{A2}$$

with (i, j) iterating over all pixel locations in the image I.

Because feature values are always nonnegative, the surround kernel bounds the center kernels—i.e., $surr \geq cent$, and thus the ratio of center to surround will always be between 0 and 1. If most of the activity for a specific feature is in the center, the ratio will be close to 1 and the feature activity at that location will remain strong. If there is significant activity for the specific feature in the surrounding regions, this ratio will be close to 0 and the feature value at that pixel will be reduced in strength.

For the feature extraction and contrast enhancement stages, the model computes the f and c values from the whole image rather than by patch. For feature extraction, this avoids boundary issues at the edges of the patch when convolving the image with the Gabor filter. For the contrast enhancement stage, this allows the surrounding areas of a location to have an influence on the contrast value even if they reside in a different patch. The standard deviations for the center and surround kernels are $\sigma_{\mathrm{c}} = 0.05N$ and $\sigma_{\mathrm{s}} = 0.5N$, where the image dimension is $N \times N$. These

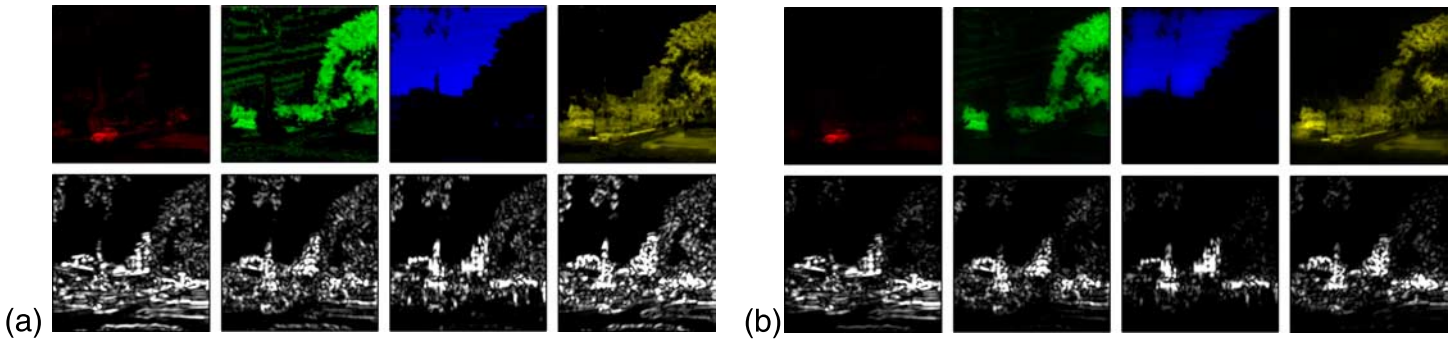


Figure A2. Activities in TASC for the first two stages of processing: (a) feature extraction and (b) contrast enhancement. The original image is shown in Figure A1a. The feature channels on top are red, green, blue, and yellow. The four channels on the bottom correspond to orientations of 0, 45, 90, and 135 degrees from left to right.

parameters lead to a contrast mechanism that compares a local region to a large part of the image, i.e., contrasting occurs at a global scale. The surround region could be shrunk to yield more local contrasting. We chose these values primarily because the content of synthetic images used in visual search tasks is usually spread across the whole image. These parameters were left fixed for all other image data sets. However, a smaller surround would perhaps be more appropriate for real-world images. The same center and surround parameters are also used across all ranges of influence in this implementation, though, these parameters could differ from one range of influence to the next. For example, at the long range of influence, a larger surround value may be preferred. It is also possible that several contrast values could be obtained by varying the center and surround percentages. This would, in essence, provide another form of multi-scale processing.

Dimensionality reduction

After the contrast enhancement stage, TASC maintains 8 feature values per pixel. Due to the redundancy of these features and redundancy of nearby locations in a patch, dimensionality reduction is called for to streamline computation. Following the methods used in TOCH for gist processing, the dimensionality of the representation in TASC is significantly reduced through subsampling and principal components analysis while preserving key information needed for scene classification.

Subsampling

As in TOCH, TASC uses a simple subsampling scheme where each pixel in the subsampled image represents the mean value of all pixels in a corresponding unique, nonoverlapping rectangular region in the original data. Subsampling is performed only within feature channels so for each subsampled region in the image the model still maintains 8 feature values. If the original patch dimension is $n \times n$ and patches are subsampled by a factor ρ, the new patch dimension becomes $n/\rho \times n/\rho$. Each pixel in the subsampled data corresponds to the average of all pixels in a $\rho \times \rho$ block in the original data.

