Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search

Antonio Torralba
Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology

Aude Oliva
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology

Monica S. Castelhano
Department of Psychology and
Cognitive Science Program,
Michigan State University

John M. Henderson
Department of Psychology and
Cognitive Science Program,
Michigan State University

Behavioral experiments have shown that the human visual system makes extensive use of contextual information for facilitating object search in natural scenes. However, the question of how to formally model contextual influences is still open. Based on a Bayesian framework, we present an original approach of attentional guidance by global scene context. Two parallel pathways comprise the model; one pathway computes local features (saliency) and the other computes global (scene-centered) features. The Contextual Guidance model of attention combines bottom-up saliency, scene context and top-down mechanisms at an early stage of visual processing, and predicts the image regions likely to be fixated by human observers performing natural search tasks in real world scenes.

Keywords: attention, eye movements, visual search, context, scene recognition

Introduction

According to feature-integration theory (Treisman & Gelade, 1980) the search for objects requires slow serial scanning since attention is necessary to integrate low-level features into single objects. Current computational models of visual attention based on saliency maps have been inspired by this approach, as it allows a simple and direct implementation of bottom-up attentional mechanisms that are not task specific. Computational models of image saliency (Itti, Kock & Niebur, 1998; Koch & Ullman, 1985; Parkhurst, Law & Niebur, 2002; Rosenholtz, 1999) provide some predictions about which regions are likely to attract observers' attention. These models work best in situations where the image itself provides little semantic information and when no specific task is driving the observer's exploration. In real-world images, the semantic content of the scene, the co-occurrence of objects, and task constraints have been shown to play a key role in modulating where attention and eye movement go (Chun & Jiang, 1998; Davenport & Potter, 2004; DeGraef, 1992; Henderson, 2003; Neider & Zelinski, 2006; Noton & Stark, 1971; Oliva, Torralba, Castelhano & Henderson, 2004; Palmer, 1975; Tsotsos, Culhane, Wai, Lai, Davis & Nuflo, 1995; Yarbus, 1967). Early work by Biederman, Mezzanotte & Rabinowitz (1982) demonstrated that the violation of typical item configuration slows object detection in a scene (e.g., a sofa floating in the air, see also DeGraef, Christianens & d'Ydewalle, 1990; Henderson, Weeks & Hollingworth, 1999). Interestingly, human observers need not be explicitly aware of the scene context to benefit from it. Chun, Jiang and colleagues have shown that repeated exposure to the same arrangement of random elements produces a form of learning that they call contextual cueing (Chun & Jiang, 1998, 1999; Chun, 2000; Jiang, & Wagner, 2004; Olson & Chun, 2002). When repeated configurations of distractor elements serve as predictors of target location, observer's are implicitly cued to the position of the target in subsequent viewing of the repeated displays. Observer's can also be implicitly cued to a target location by global properties of the image like color background (Kunar, Flusberg & Wolfe, 2006) and when learning meaningful scenes background (Brockmole & Henderson, 2006; Brockmole, Castelhano & Henderson, in press; Hidalgo-Sotelo, Oliva & Torralba, 2005; Oliva, Wolfe & Arsenio, 2004).

One common conceptualization of contextual information is based on exploiting the relationship between co-occurring objects in real world environments (Bar, 2004; Biederman, 1990; Davenport & Potter, 2004; Friedman, 1979; Henderson, Pollatsek & Rayner, 1987). In this paper we discuss an alternative representation of context that does not require parsing a scene into objects, but instead relies on global statistical properties of the image (Oliva & Torralba, 2001). The proposed representation provides the basis for feedforward processing of visual context that can be performed in parallel with object processing. Global context can thus benefit object search mechanisms by modulating the use of the features provided by local image analysis. In our Contex-

tual Guidance model, we show how contextual information can be integrated prior to the first saccade, thereby reducing the number of image locations that need to be considered by object-driven attentional mechanisms.

Recent behavioral and modeling research suggests that early scene interpretation may be influenced by global image properties that are computed by processes that do not require selective visual attention (Spatial Envelope properties of a scene, Oliva & Torralba, 2001, statistical properties of object sets, Ariely, 2001; Chong & Treisman, 2003). Behavioral studies have shown that complex scenes can be identified from a coding of spatial relationships between components like geons (Biederman, 1995) or low spatial frequency blobs (Schyns & Oliva, 1994). Here we show that the structure of a scene can be represented by the mean of global image features at a coarse spatial resolution (Oliva & Torralba, 2001, 2006). This representation is free of segmentation and object recognition stages while providing an efficient shortcut for object detection in the real world. Task information (searching for a specific object) modifies the way that contextual features are used to select relevant image regions.

The *Contextual Guidance* model (Fig. 1) combines both local and global sources of information within the same Bayesian framework (Torralba, 2003). Image saliency and global-context features are computed in parallel, in a feed-forward manner and are integrated at an early stage of visual processing (i.e., before initiating image exploration). Top-down control is represented by the specific constraints of the search task (looking for a pedestrian, a painting, or a mug) and it modifies how global-context features are used to select relevant image regions for exploration.

Model of object search and contextual guidance

Scene Context recognition without object recognition

Contextual influences can arise from different sources of visual information. On the one hand, context can be framed as the relationship between objects (Bar, 2004; Biederman, 1990; Davenport & Potter, 2004; Friedman, 1979; Henderson et al., 1987). According to this view, scene context is defined as a combination of objects that have been associated over time and are capable of priming each other to facilitate scene categorization. To acquire this type of context, the observer must perceive a number of diagnostic objects within the scene (e.g., a bed) and use this knowledge to infer the probable identities and locations of other objects (e.g., a pillow). Over the past decade, research on the change blindness has shown that in order to perceive the details of an object, one must attend to it (Henderson & Hollingworth, 1999; Hollingworth, Schrock & Henderson, 2001; Hollingworth & Henderson, 2002; Rensink, 2000; Rensink, O'Regan & Clark, 1997; Simons & Levin, 1997). In light of these results, object-to-object context would be built as a serial process that will first require perception of diagnostic objects before inferring associated objects. In theory, this process could take place within an initial glance with attention being able to grasp 3 to 4 objects within a 200 msec window (Vogel, Woodman & Luck, in press; Wolfe, 1998). Contextual influences induced by co-occurrence of objects have been observed in cognitive neuroscience studies. Recent work by Bar and collaborators (2003, 2004) demonstrates that specific cortical areas (a subregion of the parahippocampal cortex and the retrosplenial cortex) are involved in the analysis of contextual associations (e.g., a farm and a cow) and not merely in the analysis of scene layout.

Alternatively, research has shown that scene context can be built in a holistic fashion, without recognizing individual objects. The semantic category of most real-world scenes can be inferred from their spatial layout only (e.g., an arrangement of basic geometrical forms such as simple Geons clusters, Biederman, 1995; the spatial relationships between regions or blobs of particular size and aspect ratio, Oliva & Schyns, 2000; Sanocki & Epstein, 1997; Schyns & Oliva, 1994). A blurred image in which object identities cannot be inferred based solely on local information, can be very quickly interpreted by human observers (Oliva & Schyns, 2000). Recent behavioral experiments have shown that even low level features, like the spatial distribution of colored regions (Goffaux & et al., 2005; Oliva & Schyns, 2000; Rousselet, Joubert & Fabre-Thorpe, 2005) or the distribution of scales and orientations (McCotter, Gosselin, Cotter & Schyns, 2005) can reliably predict the semantic classes of real world scenes. Scene comprehension and more generally recognition of objects in scenes can occur very quickly, without much need for attentional resources. This rapid understanding phenomenon has been observed under different experimental conditions, where the perception of the image is difficult or degraded, like during RSVP tasks (Evans & Treisman, 2006; Potter, 1976; Potter, Staub & O'Connor, 2004), very short presentation time (Thorpe et al., 1996), backward masking (Bacon-Mace, Mace, Fabre-Thorpe & Thorpe, 2005), dual-task conditions (Li, VanRullen, Koch, & Perona, 2002) and blur (Schyns & Oliva, 1994; Oliva & Schyns, 1997). Cognitive neuroscience research has shown that these recognition events would occur 150 msec after image onset (Delorme, Rousselet, Mace & Fabre-Thorpe, M., 2003; Goffaux et al., 2005; Johnson and Olshausen, 2005; Thorpe, Fize & Marlot, 1996). This establishes an upper bound on how fast natural image recognition can be made by the visual system, and suggest that natural scene recognition can be implemented within a feed-forward mechanism of information processing. The global features approach described here may be part of a feed-forward mechanism of semantic scene analysis (Oliva & Torralba, 2006).

Correspondingly, computational modeling work has shown that real world scenes can be interpreted as a member of a basic-level category based on holistic mechanisms, without the need for segmentation and grouping stages (Fei Fei & Perona, 2005; Oliva & Torralba, 2001; Walker Renninger & Malik, 2004; Vogel & Schiele, in press). This scene-centered approach is consistent within a global-to-local image analysis (Navon, 1977) where the processing of the global structure and the spatial relationships among components precede

the analysis of local details. Cognitive neuroscience studies have acknowledged the possible independence between processing a whole scene and processing local objects within in an image. The parahippocampal place area (PPA) is sensitive to the scene layout and remains unaffected by the visual complexity of the image (Epstein & Kanwisher, 1998), a virtue of the global feature coding proposed in the current study. The PPA is also sensitive to scene processing that does not require attentional resources (Marois, Yi & Chun, 2004). Recently, Goh, Siong, Park, Gutchess, Hebrank & Chee (2004) showed activation in different brain regions when a picture of a scene background was processed alone, compared to backgrounds that contained a prominent and semantically-consistent object. Whether the two approaches to scene context, one based on holistic global features and the other one based on object associations recruit different brain regions (for reviews, see Bar, 2004; Epstein, 2005; Kanwisher, 2003), or instead recruit a similar mechanism processing spatial and conceptual associations (Bar, 2004), are challenging questions for insights into scene understanding.

A scene-centered approach of context would not preclude a parallel object-to-object context, rather it would serve as a feed-forward pathway of visual processing, describing spatial layout and conceptual information (e.g. scene category, function), without the need of segmenting the objects. In this paper, we provide a computational implementation of a scene-centered approach to scene context, and show its performance in predicting eye movements during a number of ecological search tasks.

In the next section we present a contextual model for object search that incorporates global features(scene-centered context representation) and local image features (salient regions).

Model of object search and contextual guidance

We summarize a probabilistic framework of attentional guidance that provides, for each image location, the probability of target presence by integrating global and local image information and task constraints. Attentional mechanisms such as image saliency and contextual modulation emerge as a natural consequence from such a model (Torralba, 2003).

There has been extensive research on the relationship between eye movements and attention and it has been well established that shifts of attention can occur independent of eye movements (for reviews see Henderson, 2005; Liversedge & Findlay, 2000; Rayner, 1998). Furthermore, the planning of an eye movement is itself thought to be preceded by a shift of overt attention to the target location before the actual movement is deployed (Deubel & Schnerider, 1996; Hoffman & Subramaniam, 1995; Kowler, Anderson, Dosher & Blaser, 1995; Rayner, McConkie & Ehrlich, 1978; Rayner, 1998; Remington, 1980). However, previous studies have also shown that with natural scenes and other complex stimuli (such as reading), the cost of moving the eyes to shift attention is less than to shift attention covertly, and led some to posit that studying covert and overt attention as separate processes in these cases is misguided (Findlay, 2004). The model proposed in the current study attempts to predict the image regions that will be explored by covert and overt attentional shifts, but performance of the model is evaluated with overt attention as measured with eye movements.

In the case of a search task in which we have to look for a target embedded in a scene, the goal is to identify whether the target is present or absent, and if present, to indicate where it is located. An ideal observer will fixate the image locations that have the highest probability of containing the target object given the available image information. Therefore, detection can be formulated as the evaluation of the probability function $p(O,X|I)$ where I is the set of features extracted from the image. O is a binary variable where $O=1$ denotes target present and $O=0$ denotes target absent in the image. X defines the location of the target in the image when the target is present ($O=1$). When the target is absent $p(O=0,X|I) \propto p(O=0|I)$.

In general, this probability will be difficult to evaluate due to the high dimensionality of the input image I. One common simplification is to make the assumption that the only features relevant for evaluating the probability of target presence are the local image features. Many experimental displays are set up in order to verify that assumption (e.g., Wolfe 1994). In the case of search in real-world scenes, local information is not the only information available and scene based context information can have a very important role when the fixation is far from the location of the target. Before attention is directed to a particular location, the non-attended object corresponds to a shapeless bundle of basic features insufficient for confident detection (Wolfe & Bennett, 1997). The role of the scene context is to provide information about past search experiences in similar environments and strategies that were successful in finding the target. In our model, we use two sets of image features: local and global features. Local features characterize a localized region of the image; global features characterize the entire image. Target detection is then achieved by estimating $p(O,X|L,G)$. This is the probability of the presence of the target object at the location $X=(x,y)$ given the set of local measurements $L(X)$ and a set of global features G. The location X is defined in an image centered coordinates frame. In our implementation, the image coordinates are normalized so that x is in the range $[0,1]$. The choice of units or the image resolution does not affect the model predictions. The global features G provide the context representation.

