

Contextual Influences on Saliency

Antonio Torralba

ABSTRACT

This article describes a model for including scene/context priors in attention guidance. In the proposed scheme, visual context information can be available early in the visual processing chain, in order to modulate the saliency of image regions and to provide an efficient shortcut for object detection and recognition. The scene is represented by means of a low-dimensional global description obtained from low-level features. The global scene features are then used to predict the probability of presence of the target object in the scene, and its location and scale, before exploring the image.

I. INTRODUCTION

What is the role of contextual information in object recognition and detection tasks? What is the influence of the scene on determining the way that attention is deployed when trying to solve a task? How is the saliency of different image regions enhanced or reduced as a function of high-level scene information?

A number of studies have shown the importance of scene factors in object search and recognition. Studies by Biederman et al. (1982) and Palmer (1975) highlight the effect of contextual information in the processing time for object recognition. Rensink et al. (1997) have shown that changes in real-world scenes are noticed most quickly for objects or regions of interest, thus suggesting a preferential deployment of attention to these parts of a scene. Henderson and Hollingworth (1999) have reported results suggesting that the choice of these regions is governed not merely by their low-level saliency but also by scene semantics. Chun and Jiang (1998; Chapter 40) showed that visual search is facilitated when a correlation exists across different

trials between the contextual configuration of the display and the target location. In a similar vein, several studies support the idea that scene semantics can be available early in the chain of information processing (Schyns and Oliva, 1994; Thorpe et al., 1996) and suggest that scene recognition may not require object recognition as a first step (Schyns and Oliva, 1994; Oliva and Torralba, 2001; Chapter 41).

Here a scheme is described in which visual context information can be available early in the visual processing chain in order to modulate the saliency of image regions and to provide an efficient shortcut for object detection and recognition. Context consists in a global description of the scene obtained from low-level features. In the proposed scheme, contextual information is used to predict the presence and absence of the target before scanning the image and to select the image regions that are relevant for the task.

II. THE SCENE CONTEXT

In Fig. 96.1A, observers describe the scenes as (left) a pedestrian in the street, (center) a car in the street, and (right) some food on a table. However, in the three images, the blob is identical (the pedestrian blob is the same shape as the car except for a 90-degs rotation). When object intrinsic information is reduced so much that an object cannot be identified based on local information, the object recognition system is not invariant to changes in pose, orientation, and location and context plays a mayor role in recognition.

In saliency models of attention (see Chapter 94), the context of the target object is considered as a collection of distractors. Figure 96.1B (left), shows a display with a salient target, in which context (distractors) does not affect target processing (Treisman and Gelade, 1983). In the central image, a person is embedded in the back-