The amount of subsampling varies across ranges of influence in TASC; we aim to achieve a roughly constant bandwidth of data across ranges of influence, leading to a greater reduction in the longer ranges of influence. This approach fits well with both NI and TOCH. Specifically, NI operates with a short range of influence and does not perform any subsampling while TOCH operates at a long range of influence and uses significant subsampling of feature data. TASC uses the following values of ρ for the four ranges of influence from short to long: $\rho_1 = 4$, $\rho_2 = 8$, $\rho_3 = 8$, and $\rho_4 = 16$. These specific values were chosen to yield a reasonable computational load in the following principal components analysis stage. However, changing these values does not significantly affect the model's performance.

Principal components analysis

After subsampling, the model uses principal components analysis (PCA) to further reduce the dimensionality of data. This transform extracts the critical feature properties of a patch and discards extraneous feature information. PCA analysis is done separately for each of the 8 feature channels and 4 ranges of influence. For each feature channel, the top ℓ principal components are preserved. The PCA projections are based on training on a set of patches extracted from our image database—using several randomly distributed patches per image. The database contains roughly 2,500 images including many varied real-world scenes and some synthetic visual search images. Note that for a given feature channel and range of influence, the same PCA projection is used for all patches, thus yielding a location-independent reduction. Additionally, the same PCA projections are used across all simulations described in the Simulations of TASC section—for both images with artificial displays and real-world scenes.

We chose to perform PCA on each feature channel to maintain a constant bandwidth per feature dimension. The alternative would involve lumping all feature data together and performing one PCA projection. In this case, PCA may find that a particular feature dimension does not have significant discrimination power and thus would not maintain any information about this feature channel in the top principal components. This could be problematic for visual search tasks where the discarded feature plays a defining role. This problem could be avoided by carefully controlling the database used for deriving the PCA projections; however, we found it preferable to instead treat the individual feature channels independently.

If the original patch size is $n \times n$, the dimensionality of each feature channel after subsampling is $(n/\rho)^2$. After PCA, there are only ℓ data points per channel. The full patch data, $\mathbf{x}$, are obtained by concatenating the data from each feature channel into a vector with dimensionality 8ℓ. The number of principal components used varies across the ranges of influence; from short range to long range, $\ell_1 = 1$, $\ell_2 = 3$, $\ell_3 = 9$, and $\ell_4 = 27$. Because there are far more patches per image at the shorter ranges of influence, the total number of data points per image decreases as the range of influence becomes longer. At the shortest range of influence, there are $15 \times 15 = 225$ patches per image (there are an odd number of patches because patches overlap). At progressively longer ranges of influence, there are $7 \times 7 = 49$, $3 \times 3 = 9$, and $1 \times 1 = 1$ patches per image. From short to long range of influence, the total

data points used per image are $225 \times 8 \times 1 = 1800$, $49 \times 8 \times 3 = 1176$, $9 \times 8 \times 9 = 648$, and $1 \times 8 \times 27 = 216$ (*patches* $\times$ *features* $\times$ ℓ).

Figure A3 provides an informative view into what the PCA stage achieves by depicting which regions each principal component favors for each feature channel. Each box in this figure displays the values of a specific principal component vector for a specific channel. Lighter values correspond to stronger weights for the features at that location in the patch. Principal components decrease in importance from top to bottom. The first principal component typically represents the DC level of activity. Subsequent principal components provide more fine-grained spatial selectivity. The principal component vectors reconstructed in this figure correspond to the second longest range of influence for which 9 principal components are used. The PCA projections used for the other ranges of influence exhibited a very similar pattern.

Figure A4 shows the representation for the image in Figure A1a after subsampling and PCA at the 4 ranges of influence. Each row in Figures A4a–A4c shows 8 composite images, one for each feature channel, formed by assembling the *n*th principal component value for each patch. The top row corresponds to the first principal component and lower rows display subsequent principal components with decreasing significance. Figure A4a is the representation for the shortest range of influence. At the longest range of influence (Figure A4d), there is only one patch per image but more principal components (now shown along the columns).