Using Bayes' rule we can split the target presence probability function into a set of components that can be interpreted in terms of different mechanisms that contribute to the guidance of attention (Torralba, 2003):

$$p(O=1,X|L,G) = \frac{1}{p(L|G)} p(L|O=1,X,G)p(X|O=1,G)p(O=1|G) \quad (1)$$

a) The first term, $1/p(L|G)$, does not depend on the target, and therefore is a pure bottom-up factor. It provides a measure of how unlikely it is to find a set of local measurements within the image. This term fits the definition of saliency

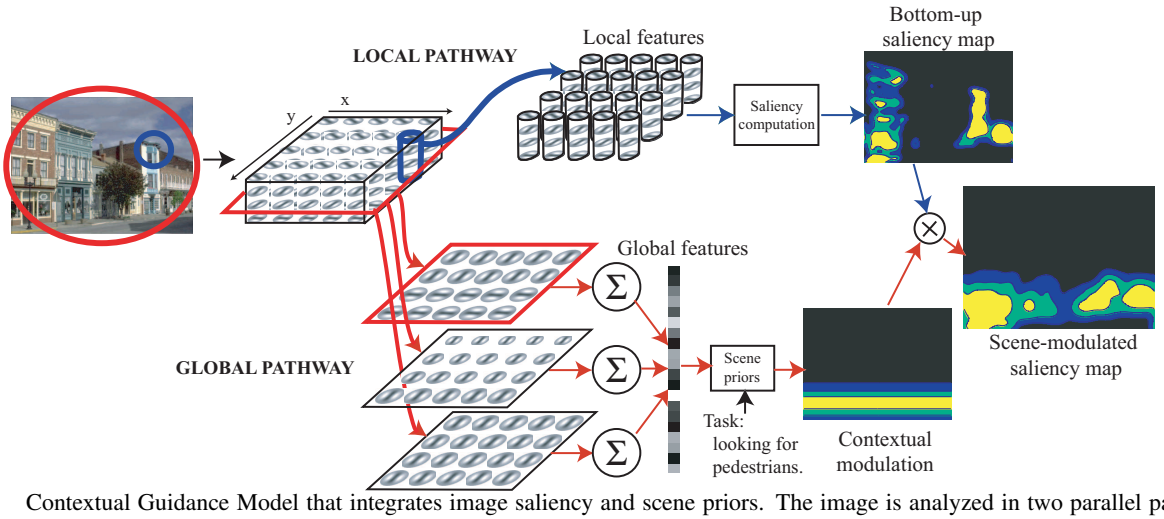

Figure 1. Contextual Guidance Model that integrates image saliency and scene priors. The image is analyzed in two parallel pathways. Both pathways share the first stage in which the image is filtered by a set of multiscale oriented filters. The local pathway represents each spatial location independently. This local representation is used to compute image saliency and to perform object recognition based on local appearance. The global pathway represents the entire image holistically by extracting global statistics from the image. This global representation can be used for scene recognition. In this model the global pathway is used to provide information about the expected location of the target in the image.

(Koch & Ullman, 1985; Itti et al., 1998; Treisman & Gelade, 1980) and emerges naturally from the probabilistic framework (Rosenholtz, 1999; Torralba, 2003).

b) The second term, $p(L|O=1,X,G)$, represents the top-down knowledge of the target appearance and how it contributes to the search. Regions of the image with features unlikely to belong to the target object are vetoed and regions with attended features are enhanced (Rao, Zelinsky, Hayhoe & Ballard, 2002; Wolfe, 1994).

c) The third term, $p(X|O=1,G)$, provides context-based priors on the location of the target. It relies on past experience to learn the relationship between target locations and global scene features (Biederman, Mezzanotte & Rabinowitz, 1982; Brockmole & Henderson, in press; Brockmole & Henderson, 2006; Brockmole, Castelhano & Henderson, in press; Chun & Jiang, 1998; 1999; Chun, 2000; Hidalgo-Sotelo, Oliva & Torralba, 2005; Kunar, Flusberg & Wolfe, 2006; Oliva, Wolfe & Arsenio, 2004; Olson & Chun, 2001; Torralba, 2003).

d) The fourth term, $p(O=1|G)$, provides the probability of presence of the target in the scene. If this probability is very small, then object search need not be initiated. In the images selected for our experiments, this probability can be assumed to be constant and therefore we have ignored it in the present study. In a general setup this distribution can be learnt from training data (Torralba, 2003).

The model given by eq. (2) does not specify the temporal dynamics for the evaluation of each term. Our hypothesis is that both saliency and global contextual factors are evaluated very quickly, before the first saccade is deployed. However, the factor that accounts for target appearance might need longer integration time, particularly when the features that define the object are complex combinations of low-level image primitives (like feature conjunctions of orientations and colors, shapes, etc.) that require attention to be focused on a local image region (we assume also that, in most cases, the objects are relatively small). This is certainly true for most real-world objects in real-world scenes, since no simple feature is likely to distinguish targets from non-targets.

In this paper we consider the contribution of saliency and contextual scene priors, excluding any contribution from the appearance of the target. Therefore, the final model used to predict fixation locations, integrating bottom-up saliency and task dependent scene priors, is described by the equation:

$$S(X) = \frac{1}{p(L|G)} p(X|O=1,G) \qquad (2)$$

The function $S(X)$ is a contextually modulated saliency map that is constrained by the task (searching the target). This model is summarized in Fig. 1. In the local pathway, each location in the visual field is represented by a vector of features. It could be a collection of templates (e.g., mid-level complexity patches, Ullman, Vidal-Naquet & Sali, 2002) or a vector composed of the output of wavelets at different orientations and scales (Itti et al., 1998; Reisenhuber & Poggio, 1999). The local pathway (object centered) refers principally to bottom-up saliency models of attention (Itti et al., 1998) and appearance-based object recognition (Rao et al., 2002). The global pathway (scene centered) is responsible for both the representation of the scene- the basis for scene recognition- and the contextual modulation of image saliency and detection response. In this model, the gist of the scene (here represented by the global features G) is acquired during the first few hundred milliseconds after the image onset (while the eyes are still looking at the location of the initial fixation point). Finding the target requires scene exploration. Eye movements are needed as the target can be small (people

in a street scene, a mug in a kitchen scene, etc.). The locations to which the first fixations are directed will be strongly driven by the scene gist when it provides expectations about the location of the target.

In the next subsections we summarize how the features and each factor of eq. (2) are evaluated.

Local features and Saliency

Bottom-up models of attention (Itti et al., 1998) provide a measure of the saliency of each location in the image computed from various low level features (contrast, color, orientation, texture, motion). In the present model, saliency is defined in terms of the probability of finding a set of local features within the image as derived from the Bayesian framework. Local image features are salient when they are statistically distinguishable from the background (Rosenholtz, 1999; Torralba, 2003). The hypothesis underlying these models is that locations with different properties from their neighboring regions are considered more informative and therefore will initially attract attention and eye movements. In the task of an object search, this interpretation of saliency follows the intuition that repetitive image features are likely to belong to the background whereas rare image features are more likely to be diagnostic in detecting objects of interest (Fig. 2)

In our implementation of saliency, each color channel (we use the raw R,G,B color channels) is passed through a bank of filters (we use the Steerable pyramid, Simoncelli & Freeman, 1995) tuned to 6 orientations and 4 scales (with 1 octave separation between scales) which provide a total of 6x4x3 = 72 features at each location. Each image location is represented by a vector of features (L) that contains the output of the multiscale oriented filters for each color band. Computing saliency requires estimating the distribution of local features in the image. In order to model this distribution, we use a multivariate power-exponential distribution, which is more general than a Gaussian distribution and accounts for the long tails of the distributions typical of natural images (Olshausen & Field, 1996):

$$\log p(L) = \log k - \frac{1}{2}\left[(L-\eta)^t \Delta^{-1}(L-\eta)\right]^{\alpha} \qquad (3)$$

where k is a normalization constant, η and Δ are the mean and covariance matrix of the local features. The exponent α (with $\alpha < 1$) accounts for the long tail of the distribution. When $\alpha = 1$ the distribution is a multivariate Gaussian. We use maximum likelihood to fit the distribution parameters η, Δ and α. For α we obtain values in the range of $[0.01, 0.1]$ for the images used in the eye movement experiments reported below. This distribution can also be fitted by constraining Δ to be diagonal and then allowing the exponent α to be different for each component of the vector of local features L. We found no differences between these two approximations when using this probability for predicting fixation points. We approximate the conditional distribution $p(L|G) \simeq p(L|\eta(I), \Delta(I), \alpha(I))$ by fitting the power-exponential distribution using the features computed at the current image I.

The computation of saliency does not take into account the target appearance, and so it will be a weak predictor of the target location for many objects. Fig. 2 shows the saliency measured in several indoor and outdoor scenes along with the relative saliency of several objects computed over a large database of annotated images (the number of images used for each object varies from 50 to 800). To provide a better local measure of saliency, the inverse probability is first raised to the power of $\gamma = 0.05$ and then the result is smoothed with a Gaussian filter (with a half-amplitude spatial width of $\sigma = 1$ degree of visual angle). The exponent γ was selected according to the description provided in eq. (7), and the smoothing filter was selected in order to maximize the saliency of people in street scenes (we found the parameters to be insensitive to the target class for the object categories used in this study). The size σ of the smoothing filter is related to the average size of the target in the scenes and to the dispersion of eye fixations around a location of interest. We found that the parameters γ and σ did not differ significantly when optimizing the model for different objects. Therefore we fixed the parameters and used them for different targets (Fig. 2).

This measure of saliency will provide the baseline model to which we will compare the results of our model, which integrates contextual information to predict the regions fixated by observers.

Global image features

The statistical regularities of band-pass filter outputs (similar to receptive fields of cells found in the visual cortex, Olshausen & Field 1996) have been shown to be correlated with high-level properties of real-world scenes (Oliva & Torralba, 2001; Oliva & Schyns, 2000; Vailaya et al., 1998). For instance, the degree of perspective or the mean depth of the space that a scene image subtends can be estimated by a configuration of low-level image features (Torralba & Oliva 2002, 2003). Evidence from the psychophysics literature suggests that our visual system computes a global statistical summary of the image in a pre-selective stage of visual processing or at least, with minimal attentional resources (mean orientation, Parkes et al., 2001; mean of set of objects, Ariely, 2001; Chong & Treisman, 2003). By pooling together the activity of local low-level feature detectors across large regions of the visual field, we can build a holistic and low-dimensional representation of the structure of a scene that does not require explicit segmentation of image regions and objects and therefore, requires low amounts of computational (or attentional) resources. This suggests that a reliable scene representation can be built, in a feed-forward manner, from the same low-level features used for local neural representations of an image (receptive fields of early visual areas, Hubel & Wiesel, 1968).