1

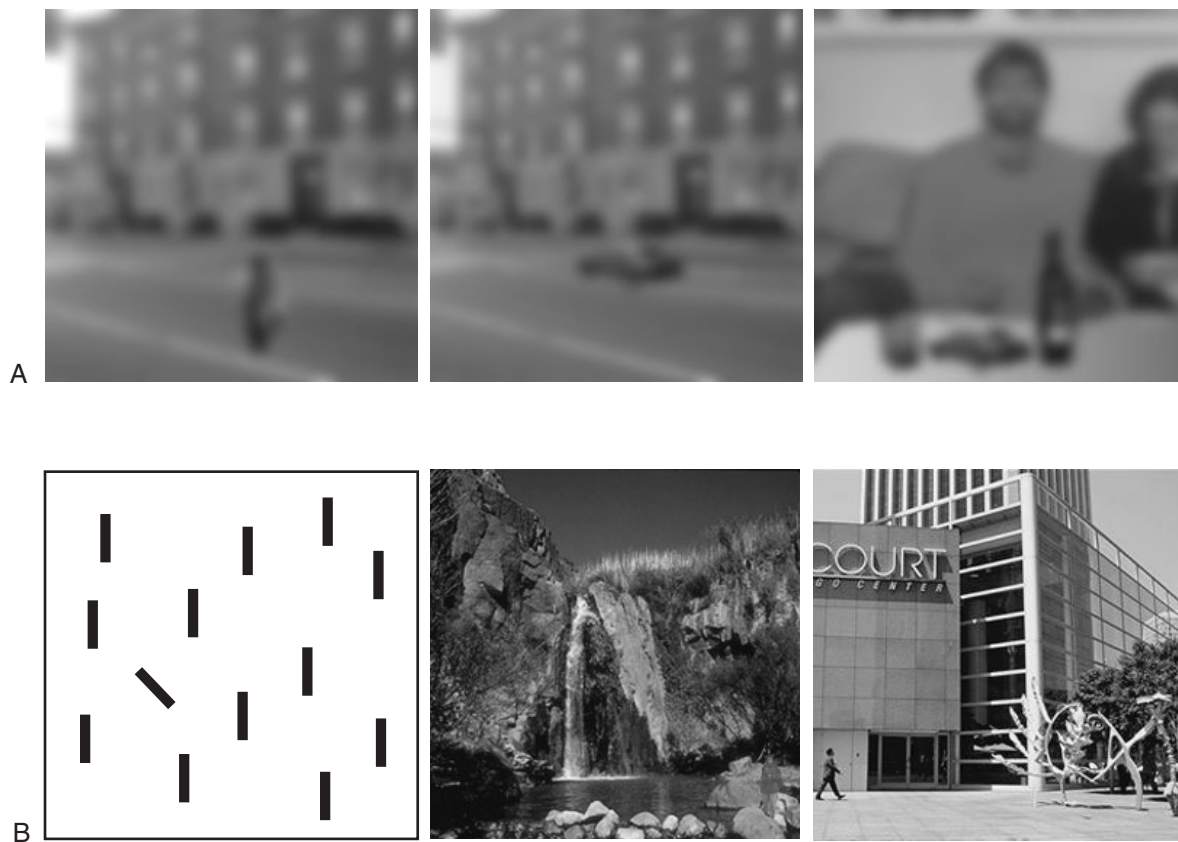


FIGURE 96.1 (A) When object intrinsic information is reduced, then context plays a major role in recognition. Now, the object recognition system is not invariant to pose, orientation, location, and background. (B) Effects of context in masking and providing priors for finding the target.

ground. The person is masked by the context and is difficult to find. In these two examples, context is noninformative and its only effect on the search is the ability of the background to mask the target. But context can also provide information about the presence of the target. In Fig. 96.2B (right) the context, instead of masking the person, provides priors about what are the expected locations and scales in which we can find the target. In the canyon scene, a person could be almost in any location. However, in the street scene, the environment imposes strong constraints on the typical locations in which people is expected to be. This use of context is the one we are interested in modeling here.

III. THE REPRESENTATION OF SCENES

We can define the context of a particular object in terms of other previously recognized objects within the scene. There, the context representation is object-centered and requires object recognition as a first step.

The context representation described here does not require parsing the image to build a representation of the scene. As suggested in (Oliva and Torralba, 2001), it is possible to build a description of the scene that bypasses object identities, in which the scene is represented as a single entity. The representation proposed is based on identifying a number of properties that are related to the scene and that do not refer to individual objects. Our goal here is to use such a scheme for including context information in object representations and to demonstrate its role in facilitating object detection (Torralba, 2003a, 2003b).

As illustrated in Fig. 96.2, the analysis of the image is performed using two parallel pathways: one local (e.g., objects) and one global (e.g., scenes). Here we describe the features that can be used in both pathways.

A. Local Features

Most models of attention and object recognition rely on the definition of sets of local features. In the local