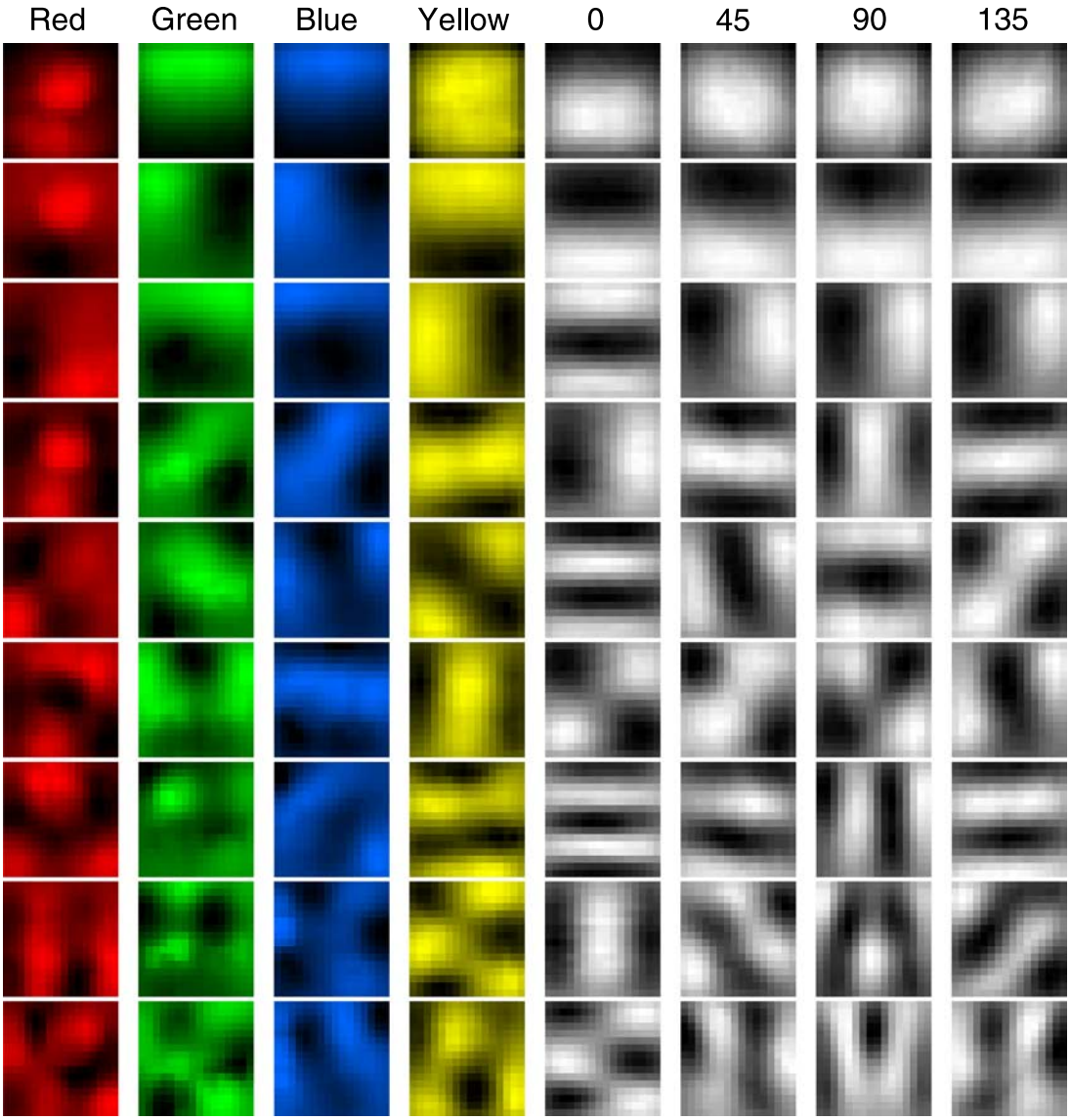
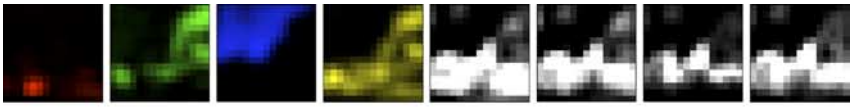
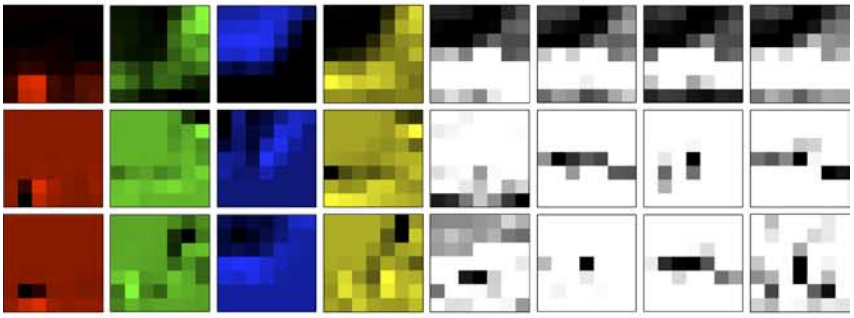