As in Oliva & Torralba (2001), we adopted a representation of the image context using a set of "global features" that provides a holistic description of the spatial organization of dominant scales and orientations in the image. The number of global features that can be computed is quite high. The most effective global features will be those that reflect the

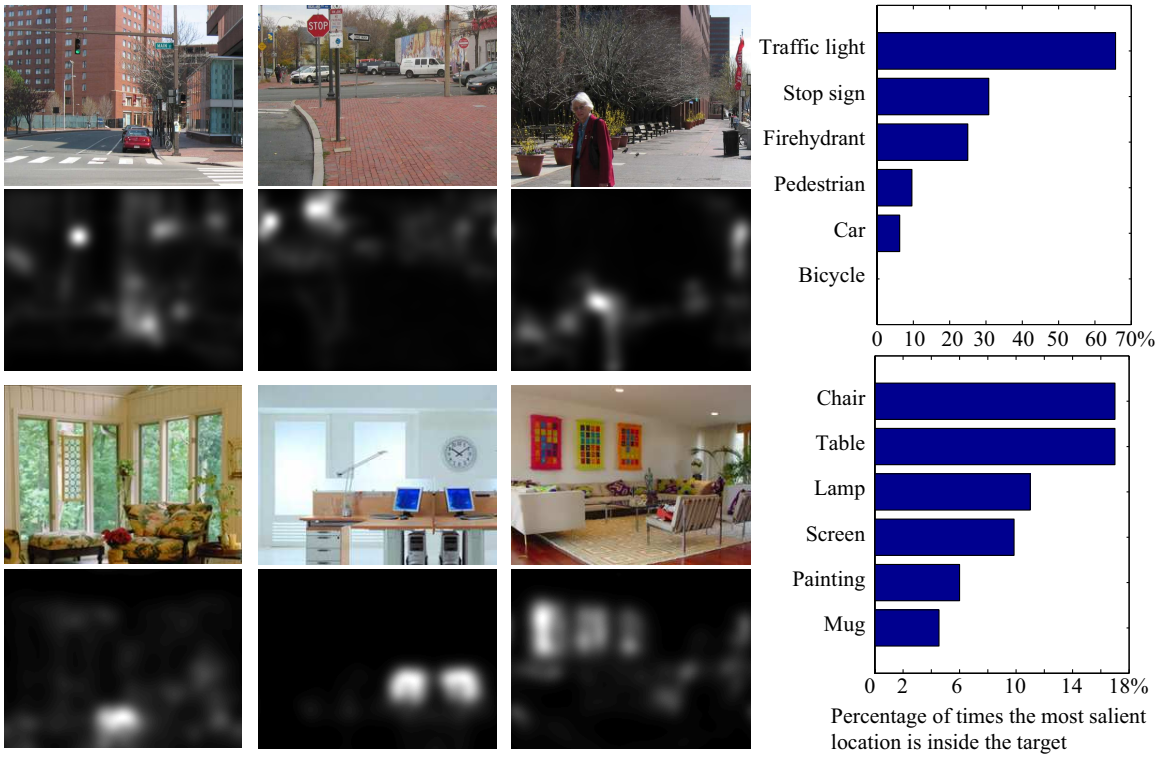

Figure 2. Examples of image saliency. The graph on the right shows a bar for each object corresponding to the percentage of times that the most salient location in the image was inside the target object. These percentages are averages computed over a database with hundred images for each object class (Russell, Torralba, Murphy, Freeman, 2005). Long bars correspond to salient objects. Traffic lights have the highest saliency with 65% of times being the most salient object in the scenes analyzed. People are less salient than many other objects in outdoor scenes: Pedestrians were the most salient object in only 10% of the scene images. Bicycles never contain the most salient point in any of the images analyzed. Tables and chairs are among the most salient objects in indoor scenes.

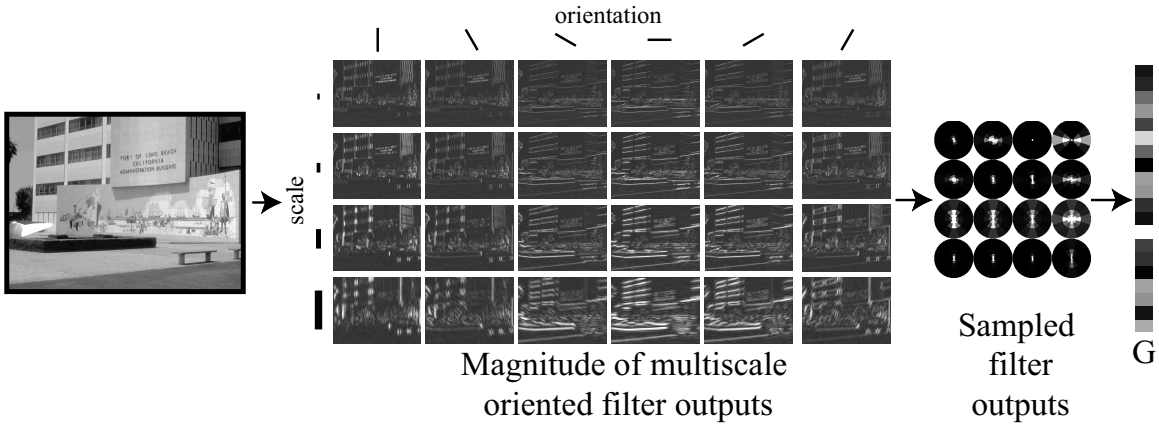

Figure 3. Computation of global features. The Luminance channel is decomposed using a Steerable pyramid with 6 orientations and 4 scales. The output of each filter is subsampled by first taking the magnitude and then computing the local average response over 4x4 non-overlapping windows. The sampled filter outputs are shown here using a polar representation at each location (the polar plots encode scale of the filter in the radius and orientation of tuning in the angle. The brightness corresponds to the output magnitude). The final representation is obtained by projecting the subsampled filter outputs (which represents a vector of 384 dimensions) into its first 64 principal components.

global structures of the visual world. Several methods of image analysis can be used to learn a suitable basis of global features (Fei-Fei & Perona, 2005; Oliva & Torralba, 2001; Vailaya, Jain & Zhang, 1998; Vogel & Schiele, in press) that capture the statistical regularities of natural images. In the modeling presented here, we only consider global features that summarize the statistics of the outputs of receptive fields measuring orientations and spatial frequencies of image components (Fig. 3).

By pooling together the activity of local low-level feature detectors across large regions of the visual field, we can build an holistic and low-dimensional representation of the scene context that is independent of the amount of clutter in the image. The global features are computed starting with the same low level features as the ones used for computing the local features. The Luminance channel (computed as the average of the R, G, B channels) is decomposed using a steerable pyramid (Simoncelli & Freeman, 1995) with 6 orientations and 4 spatial frequency scales. The output of each filter is subsampled by first taking the magnitude of the response and then computing the local average over 4x4 non-overlapping spatial windows. Each image is then represented by a vector of NxNxK=4x4x24=384 values (where K is the number of different orientations and scales; NxN is the number of samples used to encode, in low-resolution, the output magnitude of each filter). The final vector of global features (G) is obtained by projecting the subsampled filter outputs into its first 64 principal components (PC), obtained by applying principal component analysis (PCA) to a collection of 22000 images (the image collection includes scenes from a full range of views, from close-up to panoramic, for both man-made and natural environments). Fig. 4 shows the first PCs of the output magnitude of simple cells for the Luminance channel for a spatial resolution of 2 cycles per image (this resolution refers to the resolution at which the magnitude of each filter output is reduced before applying the PCA. 2 cycles/image corresponds to NxN = 4x4). Each polar plot in Fig. 4 (low spatial frequencies in the center) illustrates how the scales and orientations are weighted at each spatial location in order to calculate global features. Each of the 24 PCs shown in Fig. 4 is tuned to a particular spatial configuration of scales and orientations in the image. For instance, the second PC responds strongly to images with more texture in the upper half than on the bottom half. This global feature will represent the structure of a natural landscape well, for instance a landscape scene with a road or snow at the bottom and a lush forest at the top. Higher-order PCs have an increasing degree of complexity (Oliva & Torralba, 2006).

In order to illustrate the amount of information preserved by the global features, Fig. 5 shows noise images that are coerced to have the same global features as the target image. This constraint is imposed by an iterative algorithm. The synthetic images are initialized to be white noise. At each iteration, the noise is decomposed using the bank of multi-scale oriented filters and their outputs are modified locally to match the global features of the target image. This procedure is similar to the one used in texture synthesis (Portilla & Simoncelli, 2000). The resulting representation provides a coarse encoding of the edges, and textures in the original scene picture. Despite its shapeless representation, the sketch of the image is meaningful enough to support an inference of the probable category of the scene (Oliva & Torralba, 2002).

From a computational stance, estimating the overall structure or shape of a scene as a combination of global features is a critical advantage as it provides a mechanism of visual understanding that is independent of an image's visual complexity. Any mechanisms parsing the image into regions would be dependent on the amount of clutter and occlusions between objects: the more objects to be parsed the more computational resources needed.

Learning context and the layered structure of natural images

The role of the global features in this model is to activate the locations most likely to contain the target object, thereby reducing the saliency of image regions not relevant for the task. The use of context requires a learning stage in which the system learns the association of the scene with the target location. When searching for people, for example, the system learns the correlation between global scene features and the location of people in the image. Such an association is represented in our model by the joint density function $p(X,G|O=1)$. This function will be different for each object category.

The relationship between global scene features and target location is non-linear. We model this relationship by approximating the joint density with a mixture of gaussians. The mixture of gaussians allows for an intuitive description of the behavior of the model as using a set of scene prototypes. Each prototype is associated with one distribution of target locations. When the input image has a set of global features that are similar to one of the prototypes, the expected location of the target will be close to the location of the target associated with the prototype. In a general situation, the expected target location will be a weighted mixture of the target locations for all the prototypes, with the weights depending on how close the current image is to one of the prototypes. The joint density is written as:

$$p(X,G|O=1) = \sum_{n=1}^{N} P(n)p(X|n)p(G|n) = \sum_{n=1}^{N} \pi_n \mathcal{N}(X;\mu_n,\Lambda_n)\mathcal{N}(G;\zeta_n,\Upsilon_n) \quad (4)$$

where $\mathcal{N}$ denotes the Gaussian distribution and N is the number of clusters (prototypes). X is the target location and G are the global features of the scene picture. The first factor, $P(n)=\pi_n$, is the weight assigned to the scene prototype n. The weights are normalized such that $\sum_{n=1}^{N}\pi_n = 1$. The second factor, $p(X|n)$ is the distribution of target locations for the prototype n. This distribution is a Gaussian with mean μ_n and covariance Λ_n. The third factor, $p(G|n)$, is the distribution of global features for prototype n and is a Gaussian with

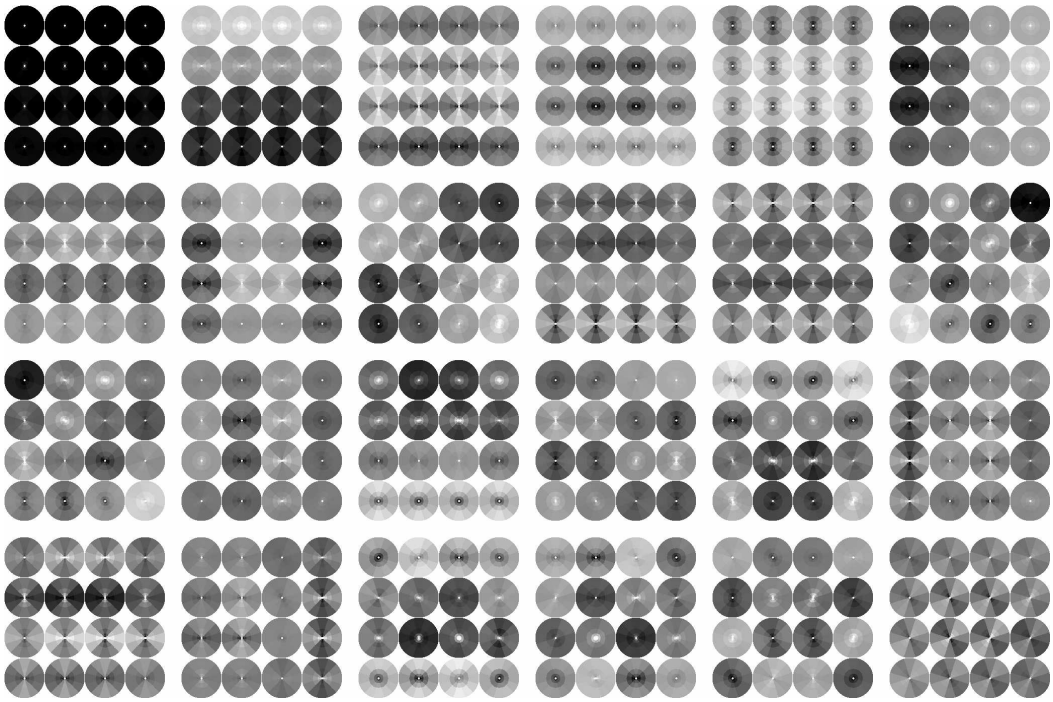

Figure 4. The figure shows the first 24 principal components (PC) of the output magnitude of a set of multiscale oriented filters tuned to six orientations and four scales at 4x4 spatial locations. Each subimage shows, in a polar plot (as in Fig. 3), how the scale and orientations are weighted at each spatial location. The first PC (shown in the top-left panel) has uniform weights. The second component weights positively energy in the upper half of the image and negatively in the bottom half (across all orientations and scales). The third component opposes horizontal (positively) and vertical (negatively) edges anywhere in the image. The fourth component opposes low spatial frequencies against high spatial frequencies anywhere in the image. High order components have more complex interactions between space and spectral content.

Figure 5. Top row: original images. Bottom row: noise images coerced to have the same global features (N=64) as the target image.

mean ζ_n and covariance Υ_n. The vector ζ_n is the vector of global features for the scene prototype n.

There is an important improvement in performance by using cluster-weighted regression instead of the mixture of Gaussians of eq. 5. This requires just a small modification to eq. 5 by replacing $p(X|n)$ with $p(X|G,n)$. In this case we allow for the distribution of target locations for each cluster to depend on the global features. The goal of this model is to learn the local mapping between variations in the target location and small variations of the global features with respect to the prototype. The simplest model is obtained by assuming that in the neighborhood of a prototype the relationship between global features and target location can be approximated by a linear function: $p(X|G,n) = \mathcal{N}(X;\mu_n + W_n G, \Lambda_n)$ where the new parameter W_n is the regression matrix. This is the model that we will use in the rest of the paper.