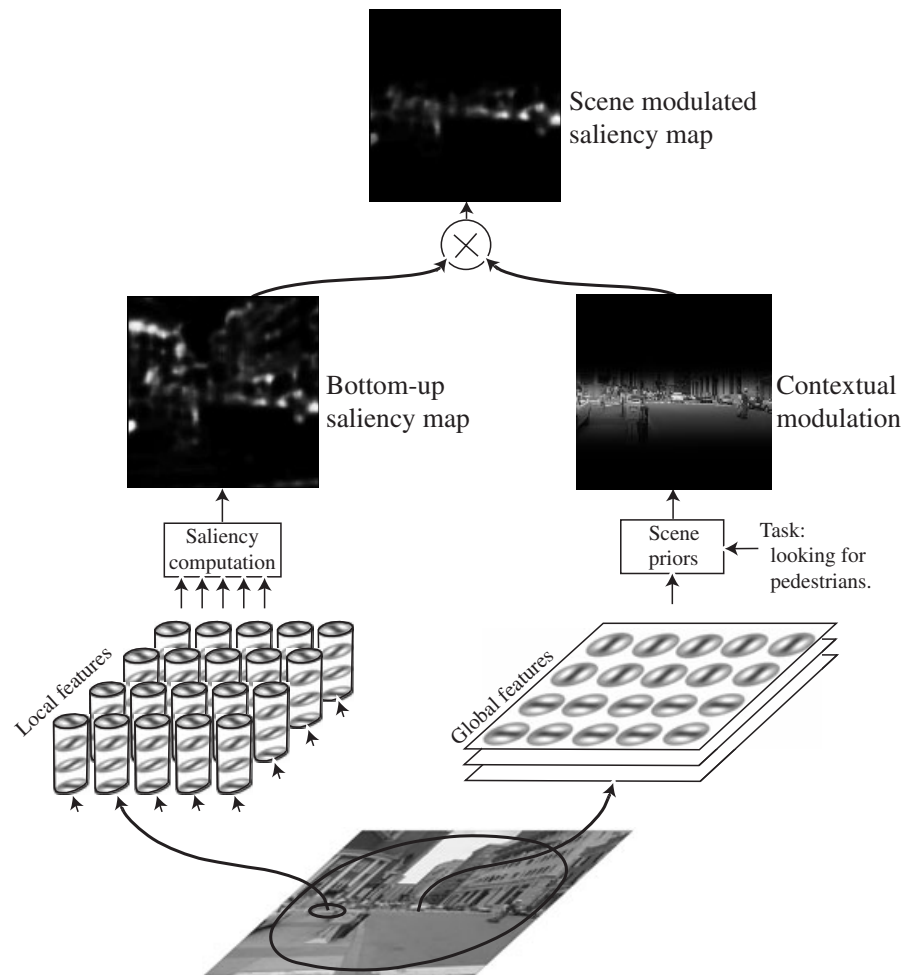


FIGURE 96.2 Contextual and local pathways for pre-attentive object search. This scheme incorporates contextual information for modulating image saliency. The scheme consists in two parallel pathways: the first one processes local image information; the second one encodes globally the pattern of activation of the feature maps. When looking for a person in the image, the saliency map, which is task independent, selects image regions that are salient in terms of local orientations and spatial frequencies. However, the contextual priming (task dependent) drives attention to the image regions that can contain the target object (sidewalks for pedestrian). The final attentional map, obtained as the product of both maps, selects the salient locations in side the image region relevant for the task.

pathway, each location is represented by a vector of features that describes local image properties. It could be a collection of templates (e.g., object detection) or a vector composed by the output of wavelets at different orientations and scales (e.g., saliency models of attention).

For instance, in Fig. 96.2, each local feature vector is a jet of filter responses: $v_1(x) = \{g_1(x), g_2(x), \dots, g_N(x)\}$. Following the structure of the receptive fields of simple and complex cells in V1, the features $g_k(x)$ used here are obtained as: $g_k(x) = |\sum_{x'} I(x') h_k(x' - x)|^2$, where $I(x)$ is the input image and $h_k(x)$ is a Gabor-like wavelet tuned in orientation and scale.

B. Global Features

In the global pathway, the entire image is represented by a unique set of features that summarizes the appearance of the scene without encoding specific objects or regions. In the example in Fig. 96.2, the global feature shown responds to a combination of the output of oriented filters at different image locations (Oliva and Torralba, 2001): $v_c = \{\sum_x \sum_k g_k(x) \phi_m(x, k); m = 1, M\}$, where $\phi_m(x, k)$ is a set of weights that specify how to combine the outputs of the local features $g_k(x)$ to build a global feature v_c . M is the total number of global features.

In the toy example in Fig. 96.2, the global feature responds strongly to images with horizontal structures in the bottom half of the image and vertical structures in the upper half of the image (this organization corresponds to the typical structure of a street scene). The global image representation is built by a collection of such kind of features.

In the next section, we describe how both local and global features can be combined to introduce contextual factors in attention.