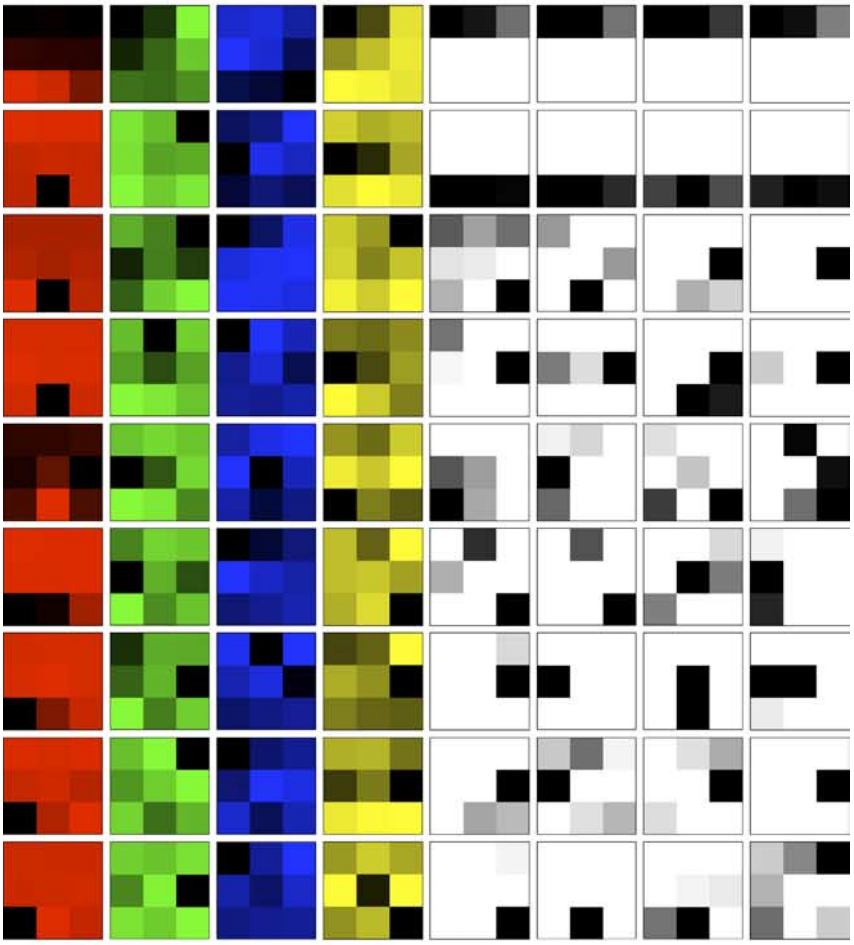
Figure A3. Reconstructed principal component weights for the 8 feature channels. The top row corresponds to the first principal component. See text for more detailed explanation.



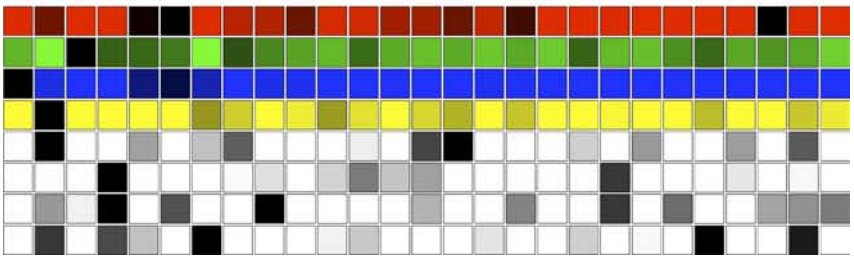
(a)



(b)



(c)



(d)

Task-specific associations

In TASC, the associative memory is implemented as a set of neural networks that process the image in patches. For each patch, there is set of input units to the network representing the patch data, $\mathbf{x}_p$, after dimensionality reduction. Each set of input units is fully connected to a set of output units via a layer of hidden units. The hidden units, however, are linear, and their purpose therefore serves to limit the rank of the mapping from input to output.