From the joint distribution we can computed the conditional density function required to compute the contextually modulated saliency (eq. 2):

$$p(X|O=1,G) = \frac{p(X,G|O=1)}{\sum_{n=1}^{N} P(n)p(G|n)} \qquad (5)$$

The conditional expected location of the target X_t, for an image with global features G, is the weighted sum of N linear regressors:

$$X_t = \frac{\sum_{n=1}^{N}(\mu_n + W_n G)w_n}{\sum_{n=1}^{N} w_n} \quad (6)$$

with weights $w_n = \pi_n \mathcal{N}(G;\zeta_n,\Upsilon_n)$. Note that X_t has a non-linear dependency with respect to the global image features.

Global context can predict the vertical location of an object class, but it is hard to predict the horizontal location of the target in a large scene. The reason is that the horizontal location of an object is essentially unconstrained by global context. Instances of one object category are likely to be within a horizontal section of the image. This is generally true for scene pictures of a large space taken by a human standing on the ground. The layered structure of images of large spaces is illustrated in Fig. 6. In order to provide an upper bound on how well the global context can constraint the location of a target in the scene, we can study how well the location of a target is constrained given that we know the location of another target of the same object class within the same image. From a large database of annotated scenes (Russell, Torralba, Murphy & Freeman, 2005) we estimated the joint distribution $p(X_1,X_2)$ where X_1 and X_2 are the locations of two object instances from the same class. We approximated this density by a full covariance Gaussian distribution. We then compared two distributions: the marginal $p(X_1)$ and the conditional $p(X_1|X_2)$. The distribution $p(X_1)$ denotes the variability of target locations within the database. The images are cropped so that this distribution is close to uniform. The dashed ellipses in Fig. 6 show the covariance matrix for the location distribution of several indoor and outdoor objects. The conditional distribution $p(X_1|X_2)$ informs about how the uncertainty on the target location X_1 decreases when we know the location X_2, another instance of the same class. The solid ellipse in Fig. 6 shows the covariance of the conditional gaussian. The variance across the vertical axis is significantly reduced for almost all of the objects, which implies that the vertical location can be estimated quite accurately. However, the variance across the horizontal axis is almost identical to the original variance showing that the horizontal locations of two target instances are largely independent. In fact, objects can move freely along a horizontal line with relatively few restrictions. In particular, this is the case for pedestrians in street pictures.

Therefore, for most object classes we can approximate $p(X|O=1,G) = p(x|O=1,G)p(y|O=1,G)$ and set $p(x|O=1,G)$ to be uniform and just learn $p(y|O=1,G)$. This drastically reduces the amount of training data required to learn the relationship between global features and target location.

The parameters of the model are obtained using a training dataset and the EM algorithm for fitting Gaussian mixtures (Dempster, Laird & Rubin, 1977). We trained the model to predict the locations of three different objects: people in street scenes, and paintings and mugs in indoor scenes. For the people detection task, the training set consists of 279 high resolution pictures of urban environments in the Boston area. For the mug and painting search, the training set was composed respectively of 341 and 339 images of indoors scenes. The images were labeled in order to provide the location of people, mugs, and paintings.

From each image in the training dataset we generated 20 images, of size 320x240 pixels, by randomly cropping the original image in order to create a larger training set with a uniform distribution of target locations. The number of prototypes (N) was selected by cross-validation and depended on the task and scene variability. For the three objects (people, paintings, and mugs), results obtained with $N = 4$ were satisfactory, with no improvement added with the use of more prototypes. Fig. 7 shows a set of images that have similar features to the prototypes selected by the learning stage for solving the task of people detection in urban scenes.

Finally, the combination of saliency and scene priors requires weighting the two factors so that the product is not constantly dominated by one factor. This is a common problem when combining distributions with high dimensional inputs that were independently trained. One common solution is to apply an exponent to the local evidence:

$$S(X) = p(L|G)^{-\gamma} p(X|O=1,G) \quad (7)$$

The parameter γ is set by sequentially searching for the best γ on a validation set. The optimization was achieved by using people as the target object. However, we found this parameter had a small effect when the target object was changed. The parameter γ was then fixed for all the experiments. The best value for γ is 0.05 (performance is similar for γ in the range $[0.01, 0.3]$). A small value for γ has the effect of down-weighting the importance of saliency with respect to contextual information. Note that this exponent has no effect on the performance of each independent module, and only affects the performance of the final model. We smooth the map $S(X)$ using a Gaussian window with a half-amplitude spatial width of 1 degree of visual angle. This provides an estimation of the probability mass across image regions of 1 degree of visual angle. Only two parameters that have been tuned to combine saliency and the scene prior: the width of the blur filter (that specifies over which region the saliency will be integrated) and the exponent (to weight the mixture of saliency and scene priors). Despite the fact that those parameters were optimized in a first instance in order to maximize the saliency of people in outdoor scenes, we found that the optimal parameters do not change from object to object. In all our experiments, those parameters are fixed. Therefore, they are not object specific.

Fig. 8 depicts the system's performance on a novel image. Two models are computed, one using salient regions alone and one using the contextual guidance model. The red dots indicate the real location of the target objects (pedestrians) in the image. The bar plot indicates the percentage of target objects that are within the attended region (set to be 20% of the image size) when using low-level saliency alone, contextual priors alone, or a combination of both factors. In each of the three cases performance is clearly above chance (20%), with the saliency model performing at 50 %. Performance

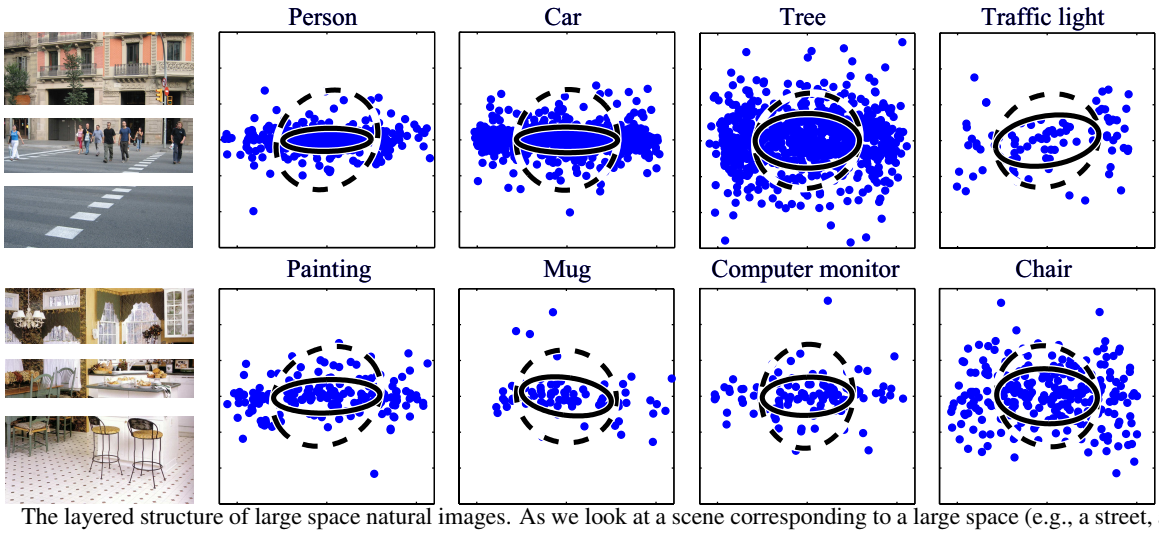

Figure 6. The layered structure of large space natural images. As we look at a scene corresponding to a large space (e.g., a street, an office, a living room), the objects on the scene seem to be organized along horizontal layers. For instance, in a street scene, we will have the road in the bottom; in the center we will have cars, pedestrians. Above this layer we will have trees, buildings and at the top the sky. If we move out eyes horizontally we will encounter objects of similar categories. On the other hand, if we move the eyes vertically we will encounter objects of quite different categories. This figure shows, by collecting statistics from a large database of annotated images, that objects of the same category are clustered along a similar vertical position while their horizontal location is mostly unconstrained. Each plot shows the covariances of the distributions $p(X_1)$ (dashed line) and $p(X_1|X_2)$ (solid line) for eight object categories. X_1 and X_2 are the locations of two object instances from the same class. The dots represent $X_1 - E[X_1|X_2]$; the location of each object relative to its expected location given that we know the location of another instance of the same object class in the same image. For each plot, the center corresponds to the coordinates $(0,0)$.

Figure 7. Scene prototypes selected for the people search task in urban scenes. The top row shows the images from the training set that are the closest to the four prototypes found by the learning algorithm. The bottom row shows the expected location of pedestrians associated with each prototype. The selected regions are aligned with the location of the horizon line.

reaches 83% when both saliency and scene priors are integrated. These results show that the use of contextual information in a search task provides a significant benefit over models that use bottom-up saliency alone for predicting the location of the target.

Eye Movement Experiment

A search experiment was designed to test the assumptions made by the model by having three groups of participants search for respectively, people, paintings and mugs in scenes images. The three tasks were selected to correspond to different contextual constraints encountered in the real world: people were defined as pedestrians, who are naturally found on ground surfaces; paintings are located on horizontal wall surfaces and mugs are located on horizontal support surfaces. The recording of the eye movements during the counting search task served as a method of validating the proposed contextual guidance model as well as a point of comparison between the model and a purely saliency-based model.

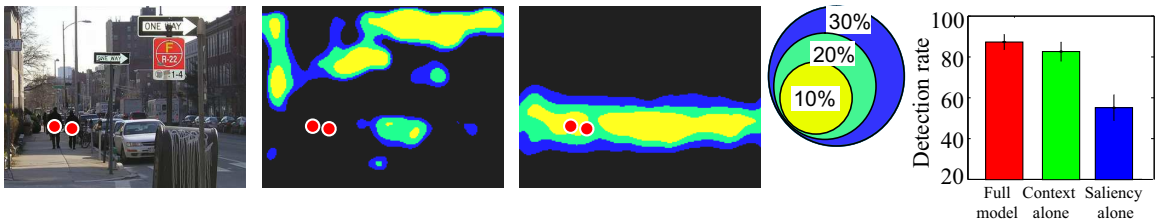

Figure 8. Comparison of performance on a detection task between a saliency model and the contextual guidance model. From left to right: 1) input image, 2) image regions selected by a saliency map, and 3) by the contextual guidance model. The red dots indicate the location of two search targets (people). The output of the two models (saliency and context) are thresholded and encoded using a color code. The graph with circles indicates how the coding of the different image areas: the yellow (lighter) region corresponds to the 10% of image pixels with higher saliency. The plot on the right shows the detection rate for pedestrians. The detection rate corresponds to the number of targets within a region of size 20% of the size of the image. Each bar corresponds (from right to left) to the detection rates of a system using saliency alone, using context priors alone, and a system using contextual guidance of saliency (integrating both context priors and bottom-up saliency). This result illustrates the power of a vision system that does not incorporate a model of the target. Informative regions are selected before processing the target.

Participants

A total of 24 Michigan State University undergraduates participated in the experiment (eight participants per search task) and received either credit toward an introductory psychology course or $7 as compensation. All participants had normal vision.

Apparatus

Eyetracking was performed by a Generation 5.5 SRI Dual Purkinje Image Eyetracker, sampling at 1000Hz. The eyetracker recorded the position and duration of eye movements during the search and the input of the participant's response. Full-color photographs were displayed on a NEC Multisync P750 monitor (refresh rate = 143 Hz).

Stimuli

The images used in the eye movements experiments consisted of two sets of 36 digitized full-color photographs taken from various urban locations (for the people search task) and various indoor scenes (for the mug and painting search tasks). For the people search task, the 36 images included 14 scenes without people and 22 scenes containing 1-6 people. A representative sample of the types of scenes used is shown in Figure 13 (people could be found on roads, pavements, grass, stairs, sidewalks, benches, bridges, etc). The same set of 36 images of indoors was used for the mug and painting tasks, as both objects are consistent in a variety of indoors categories (cf. Figure 14). Paintings were found hanging on walls and mugs were located on horizontal support-type surfaces, like kitchen islands and counters, desks, and dining, coffee, and end tables). There were respectively 17 images without paintings and 19 containing 1-6 paintings; 18 images without mugs and 18 images containing between 1-6 mugs. Mean target sizes and standard deviation (in brackets) were 1.05% (1.24 %) of the image size for people, 7.3% (7.63%) for painting and 0.5% (0.4%) for mugs. The set of images used for the eyetracking experiments was independent of the set used for adjusting the parameters and training the model. Note that we trained one model per task, independently of each other. All images subtended 15.8 deg. x 11.9 deg. of visual angle.