IV. MODEL FOR SCENE PRIORS AND THE MODULATION OF SALIENCY

Here we describe a Bayesian framework (e.g., Kersten and Yuille, 2003) for object search that integrates saliency, object appearance, and scene priors in order to guide attention (Torralba 2003a, 2003b). In a statistical framework, when looking for a target (o represents the object category), at each image location (x) and scale of analysis (σ) a probability of containing the target is assigned: $p(o, x, \sigma, \alpha | \mathbf{v}_l, \mathbf{v}_c)$. t is a vector whose parameters describe the appearance of the target (e.g., point of view). The probability is conditional on the local and global image features. The object probability function can be decomposed applying Bayes rule as:

3

$$p(O | \mathbf{v}_l, \mathbf{v}_c) = \frac{1}{p(\mathbf{v}_l | \mathbf{v}_c)} p(\mathbf{v}_l | O, \mathbf{v}_c) p(O | \mathbf{v}_c) \quad (1)$$

For simplicity of notation we have grouped all the variables that describe the appearance of the object in the image as: $O = \{o, x, \sigma, \alpha\}$. The three factors in Eq. (1) provide a simplified framework for representing three levels of attention guidance (Torralba, 2003b).

A. Saliency

The normalization factor, $1/p(\mathbf{v}_l | \mathbf{v}_c)$, does not depend on the target or task constraints and therefore is a bottom-up factor. It provides a measure of how unlikely it is to find a set of local measurements \mathbf{v}_l within the context \mathbf{v}_c . We can define local saliency as $S(x) = 1/p(v_l(x) | \mathbf{v}_c)$. This probabilistic definition of saliency fits more naturally with object detection and recognition formulations.

This formulation follows the hypothesis that frequent image features are more likely to belong to the background, whereas rare image features are more likely to be diagnostic features for the detection of (interesting) objects. Note that the term $S(x)$ does not incorporate any information about the appearance of the target. We approximate $S(x)$ by fitting a Gaussian to the distribution of local features in the image.

B. Target-Driven Control of Attention

The second factor, $p(\mathbf{v}_l | O, \mathbf{v}_c)$, gives the likelihood of the local measurements \mathbf{v}_l when the object O is present in a particular context. This factor represents the top-down knowledge of the target appearance and how it contributes to the search (Rao et al., 1996). Regions of the image with features unlikely to belong to the target object are vetoed and regions with attended features are enhanced (see Chapter 17). Note that when the object properties O fully constraint the object appearance, it is possible to approximate $p(v_l | O, \mathbf{v}_c) \approx p(v_l | O)$. This approximation allows the dissociation of the contribution of local image features and global (contextual) image features.

C. Scene Priors

The third factor, the PDF $p(O | \mathbf{v}_c)$, provides context-based priors on object class, location, scale, and appearance. This term is of capital importance for ensuring reliable inferences in situations in which the local image measurements \mathbf{v}_l produce ambiguous interpretations. This factor does not depend on local measurements or target models.

Using the definition of an object in a scene, $O = \{o, x, \sigma, \alpha\}$, contextual influences become more evident if we apply Bayes rule successively in order to split the PDF $p(O | \mathbf{v}_c)$ into several factors that model different kinds of scene priors for object search:

$$p(O | \mathbf{v}_c) = p(\alpha | x, \mathbf{v}_c, o) p(\sigma | x, \mathbf{v}_c, o) p(x | \mathbf{v}_c, o) P(o | \mathbf{v}_c) \quad (2)$$

According to this decomposition of the PDF, the contextual modulation of target saliency is a function of four factors:

- *Object-class priming:* $P(o | \mathbf{v}_c)$ provides the probability of presence of the object class o in the scene. If $P(o | \mathbf{v}_c)$ is very small, then object search need not be initiated.
- *Contextual control of focus of attention:* $p(x | o, \mathbf{v}_c)$. This PDF gives the most likely locations for the presence of object o given context information.
- *Contextual selection of scale:* $p(\sigma | x, o, \mathbf{v}_c)$. This gives the likely size of the object o in the context \mathbf{v}_c . When looking for an object, the expected size of the target determines the scanning resolution that needs to be used when exploring the image.
- *Contextual selection of target appearance:* $p(\alpha | x, o, \mathbf{v}_c)$. This gives the expected shapes (point of views, aspect ratio) of the object.

Most popular computational models of object recognition focus on modeling the probability function $p(O | \mathbf{v}_l)$ without taking into account contextual priors.