Because the shorter ranges of influence have far more patches than the longer ranges of influence, we make the hidden-layer bottleneck smaller at the shorter ranges, just as we chose a smaller number of principal components at the shorter ranges of influence. In fact, for simplicity, we simply chose the number of hidden units at a given range of influence, i, denoted r_i, to be the same as the number of principal components, previously termed ℓ_i: $r_1 = 1$, $r_2 = 3$, $r_3 = 9$, and $r_4 = 27$ for the 4 ranges of influence from short to long. (This choice results in a factor of eight compression of the image data, because the ℓ_i principal components are computed for each of the eight feature dimensions.) The rank of the mapping from input to output is determined by r_i, and when the number of patches is taken into account, our choice of the $\{r_i\}$ obtains a roughly equal number of descriptive features per pixel across the four ranges of influence.

If we use $\mathbf{V}_p$ to denote the weight matrix for the mapping from input to hidden units, and $\mathbf{W}_p$ to denote the weight matrix for the mapping from hidden to output, and the input-to-hidden mapping is linear and the hidden-to-output mapping includes a logistic squashing function, the output vector for a patch, $\mathbf{o}_p$ can be computed as

$$\mathbf{o}_p = h(\mathbf{W}_p \mathbf{V}_p \mathbf{x}_p), \tag{A3}$$

where $h(x)$ is the logistic function, $h(\mathbf{x}) = 1/(1 + e^{-\mathbf{x}})$. The logistic function yields output values bounded in the [0, 1] range, which are readily interpreted as degrees of saliency. For computational purposes, we constrain the dimension of $\mathbf{o}_p$ to be smaller than the original patch dimensions because saliency predictions do not need to be as fine-grained as the original pixel information. The reduction is such that the final saliency map has 32×32 pixels.

To obtain the final saliency map, the outputs from the individual patches must be combined. Because the patches overlap (except at the longest range of influence, which has only a single patch), the patchwise outputs must be synthesized. Either two or four patches predict the saliency at a given location, for most locations. (The corners are predicted by only a single patch.) The patchwise outputs are combined by averaging. That is, the final saliency value at a particular (x, y) location, $s(x, y)$, is computed as an average of all patch output values at that location:

$$s(x,y) = \frac{1}{|\Omega(x,y)|} \sum_{p \in \Omega(x,y)} \mathbf{o}_p\big(g_p(x,y)\big), \tag{A4}$$

where $\Omega(x, y)$ is the set of patches that contribute to point (x, y) in the saliency map, $|\Omega(x, y)|$ is the cardinality of that set, and $g_p(x, y)$ is a function that maps the coordinates in the saliency map to patch coordinates for a particular patch p. The final saliency map for a module is then smoothed via convolution with a Gaussian kernel to reduce artifacts due to patch edges. This smoothing also helps assure consistency across neighboring saliency values.

Figure A4. PCA activities in TASC at the four ranges of influence (shortest (a) to longest (d)) corresponding to the image in Figure A1a. A separate representation is maintained for each feature channel (here the ordering is: red, green, blue, yellow, 0, 45, 90, 135). PCA is computed at the patch level. Each inner square in these figures corresponds to a PCA value for a specific feature channel and patch. At the shortest range of influence, there are 15×15 patches. At the longest range of influence, there is only one patch per image. In (a)–(c), the most important component is in the top row and less important components are in lower rows. In (d), the principal components decrease in importance from left to right. Lighter values correspond to greater activation.

Combining multiple modules

Several different approaches are possible for combining the saliency maps from multiple modules. A weighted averaging rule will produce the same salience whether many modules have modest activity or a single module has high activity. Consider the task of searching for wheeled vehicles. A strong response obtained from a bike module will be moderated by weak responses from a car or bus module, due to the averaging of module outputs. Instead of averaging, perhaps it makes sense to consider a module combination rule in which the overall saliency at a location is the maximum activation at that location across the subset of modules deemed relevant for a task.

The average and maximum activation rules can be thought of as implementing soft *conjunction* and *disjunction* operators. In the former case, a location is salient to the extent that *every* module produces strong output; in the latter case, a location is salient to the extent that *any* module produces strong output. In the case where the goal is highly specific—requiring only one or a small number of modules—the two rules will produce similar results, because as fewer modules are involved, the average module output and the maximum module output converge on the same value. When goals are less constrained, however, the predictions of these approaches diverge. As described in the body of the paper, we use the max rule in this implementation.