Procedure

Three groups of eight observers each participated in the people, painting, and mug search tasks. They were seated at a viewing distance of 1.13 m from the monitor. The right eye was tracked, but viewing was binocular. After the participant centered their fixation, a scene appeared and observers counted the number of people present (group 1), counted the number of paintings present (group 2), or counted the number of mugs (group 3). A scene was displayed until the participant responded or for a maximum of 10s. Once the participants pressed the response button the search was terminated and the scene was replaced with a number array. The number array consisted of 8 digits (0-7) presented in two rows. Participants made their response by fixating on the selected digit and pressing a response button. Responses were scored as the digit closest to the last fixation on the screen at the time the button was pressed. The eyetracker was used to record the position and duration of eye movements during the search task, and response to the number array. The experimenter initiated each trial when calibration was deemed satisfactory, which was determined as $+/-$ 4 pixels from each calibration point. Saccades were defined by a combination of velocity and distance criteria (Henderson, McClure, Pierce & Schrock, 1997). Eye movements smaller than the predetermined criteria were considered drift within a fixation. Individual fixation durations were computed as elapsed time between saccades. The position of each fixation was computed from the average position of each data point within the fixation and weighted by the duration of each of those data points. The experiment lasted about 40 minutes.

Results: Eye movements evaluation

The task of counting target objects within pictures is similar to an exhaustive visual search task (Sternberg, 1966). In our design, each scene could contain up to 6 targets, target size was not pre-specified and varied among the stimuli set.

Under these circumstances, we expected participants to exhaustively search each scene, regardless of the true number of targets present. As expected, reaction times and fixation counts did not differ between target present and target absent conditions (cf. Table 1 and detailed analysis below).

On average, participants ended the search before the 10 s time limit for 97% of the people trials, 85% of the paintings trials, and 66% of the mug trials. Accordingly, participants' responses times were higher in the mug search than the other two conditions (cf. Table 1), a result which is not surprising in light of the very small size of mug targets in the scenes and the diversity of their locations.

As the task consisted in counting the target objects and not merely indicating their presence, we did not expect participants to terminate the search earlier on target present than target absent trials. Indeed, responses times did not differ between target present and absent trials (cf. Table 1).

The number of fixations summarized in Table 1 is consistent with mean reaction times: participants made an average of 22 fixations in the mug condition, 15 in the painting condition, and 13 in the people conditions. Rayner (1998) reported that fixation durations averaged about 275 ms for visual search tasks and 330 ms during scene perception. Fixation durations in the counting task were slightly shorter, overall averaging 236 msec (240 msec for present trials and 232 msec for absent trials, not significantly different, $F < 1$). These values are similar to the mean fixation duration of 247 ms observed in an object search task in line drawings of real-world scenes using the same eyetracker and analysis criteria (Henderson et al., 1999).

The average saccade amplitude (measured in degrees of visual angle) was negatively correlated with fixation count across tasks: more fixations were accompanied by shorter saccade amplitudes in the mug search task than in the other tasks. Visual search tasks have been found to exhibit more long-amplitude saccades than free viewing of natural images (on average, about 3 degrees for search and 4 degrees for scene perception, Rayner, 1998; Table 1, see also Tatler, Baddeley and Vincent, 2006). The counting search tasks resulted in an averaged saccade length of 3 degrees.

An ANOVA comparing the effects of the three tasks and target status (present-absent) on saccade length showed that there was a main effect of search condition ($F(2) = 1588$, $p < 0.001$), no effect of target presence ($F < 1$), and a significant interaction between the search task condition and target presence ($F(2) = 30.2$, $p < 0.001$). In the people condition, saccade amplitude was larger in the target absent than target present condition (3.08 deg vs 2.57 deg, $t(7) = 4.49$, $p < 0.01$) but the reverse was true for the mug condition (2.79 deg. vs. 2.28 deg, $t(7) = 6.6$, $p < 0.01$). No effect of saccade amplitude was found in the painting search.

Results: Consistency across participants

In this section, we evaluate how consistent the fixation positions, that will later be compared with the models, were across participants. Analysis of the eye movement patterns across participants showed that the fixations were strongly constrained by the search task and the scene context.

To evaluate quantitatively the consistency across participants, we studied how well the fixations of 7 participants can be used to predict the locations fixated by the eighth participant. To illustrate, Fig. 9.A shows the fixations of 7 participants superimposed on a scene for the people search task. From each subset of 7 participants, we created a mixture of Gaussians by putting a Gaussian of 1 degree of visual angle centered on each fixation. This mixture defines the distribution:

$$p(x_i^t = x) = \frac{1}{M-1} \sum_{j \setminus i} \frac{1}{N_j} \sum_{t=1}^{N_j} \mathcal{N}(x; x_j^t, \sigma) \qquad (8)$$

where x_j^t denotes the location of the fixation number t for participant j. The notation $j \setminus i$ denotes the sum over all the participants excluding participant i. M is the number of participants and N_j is the number of fixations of participant j. The obtained distribution $p(x_i^t = x)$ is an approximation for the distribution over fixated locations. Note that the ordering of the fixations is not important for the analysis here (therefore, this distribution ignores the temporal ordering of the fixations).

To evaluate consistency across participants in a way that is consistent with the evaluation of model performance (see next section), the density $p(x_i^t = x)$ is thresholded to select an image region with the highest probability of being fixated that has an area of 20% of the image size (Fig. 9.B). The consistency across participants is determined by the percentage of fixations of the i-th participant that fell within the selected image region (chance is at 20%). The final result is obtained by averaging the consistency obtained for all participants and images. The results are summarized in Fig. 9. First, the results show that participants are very consistent with one another in the fixated locations in the target present conditions (Fig. 9.D-F). Considering the five first fixations, participants have a very high level of consistency both in the target absent and target present case for the people search (over 90 % in both cases). In the two other search conditions, the consistency across participants is significantly higher when the target is present than absent (painting, $t(34) = 2.9$, $p < .01$; mug, $t(34) = 3, p < .01$).

For the target present images, we can also evaluate how well the location of the target can predict image regions that will be fixated. We define the target selected region using the target mask (for all the images the targets were previously segmented) and blurring the binary mask with a gaussian of 1 degree of width at half amplitude (Fig. 9.C). As before, we threshold the blurred mask in order to select an image region with an area equal to 20% of the image size. Then, we counted the number of fixations that fell within the target region. The results are shown in Figs. 9.D-F. Surprisingly, the region defined by the target only marginally predicted participants' fixations (on average, 76% for people, 48% for painting and 63% for mug conditions, all significantly lower than the consistency across participants).

It is interesting to note that using other participants to predict the image locations fixated by an additional participant

Table 1
Summary of the eye movement patterns for the three search tasks.

		People		Paintings		Mug	
		Absent	Present	Absent	Present	Absent	Present
RT (ms)	Avg	4546	4360	3974	3817	6444	6775
	SD	605	957	1097	778	2176	1966
Fix. Duration (ms)	Avg	229	237	228	236	239	247
	SD	37	41	26	22	22	18
Fix. Count	Avg	13.9	13	15.7	14.9	21.8	21.8
	SD	2	3	4.3	4.7	6.9	6.5
Sac. Length (deg)	Avg	3.1	2.6	3.3	3.2	2.3	2.8
	SD	0.4	0.4	0.4	0.4	0.3	0.4

provides more accurate predictions than the target location itself (for the three objects studied). This suggests that the locations fixated by observers in target present images are not only driven by the target location or the target features but also by other image components. The next section compares the predictions generated by the models based on two image components: saliency and global context features.

Results: Comparison human observers and models

To assess the respective role of saliency and scene context in guiding eye movements, we compared a model using bottom-up saliency alone (Fig. 10) and the Contextual Guidance model (Fig. 11) that integrates saliency and scene information (eq. 7) with the fixations of participants for the three search tasks. The output of both models is a map in which each location is assigned a value that indicates how relevant that location is with respect to the task.

As in the previous results section, we apply a threshold to the outputs of the models in order to define predicted regions with a predefined size that allows for comparing the different algorithms. The threshold is set so that the selected image region occupies a fixed proportion of the image size (set to 20% for the results shown in Fig. 12). The efficiency of each model is determined by the percentage of human fixations that fall within the predicted region. Fig. 12 summarizes the results obtained in the search experiment for the three target objects, and compares two instances of the model (Fig. 1): a model using saliency alone (local pathway), and the Contextual Guidance model (full model) integrating both sources of information (eq. 7). We also plotted the consistency across participants from Fig. 9 on the same graph as it provides an upper bound on the performance that can be obtained.

First of all, the two models performed well above chance level (20%) for target present and absent conditions, in their predictions of locations of human fixations. The differences seen on Fig. 12 are statistically significant: for the target present case, an ANOVA considering the first five fixations for the three groups and the two models showed an effect of models ($F(1,55) = 28.7$, $p < .0001$) with the full model better predicting human fixations than the saliency only model (respectively 73% and 58%). A significant main effect of groups ($F(2,55) = 12.4$, $p < .0001$) was mostly driven by differences in saliency model performance. The same trend was found for the target absent conditions.

As our models are expected to be more representative of the early stages of the search, before decision factors start playing a dominant role in the scan pattern, we considered first the first two fixations for the statistical analysis. For the people search task, the graphs in Fig. 12 clearly show that the full model performed better than the saliency model ($t(20) = 4.3$, $p < .001$ for target present, and $t(14) = 3.6$, $p < .01$, for target absent). The full model's advantage remains for the painting search task ($t(18) = 4.2$, $p < .001$ for target present, and $t(16) = 2.7$, $p < .02$, for target absent) and the mug search task (for target present, $t(17) = 2.8$, $p < .02$, and for target absent, $t(17) = 2.2$, $p < .05$).

When considering the first five fixations for the analysis, for the people search task, the graphs in Fig. 12 clearly indicate that the full model performed better than the saliency only model for both target present($t(20) = 3.6$, $p < .01$) and target absent conditions ($t(14) = 6.3$, $p < .01$). This remains true for the painting and the mug search tasks (respectively, $t(18) = 2.7$, $p < .02$ and $t(17) = 3.5$, $p < .01$) but for target present only.

Interpretation

The comparison of the contextual guidance model and the saliency-based model with participants' consistency provides a very rich set of results. The contextual model was able to consistently predict the locations of the first few fixations in the three tasks, despite the fact that some target objects were very small (e.g., people and mugs were representing only 1% of the image pixels) and that objects location varied greatly, even when the target object was absent. Participants had a tendency to start fixating image locations that contained salient local features within the region selected by global contextual features. This effect was strongest in the people task search (Fig. 11.A, and Fig. 13), showing that par-

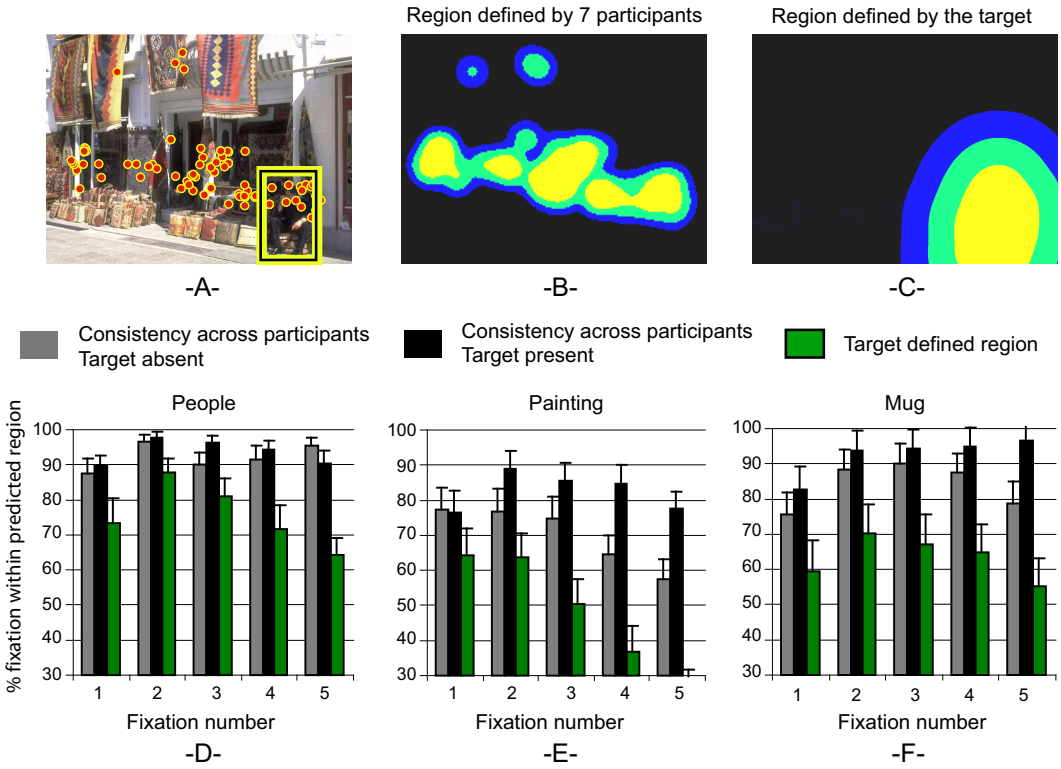

Figure 9. Analysis of the regularity of fixations. The first row illustrates how the consistency among participants was computed and also how well the location of the target predicted the image regions fixated. A) Example of an image and all the fixations of seven participants for the people search task. B) To analyze consistency among participants we iteratively defined a region using seven participants to predict the fixations of the eighth participant. C) For images with target present we defined a region using the support of the target. For the three search tasks we evaluated the consistency among participants and also how well the region occupied by the target explained the locations fixated by the participants: D) people, E) painting and F) mug search task. In all the cases the consistency among participants was high from the first fixation. In all the cases, the consistency among participants was higher than the predictions made by the target location.

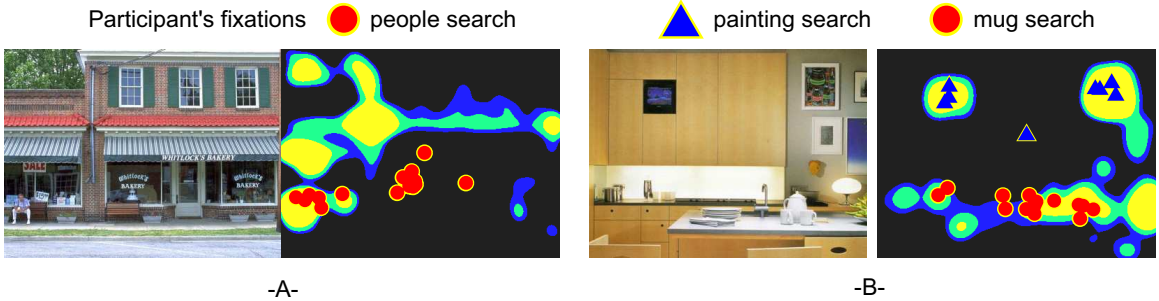

Figure 10. A) People and B) mug and painting search tasks. For the two example images, we show the regions predicted by the saliency model and superimposed the first 2 locations fixated by 8 participants. A model based only on image saliency does not provides accurate predictions for the fixated locations and it is not able to explain changes in search when the target object changes.

ticipants kept exploring the regions predicted by the Contextual Guidance Model. Pedestrians were relatively small target, embedded in large scenes with clutter, forcing observers to scrutinize multiple ground surface locations.

In the painting and mugs conditions, participants also start by exploring the image regions that are salient and the most contextually relevant locations, but then continue exploring the entirety of the scene after the second or third fixation, resulting in lower performance of the contextual model as search progresses. Small objects like mugs can in practice be placed almost anywhere in a room, so it is possible that participants continued exploring regions of the scene that, despite not being strongly associated with typical positions of mugs, are not unlikely to contain the target (e.g., on a chair, a stack of books). Participants were very consistent with each other for the first five fixations (cf. Fig. 9), suggesting that they were looking indeed at the same regions. The mug condition showed another interesting pattern (see Fig. 12.E): the

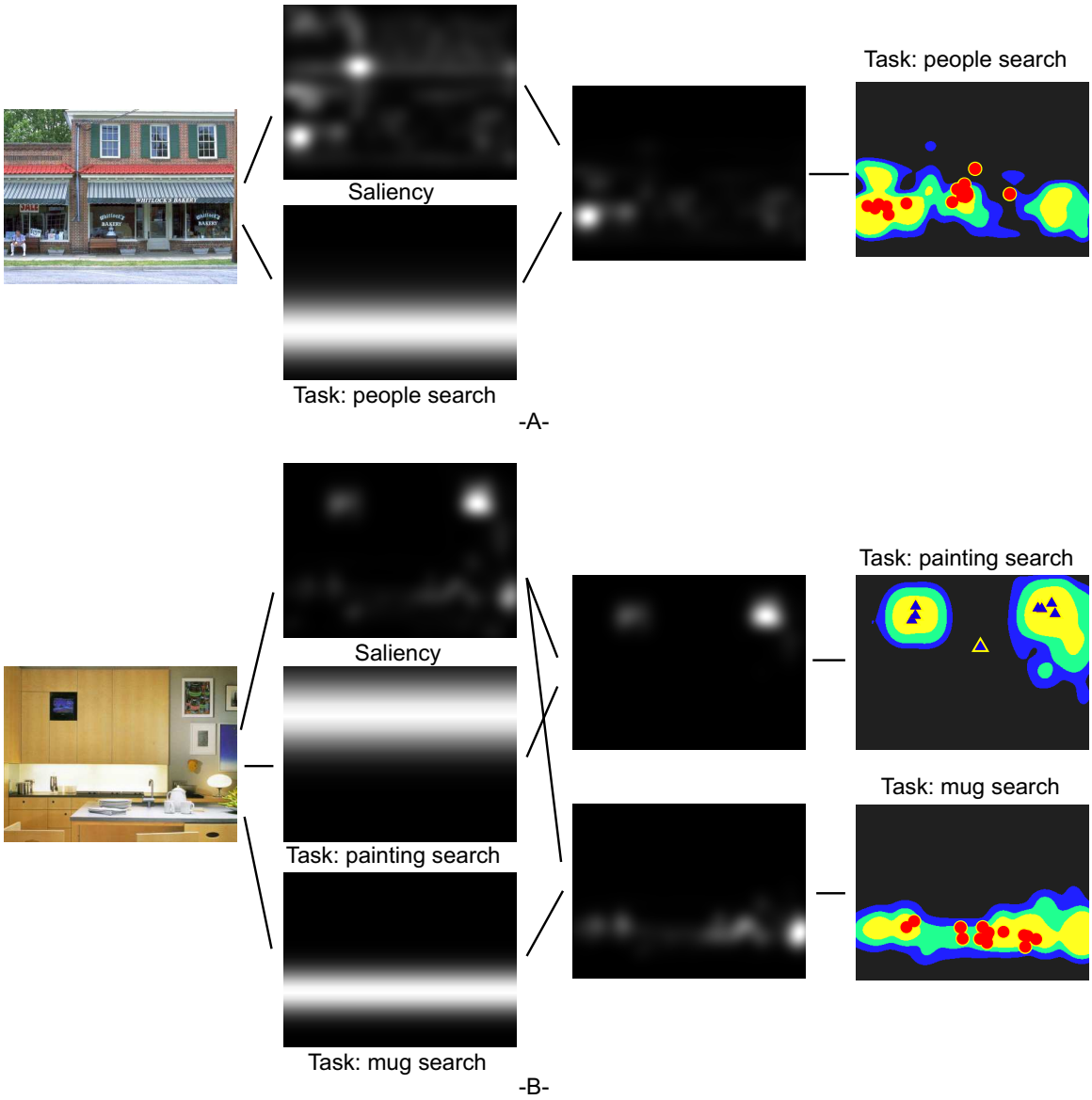

Figure 11. The full model presented here incorporated scene priors to modulate the salient regions taking into account the expected location of the target given its scene context. A) In the people search task, the two factors are combined resulting in a saliency map modulated by the task. For evaluating the performance of the models, we compared the locations fixated by 8 participants with a thresholded map. B) Here, it is illustrated how the task modulates the salient regions. The same image was used on two tasks: Painting and mug search. In this example, the result show that the scene context is able to predict which regions will be fixated and how the task produces a change of the fixations.

saliency model performed almost as well as the full model in both target present and absent conditions. This suggests that the saliency model performance in this task was not due to the saliency of the mugs themselves but instead was driven by other salient objects spatially associated with the mugs (cf. table, chair, see Fig. 2). Figures 13 and 14 qualitatively illustrate the performance of the models: both figures show a subset of the images used in the experiment and the regions selected by a model based on saliency alone and the full model, integrating contextual information.

Interestingly, the best predictor of any participants' fixations in the search counting tasks was the locations fixated by other participants, and not the location of the target object per se (Fig. 9). This effect was found for the three search tasks and suggests that the task and the scene context impose stronger constraints on fixation locations than the actual position of the target. It is possible that the requirement of the counting task had amplified the consistency between fixations, focusing overt attention on all the regions that were potentially associated with the target. Despite its outstanding performance over a saliency model, the global context model does not perform as well as the participants themselves, suggesting room for improvement in modeling additional sources of contextual (e.g., object-to-object local

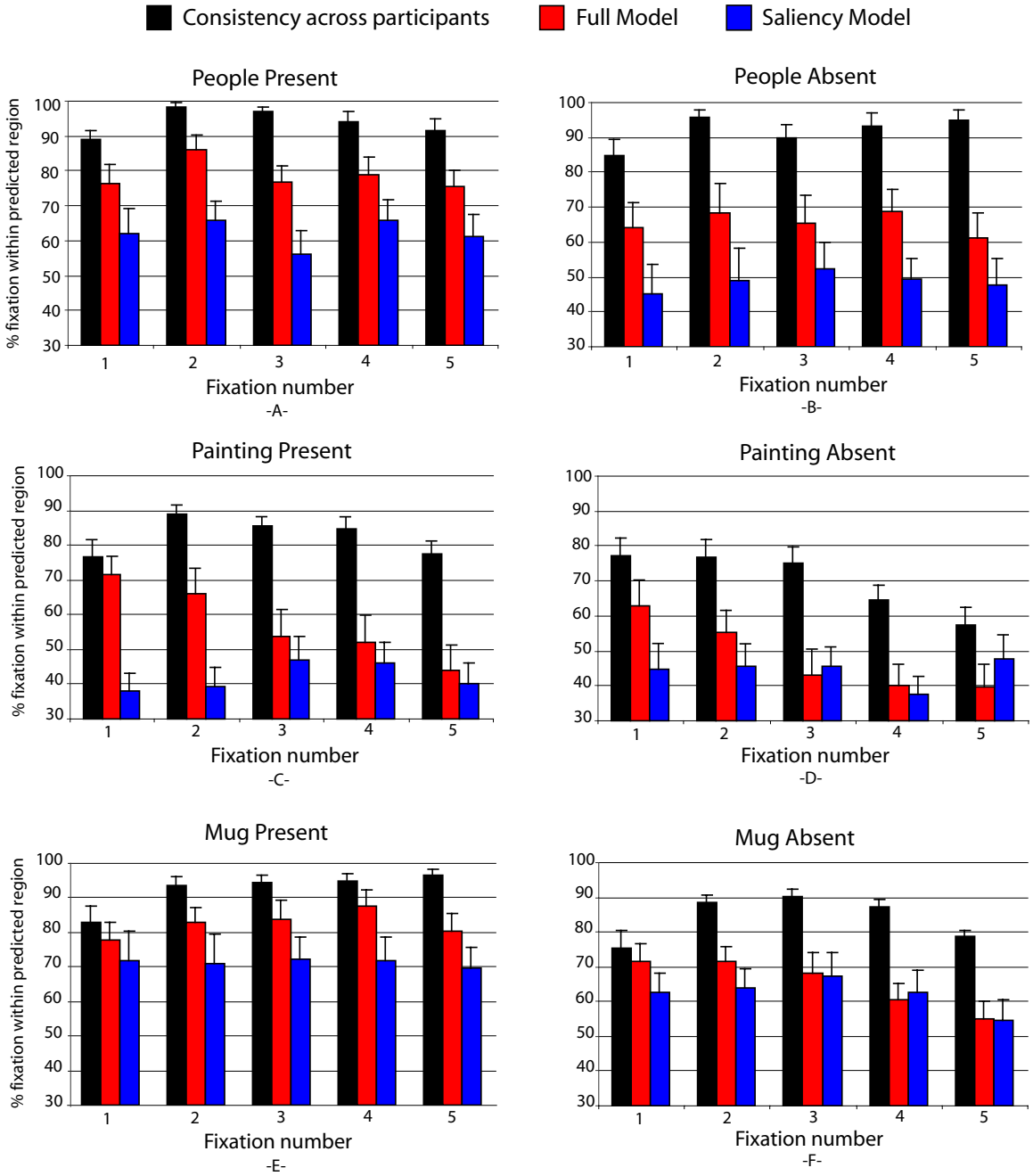

Figure 12. Comparison of participant's fixations and the models. The vertical axis is the performance of each model measured by counting the number of fixations that fall within the 20% of the image with the highest score given by each model. The horizontal axis corresponds to the fixation number (with the central fixation removed). A) Performance of the saliency model. B) Performance of the context model. Figures C) and D) compare the performance of the saliency model and the model that integrates contextual information and saliency. In addition, the consistency between observers is shown. Participants can better predict the fixations of other participants than any of the models.