Formally, the max combination rule is defined as follows. Let $s_{o,r}(x, y)$ be the saliency map value for target object o and range of influence r at location (x, y). The values in the combined saliency map are given by

$$s_c(x, y) = \max_{o,r \in M} s_{o,r}(x, y), \quad \text{(A5)}$$

where M is the set of modules relevant to the current goal. In each of the simulations, M is selected to fit the constraints of the search task. In most cases, M contains modules associated with one specific target object, though at times the model combines across multiple target objects.

Acknowledgments

The authors thank Garrison Cottrell and Jeremy Wolfe for their insightful suggestions. The authors also thank Zhaoping Li for comments on an earlier draft of the manuscript. This research was supported by NSF BCS 0339103, NSF CSE-SMA 0509521, and NSF BCS-0720375.

Commercial relationships: none.
Corresponding author: Matthew Wilder.
Email: mattwilder.cu@gmail.com.
Address: Department of Computer Science, University of Colorado, 4595 Brookfield Dr., Boulder, CO 80309, USA.

References

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

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

Baldi, P., & Hornik, K. (1989). Neural networks and principal components analysis: Learning from examples without local minima. *Neural Networks, 2,* 53–58.

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

Broadbent, D. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology, 47,* 191–196.

Brockmole, J., Castelhano, M., & Henderson, J. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 699–706.

Brockmole, J. R., Hambrick, D. Z., Windisch, D. J., & Henderson, J. M. (2008). The role of meaning in contextual cueing: Evidence from chess expertise. *Quarterly Journal of Experimental Psychology, 61,* 1886–1896.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. *Advances in Neural Information Processing Systems, 18,* 155–1632.

Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3):5, 1–24, http://www.journalofvision.org/content/9/3/5, doi:10.1167/9.3.5. [PubMed] [Article]

Bullier, J. (2001). Feedback connections and conscious vision. *Trends in Cognitive Sciences, 5,* 369–370.

Caerwinski, M., Lightfoot, N., & Shiffrin, R. (1992). Automatization and training in visual search. *American Journal of Psychology, 105,* 271–315.

Cerf, M., Frady, E., & Koch, C. (2008). Using semantic content as cues for better scanpath prediction. *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (pp. 143–146). New York, NY, USA: ACM.

Cerf, M., Harel, J., Einhauser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems, 20,* 241–248.

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

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition, 17,* 945–978.

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

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center–surround hypothesis for visual saliency. *Journal of Vision, 8*(7):13, 1–18, http://www.journalofvision.org/content/8/7/13, doi:10.1167/8.7.13. [PubMed] [Article]

Gao, D., & Vasconcelos, N. (2007). *Bottom-up saliency is a discriminant process*. Paper presented at the IEEE International Conference on Computer Vision.

Gawne, T., & Martin, J. (2002). Responses of primate visual cortical v4 neurons to simultaneously presented stimuli. *Journal of Neurophysiology, 88,* 1128–1135.

Geng, J., & Behrmann, M. (2005). Spatial probability as an attentional cue in visual search. *Perception & Psychophysics, 67,* 1252–1268.

Heinke, D., & Humphreys, G. W. (1997). SAIM: A model of visual attention and neglect. In *Proceedings of the 7th*

International Conference on Artificial Neural Networks (pp. 913–918). Lausanne, Switzerland: Springer Verlag.

Heinke, D., & Humphreys, G. (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.

Henderson, J., & Hollingworth, A. (1998). Eye guidance while reading and while watching dynamic scenes. In G. Underwood (Ed.), *Eye movements during scene viewing: An overview* (pp. 269–293). Oxford, UK: Elsevier.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49,* 1295–1306.

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

Itti, L., Koch, C., & Neiber, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 1254–1259.

Julesz, B. (1984). Toward an automatic theory of preattentive vision. In G. M. Edelman, W. E. Gall, & W. M. Cownan (Eds.), *Dynamic aspects of neocortical function* (pp. 595–612). New York: Neurosciences Research Foundation.

Kanan, C., Tong, M., Zhang, L., & Cottrell, G. (2009). Sun: Top-down saliency using natural statistics. *Visual Cognition, 17,* 979–1003.

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

Kristjansson, A. (2006). Rapid learning in attention shifts—A review. *Visual Cognition, 13,* 324–362.