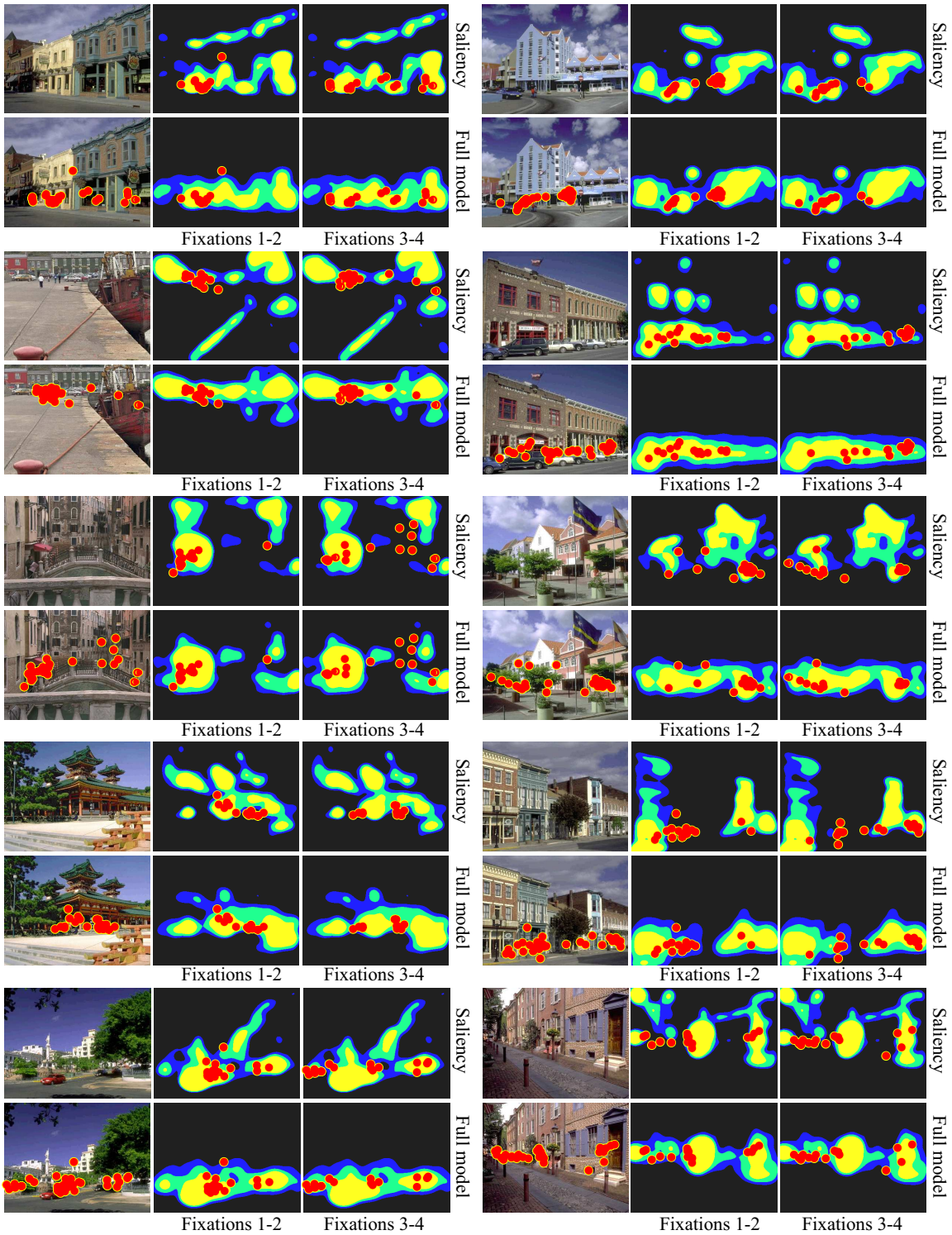

Figure 13. Comparison between regions selected by a model using saliency alone and by the full model for the people search task. Each panel shows on the top left the input image, and on the bottom left the image with the first 4 fixations for all 8 participants superimposed. The top row shows the regions predicted by saliency alone (the images show fixations 1-2 and 3-4 for the 8 participants). The bottom row shows the regions predicted by the full model that integrates context and saliency.

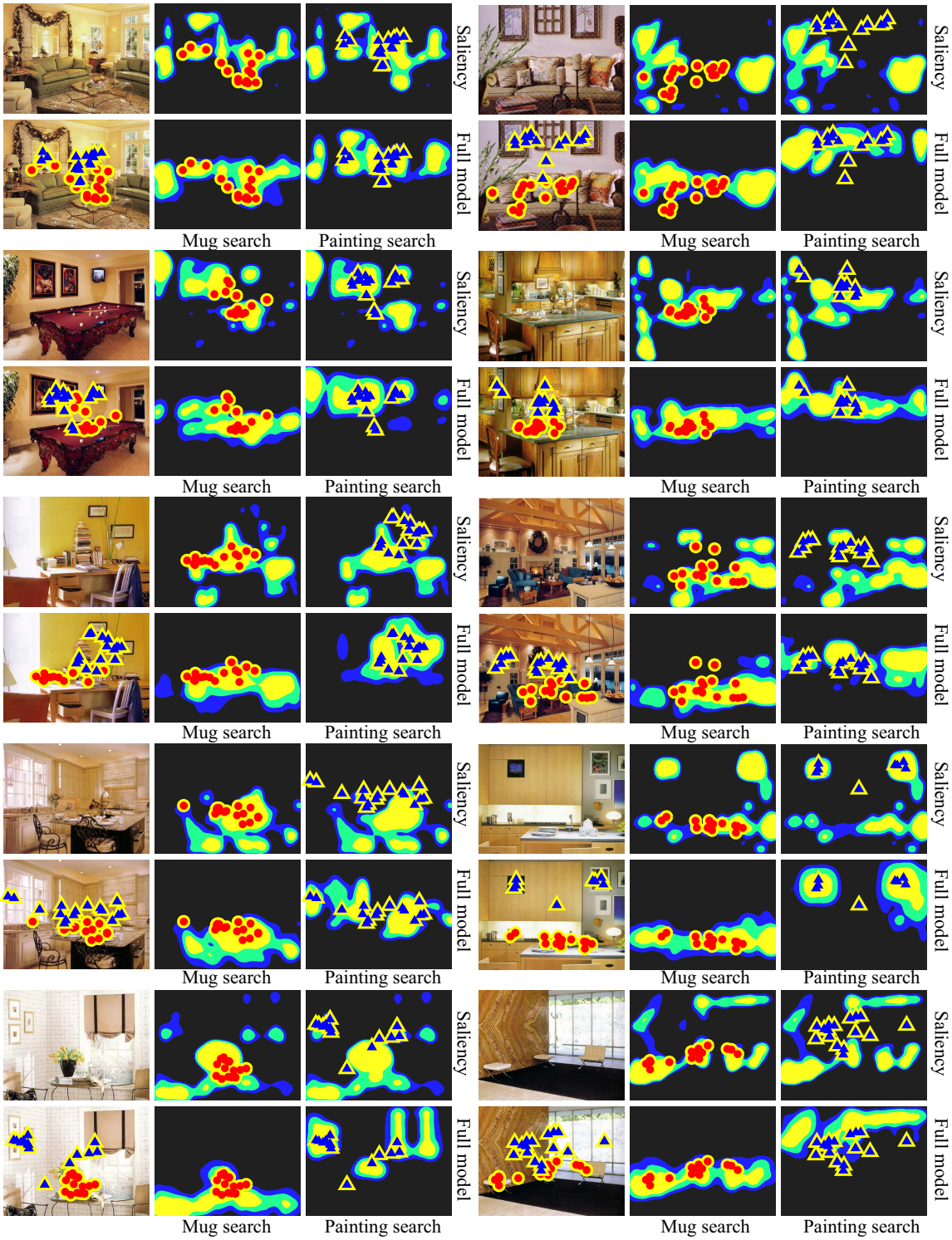

Figure 14. Comparison between regions selected by a model using saliency alone and by the full model for the mug and painting search tasks. The images show fixations 1-2 for the 8 participants on the mug search task (center) and painting search task (left). The top row shows the regions predicted by the saliency alone and, therefore, the predicted regions do not change with the task. The bottom row shows the regions predicted by the full model that integrates context and saliency. The full model selects regions that are relevant for the task and are a better predictor of eye fixations than saliency alone.

associations).

General discussion

This paper proposes a computational instantiation of a Bayesian model of attention, demonstrating the mandatory role of scene context for search tasks in real-world images. Attentional mechanisms driven by image saliency and contextual guidance emerge as a natural consequence of the probabilistic framework providing an integrated and formal scheme into which local and global features can be combined automatically to guide subsequent object detection and recognition.

Our approach suggests that a robust holistic representation of scene context can be computed from the same ensemble of low-level features used to construct other low-level image representations (e.g., junctions, surfaces), and can be integrated with saliency computation early enough to guide the deployment of attention and first eye movements toward likely locations of target objects. From an algorithmic point of view, early contextual control of the focus of attention is important as it avoids expending computational resources in analyzing spatial locations with low probability of containing the target based on prior experience. In the Contextual Guidance model, task-related information modulates the selection of the image regions that are relevant. We demonstrated the effectiveness of the Contextual Guidance model for predicting the locations of the first few fixations in three different search tasks, performed on various types of scenes categories (urban environments, variety of rooms), and for various object's size conditions.

Behavioral research has shown that contextual information plays an important role in object detection (Biederman et al., 1982; Boyce & Pollatsek, 1992; Oliva et al., 2003; Palmer, 1975). Changes in real world scenes are noticed more quickly for objects and regions of interest (Rensink et al., 1997), and scene context can even influence the detection of a change (Hollingworth & Henderson, 2000) suggesting a preferential deployment of attention to these parts of a scene. Experimental results suggest that the selection of these regions is governed not merely by low-level saliency, but also by scene semantics (Henderson & Hollingworth, 1999). Visual search is facilitated when there is a correlation across different trials between the contextual configuration of the scene display and the target location (Brockmole & Henderson, 2005; Chun & Jiang, 1998 1999; Hidalgo-Sotelo et al., 2005; Jiang & Wagner, 2004; Oliva et al., 2004; Olson & Chun, 2001). In a similar vein, several studies support the idea that scene semantics can be available early in the chain of information processing (Potter, 1976) and suggest that scene recognition may not require object recognition as a first step (Fei-Fei & Perona, 2005; McCotter et al., 2005; Oliva & Torralba, 2001; Schyns & Oliva, 1994). The present approach proposes a feedforward processing of context (Fig. 1) that is independent of object-related processing mechanisms. The global scene representation delivers contextual information in parallel with the processing of local features, providing a formal realization of an efficient feed-forward mechanism for the guidance of attention. An early impact of scene context is also compatible with the Reverse Hierarchy Theory (Hochstein & Ahissar 2002) in which properties that are abstracted late in visual processing (like object shapes, categorical scene description) rapidly feed back into early stages and constrain local processing.

It is important to note that our scene-centered approach of context modeling is complementary and not opposed to an object-centered approach of Context. The advantage of using a scene-centered approach is that contextual influences occur independent of the level of visual complexity of the image (a drawback of a contextual definition based on identification of one or more objects), and is robust at many levels of the ease of target detectability (e.g., when the target is very small or camouflaged). The global-to-local scheme of visual processing could conceivably be applied to the mechanism of object contextual influences (DeGraef, 1992; Henderson et al., 1987; Palmer, 1975) advocating for a two-stage temporal development of contextual effects: global scene features would account for an initial impact of context, quickly constraining some local analysis, while object-to-object association would be build in a more progressive way, depending on which objects were initially segmented. A more local-based approach to context is consistent with recent developments in contextual cuing tasks, showing that local associations and spatially grouped clusters of objects can also facilitate localization of the target (Jiang & Wagner, 2004; Olson & Chun, 2002), though global influences seem to have more effect in contextual cueing of real-world scenes (Brockmole et al., in press). Both levels of contextual analysis could theoretically occur within a single fixation, and their relative contribution for determining search performance is a challenging question to further models and theories of visual context.

The inclusion of object-driven representations and their interaction with attentional mechanisms is beyond the scope of this paper. Simplified experimental setups (Wolfe, 1994) and natural but simplified worlds (Rao et al., 2002) have begun to show how a model of the target object influences the allocation of attention. In large part, however, identifying the relevant features of object categories in real-world scenes remains an open issue (Riesenhuber & Poggio, 1999; Torralba, Murphy & 2004b; Ullman et al., 2002). Our claim in this paper is that when the target is very small (the people and the mugs occupy a region that has a size of 1% the size of the image on average), the target's appearance will play a secondary role in guiding the eye movements, for at least the initial few fixations. This assumption is supported by the finding that the location of the target itself did not predict well the locations of search fixations (cf. Fig. 9). If target appearance drove fixations, then fixations would be expected to be attracted to the targets when they were present rather than to fall on contextually expected locations. The current study emphasizes how much of eye movement location can be explained when a target model is not implemented.