Lampl, I., Ferster, D., Poggio, T., & Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *Journal of Neurophysiology, 92,* 2704–2713.

Lamy, D., Carmel, T., Egeth, H., & Leber, A. (2006). Effects of search mode and intertrial priming on singleton search. *Perception & Psychophysics, 68,* 919–932.

Leonards, U., Rettenbach, R., Nase, G., & Sireteanu, R. (2002). Perceptual learning of highly demanding visual search tasks. *Vision Research, 42,* 2193–2204.

Lowe, D. G. (1991). Fitting parameterized three-dimensional models to images. *Pattern Analysis and Machine Intelligence, 13,* 441–450.

Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. role of features. *Memory and Cognition, 22,* 657–672.

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

Mozer, M., & Baldwin, D. (2008). Experience-guided search: A theory of attentional control. In J. Platt, D. Koller, & Y. Singer (Eds.), *Advances in neural information processing systems* (vol. 20, pp. 1033–1040). Cambridge, MA: MIT Press.

Mozer, M., Shettel, M., & Vecera, S. (2006). Control of visual attention: A rational account. In Y. Weiss, B. Schoelkopf, & J. Plattr (Eds.), *Neural information processing systems* (vol. 18, pp. 923–930). Cambridge, MA: MIT Press.

Mruczek, R., & Sheinberg, D. (2005). Distractor familiarity leads to more efficient visual search for complex stimuli. *Perception & Psychophysics, 67,* 1016–1031.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron, 53,* 605–617.

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

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42,* 145–175.

Olson, I., & Chun, M. (2002). Perceptual constraints on implicit learning of spatial context. *Visual Cognition, 9,* 273–302.

Peterson, M., & Skow-Grant, E. (2003). Memory and learning in figure–ground perception. In B. Ross & D. Irwin (Eds.), *Cognitive vision: Psychology of learning and motivation* (vol. 42, pp. 1–34). New York: Academic Press.

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

Rao, R. P., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research, 42,* 1447–1463.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025.

Russel, B., Torralba, A., & Murphy, K. (2008, May). Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision, 77,* 157–173.

Schmidt, J., & Zelinsky, G. J. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology, 62,* 1904–1914.

Senders, J. (1964). The human operator as a monitor and controller of multidegree of freedom systems. *IEEE Transactions on Human Factors in Electronics, 5,* 2–6.

Siagian, C., & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29,* 300–312.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. *Proceedings of the 2nd IEEE International Conference on Image Processing*. Washington, DC.

Sireteanu, R., & Rettenbach, R. (2000). Perceptual learning in visual search generalizes over tasks, locations, and eyes. *Vision Research, 40,* 2925–2949.

Steinman, S. (1987). Serial and parallel search in pattern vision. *Perception, 16,* 389–399.

Torralba, A. (2001). Contextual modulation of target saliency. *Advances in Neural Information Processing Systems, 14,* 1303–1310.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of Optical Society of America, 20,* 1407–1418.

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

Treisman, A. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology, 12,* 242–248.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and objects. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 194–214.

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

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

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

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

Wolfe, J. (1998b). What can 1,000,000 trials tell us about visual search? *Psychological Science, 9,* 33–39.

Wolfe, J. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 99–119). New York: Oxford.

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

Wolfe, J., & Horowitz, T. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5,* 1–7.

Yu, A., Dayan, P., & Cohen, J. (2009). Dynamics of attentional selection under conflict: Toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 700–717.

Yu, A., Giese, M., & Poggio, T. (2002). Biophysiologically plausible implementations of maximum operation. *Neural Computation, 14,* 2857–2881.

Zhang, L., Tong, M. H., & Cottrell, G. W. (2009). SUNDAy: Saliency using natural statistics for dynamic analysis of scenes. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2944–2949). Austin, TX: Cognitive Science Society.

Zhang, L., Tong, M., Marks, T., Shan, H., & Cottrell, G. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7):32, 1–20, http://www.journalofvision.org/content/8/7/32, doi:10.1167/8.7.32. [PubMed] [Article]

Zhaoping, L., & May, K. A. (2007). Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex. *PLoS Computational Biology, 3*.

Zhaoping, L., & Snowden, R. (2006). A theory of a saliency map in primary visual cortex (V1) tested by psychophysics of colour-orientation interference in texture segmentation. *Visual Cognition, 14,* 911–933.