Our study provides the lower bound of the expected performance that can be achieved by a computational model of context, when the target is small, embedded in high level clutter, or even not present at all. In Murphy, Torralba &

Freeman (2003), global and local features (including a model of the target) are used to detect objects in scenes. The inclusion of global features helps to improve the performance of the final detection. However, use of these models to predict fixations, would require that false alarms of such models were similar to the errors made by participants. This is still beyond the current state of computer vision for general object recognition. In Torralba et al. (2003b, 2004) local objects are used to focus computations into image regions likely to contain a target object. This strategy is only very efficient when trying to detect targets that are strongly linked to other objects. The system learns to first detect objects defined by simple features (e.g., a computer screen) that provide strong contextual information in order to facilitate localization of small targets (e.g., a computer mouse). Objects that are not within expected locations defined by context may still be detected but would require strong local evidence to produce confident detections.

In this paper we demonstrate the robustness of global contextual information in predicting observer's eye movements in a search counting task of cluttered, real-world scenes. The feed-forward scheme that computes these global features successfully provides the relevant contextual information to direct attention very early in the visual processing stream.

References

Ariely, D., (2001), Seeing sets: Representation by statistical properties, *Psychological Science*, 12 (2), 157- 162

Bacon-Mace, N., Mace, M.J.M, Fabre-Thorpe, M., & Thorpe, S. (2005). The time course of visual processing: backward masking and natural scene categorization. *Vision Research*, 45, 1459-1469.

Bar, M. (2004). Visual objects in context. *Nature Neuroscience Reviews*, 5, 617-629.

Bar, M. & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, 38, 347-358.

Biederman, I., Mezzanotte, R.J., & Rabinowitz, J.C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology* 14:143-177.

Biederman, I. (1995). Visual object recognition. In An Invitation to Cognitive Science: Visual Cognition (2nd edition). M.Kosslyn & D.N. Osherson (eds.), vol 2, 121-165.

Boyce, S. J., & Pollatsek, A. (1992). Identification of objects in scenes: The role of scene background in object naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 531-543.

Brockmole, J. R., & Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, 13, 99-108.

Brockmole, J. R., & Henderson, J. M. (in press). Recognition and attention guidance during contextual cueing in real-world scenes: Evidence from eye movements. *Quarterly Journal of Experimental Psychology*.

Brockmole, J. R., Castelhano, M. S., & Henderson, J. M. (in press). Contextual cueing in naturalistic scenes: Global and local contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Carson, C., Belongie, S., Greenspan, H., & Malik, J. (2002). Blobworld: image segmentation using expectation-maximization and its expectation to image querying. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, 24, 1026-1038.

Chong S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision research*. Volume: 43, Issue: 4 pp. 393-404 February.

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

Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. Psychological Science, 10, 360-365.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, 4, 170-178.

De Graef, P. (1992). Scene-context effects and models of real-world perception. In K. Rayner (Ed.), *Eye Movements and Visual Cognition: Scene perception and Reading*, Springer-Verlag (pp. 243-259).

De Graef, P., Christiaens, D., & dYdevalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, 52, 317-329.

Delorme, A., Rousselet, G.A., Mace, M.J.M, & Fabre-Thorpe, M. (2003). Interaction of top-down and bottom-up processing in the fast analysis of natural scenes. *Cognitive Brain Research*, 19:103-113.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, vol. 34, pp. 1-38.

Deubel, H., Schneider, W. X., & Bridgeman, B. (1996). Postsaccadic target blanking prevents saccadic suppression of image displacement. *Vision Research*, 36, 985-996.

Epstein, R.A. (2005).The cortical basis of visual scene processing.*Visual Cognition*, 12: 954-978.

Epstein, R., & Kanwisher, N. (1998). A Cortical Representation of the Local Visual Environment. *Nature*, 392, 598-601.

Evans, K.K., & Treisman, A. (2005). Perception of objects in natural scenes: is it really attention free? *Journal of Experimental Psychology: Human Perception and Performance*, 31, 1476-1492.

Fei-Fei, L., & Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. *IEEE Computer Vision and Pattern Recognition*, vol. 2, pp. 524-531.

Findlay, J.M. (2004). Eye scanning and visual search. In Henderson J. M. and Ferreira F. (Eds.) *The interface of language, vision and action: Eye movements and the visual world*. New York, Psychology Press (pp 135-159).

Greene, M.R., & Oliva. A. (submitted). Natural Scene Categorization from Conjunctions of Ecological Global Properties.

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Rossion, B., & Schyns. P.G. (2005). Diagnostic colors contribute to early stages of scene categorization: behavioral and neurophysiological evidences. *Visual Cognition*, 12, 878-892.

Goh, J.O.S., Siong, S.C., Park, D., Gutchess, A., Hebrank, A., & Chee, M.W.L. (2004). Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. *The Journal of Neuroscience*, 24, 10223-10228.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *Trends in Cognitive Sciences*, 7, 498-504.

Henderson, J.M., & Hollingworth, A. (1999). High level scene perception. *Annual Review of Psychology*, 50, 243-271.

Henderson, J.M., & Hollingworth, A. (1999b). The role of fixation position in detecting scene changes across saccades. *Psychological Science*, 10, 438-443.

Henderson, J. M., Weeks, P. A. Jr., & Hollingworth, A. (1999). Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 210-228.

Henderson, J. M., McClure, K., Pierce, S., & Schrock, G. (1997). Object identification without foveal vision: Evidence from an artificial scotoma paradigm. *Perception & Psychophysics*, 59, 323-346.

Henderson, J. M., Pollatsek, A., & Rayner, K. (1987). The effects of foveal priming and extrafoveal preview on object identification. Journal of Experimental Psychology: Human Perception and Performance, 13, 449-463.

Hidalgo-Sotelo, B.,Oliva, A., & Torralba, A. (2005). Human Learning of Contextual Priors for Object Search: Where does the time go? Proceedings of the *3rd Workshop on Attention and Performance in Computer Vision*.

Hochstein, S., & Ahissar, M. (2002). View from the top: hierarchies and reverse hierarchies in the visual system. *Neuron*, 36, 791-804.

Hollingworth, A. & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition*, Special Issue on Change Blindness and Visual Memory, 7, 213-235.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. Journal of Experimental Psychology: Human Perception and Performance, 28, 113-136

Hollingworth, A., Schrock, G., & Henderson, J. M. (2001). Change detection in the flicker paradigm: the role of fixation position within the scene. *Memory & Cognition*, 29, 296-304.

Hoffman, J.E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57: 787-795.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Vision*, 20(11):12-54.

Jiang, Y., & Wagner, L.C. (2004). What is learned in spatial contextual cuing configuration or individual locations?. *Perception & Psychophysics*, 66, 454-463.

Johnson, J. S., & Olshausen, B. A. (2003). Time course of neural signatures of object recognition. *Journal of Vision*, 3(7), 499-512.

Kanwisher, N. (2003) The Ventral Visual Object Pathway in Humans: Evidence from fMRI. In The Visual Neurosciences. Edited by Chalupa, L. & Werner, J. MIT Press 1179-1189.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, 13(2): 150-158.

Kirchner, H., & Thorpe, S.J.(2006). Ultra-rapid object detection with saccadic eye movements: visual processing revisited. *Vision Research*, 46: 1762-1776.

Knill, D. & Richards, W. Perception as Bayesian Inference (Cambridge Univ. Press, Cambridge, 1996).

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

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research*, 35, 1897-1916.

Kunar, M.A., Flusberg, S.J., & Wolfe, J.M.(2006). Contextual cueing by global features. *Perception & Psychophysics*.

Land, M.F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision research*, 41, 3559-3565.

Li, F. F., VanRullen, R., Koch, C. & Perona, P. (2002) Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academia of Sciences*, A, 99: 9596-9601.

Liversedge, S.P., & Findlay, J.M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4, 6-14.

Loftus, G.R., & Mackworth, N.H., (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 565-572.

Marois, R., Yi, D.J., & Chun, M. (2004). The neural fate of consciously perceived and missed events in the attentional blink. *Neuron*, 41: 465-472.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P.G. (2005). The use of visual information in natural scenes, *Visual Cognition*, 12: 938-953.

Murphy, K. P., Torralba, A., & Freeman, W. T. (2003). Using the forest to see the trees: a graphical model relating features, objects and scenes. Adv. in Neural Information Processing Systems 16 (NIPS), Vancouver, BC, MIT Press.

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cognitive Psychology*, 9: 353-383.

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

Noton, D., & Stark, L. (1971). Scanpaths in Eye Movements during Pattern Perception. *Science*, Vol. 171, 3968, 308-311.

Oliva, A. (2005). Gist of the Scene. In Neurobiology of Attention, L. Itti, G. Rees & J. K. Tsotsos (Eds.), Elsevier, San Diego, CA (pp 251-256).

Oliva, A., & Schyns, P.G. (1997) Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 34: 72-107.

Oliva, A., & Schyns, P.G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41, 176-210.

Oliva, A., & Torralba, A. (2002). Scene-Centered Description from Spatial Envelope Properties. Lecture Note in Computer Science Serie. *Proc. 2nd Workshop on Biologically Motivated Computer Vision*, Springer-Verlag (pp 263-272).

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.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Reasearch: Special Issue on Visual Perception*.

Oliva, A., Wolfe, J. M, & Arsenio, H. (2004). Panoramic Search: The interaction of Memory and Vision in Search through a Familiar Scene. *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 30, 6, 1132-1146.

Oliva, A., Torralba, A., Castelhano, M. S. & Henderson, J. M. (2003). Top-Down control of visual attention in object detection. Proc. *IEEE Int. Conf. Image Processing*, 1, 253-256.

Olshausen, B. A. & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381, 6079

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

Palmer, S.T. (1975). The effects of contextual scenes on the identification of objects. *Memory and Cognition*, 3:519-526.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention, *Vision research* 42, 107-123.

Portilla, J., and Simoncelli, E. P. (2000) A parametric texture model based on joint statistics of complex wavelets coefficients. *International Journal of Computer Vision*, vol. 40, pp. 49-71.

Potter, M.C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, 2, 509-522.

Potter, M.C., Staub, A., & O Connor, D.H. (2004) Pictorial and Conceptual Representation of Glimpsed Pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 30: 478-489.

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 371-422.

Rayner, K., McConkie, G. W., & Ehrlich, S. F. (1978). Eye movements and integrating information across fixations. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 529544.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8, 368-373.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, 7, 17-42.

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

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39, 3157-3163.

Rousselet, G.A., Joubert, O.R., & Fabre-Thorpe, M. (2005). How long to get to the gist of real-world natural scenes? *Visual Cognition*, 12, 852-877.

Russell, B. C., Torralba, A., Murphy, K. P., Freeman, W. T. LabelMe: a database and web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025, September, 2005.

Schyns, P.G. & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5, 195-200.

Simoncelli, E.P., & Freeman, W.T. (1995) The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *2nd Annual Intl. Conf. on Image Processing*, Washington, DC.

Simons, D.J., & Levin, D.T.(1966). Change blindness. *Trends in Cognitive Science*, 1: 261-267.

Sternberg,S. (1966). High-speed scanning in human memory. *Science*, 153: 652-654.

Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2005). Learning Hierarchical Models of Scenes, Objects, and Parts. IEEE Int. Conf. on Computer Vision.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381: 520-22.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of Optical Society of America A*. Special Issue on Bayesian and Statistical Approaches to Vision, Vol.20(7), 1407-1418

Torralba, A. (2003b). Contextual priming for object detection. *International Journal of Computer Vision*. Vol. 53(2), 169-191.

Torralba, A., Murphy, K. P., & Freeman, W. T. (2004). Contextual Models for Object Detection using Boosted Random Fields. *Adv. in Neural Information Processing Systems*.

Torralba, A., Murphy, K. P. , & Freeman, W. T. (2004b). Sharing features: efficient boosting procedures for multiclass object detection. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp 762- 769).

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

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N., & Nuflo, F. (1995). Modeling visual-attention via selective tuning. *Artificial Intelligence*, 78, 507545.
Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification, *Nature Neuroscience* 5, 682687.

Vailaya, A., Jain, A., & Zhang, H.J. (1998). On image classification: City images vs. landscapes. *Pattern Recognition*, 31:1921-1935.

Vogel, E.K., Woodman, G.F., & Luck, S.J. (in press). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*.

Vogel, J., & Schiele, B. (in press). Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*.

Remington, R. (1980). Attention and saccadic eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 726-744.

Walker Renninger, L., & Malik, J. (2004) When is scene identification just texture recognition? *Vision research*, 44: 2301-2311.

Wolfe, J. M. (1994). Guided search 2.0. A revised model of visual search. *Psychonomic Bulletin and Review*, 1:202-228.

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

Wolfe, J. M., & Bennett, S.C. (1997). Preattentive object files: shapeless bundles of basic features. *Vision Research*, 37, 25-44.

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

Yarbus, A. L. (1967). Eye movements and vision. Plenum Press.