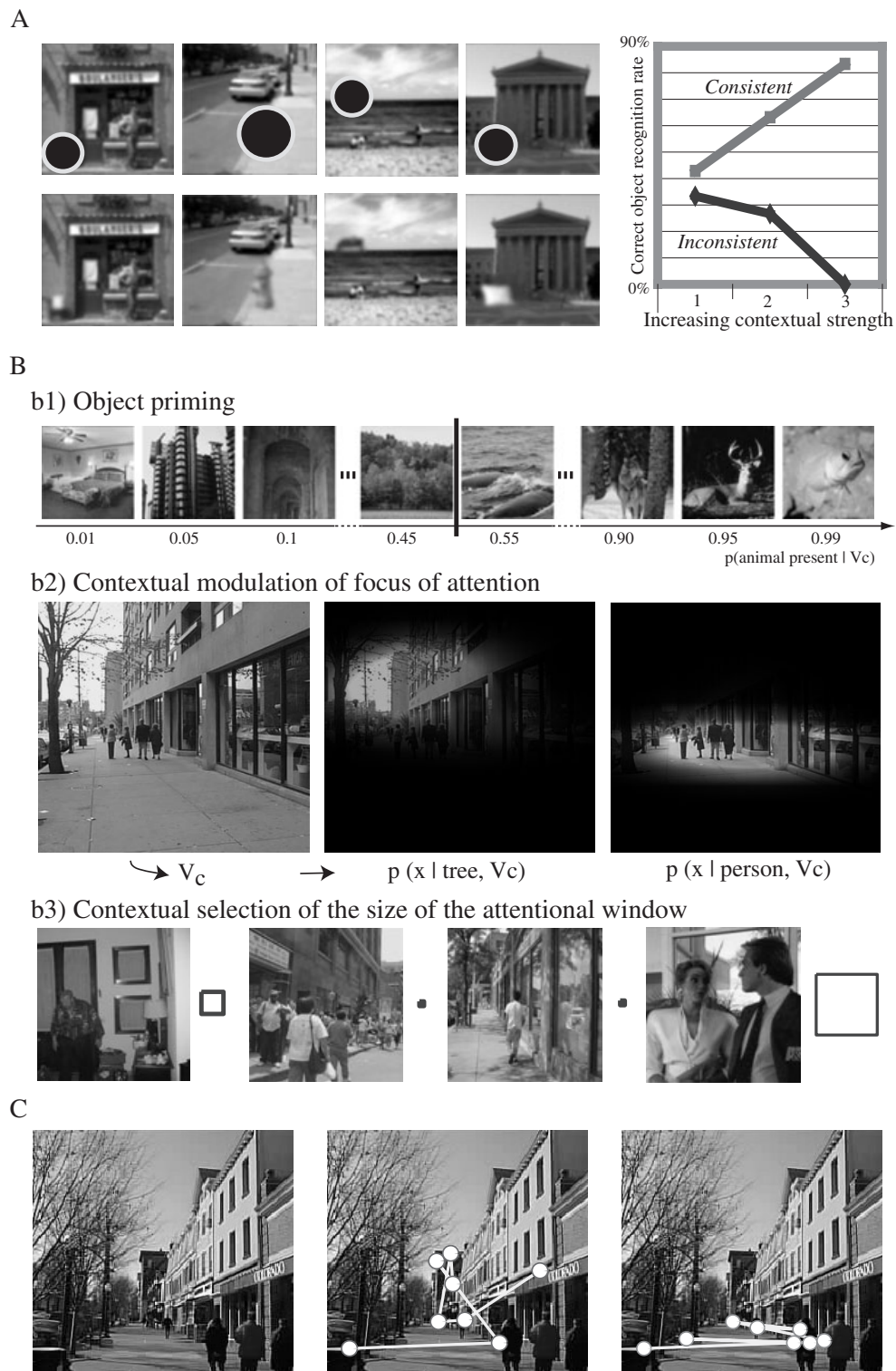


FIGURE 96.3 (A) Role of contextual priors on object recognition by subjects. (B) Scene priors obtained from global image features. (C) Examples of salient locations obtained from saliency alone (center) and combining both context and saliency (right). Including scene priors provides better predictions for the location of the target.

V. RESULTS

Figure 96.3A shows the effect that contextual priors $p(O | \mathbf{v}_c)$ have on subject performances for recognition. First, subjects are asked to guess the identity of the objects behind the masks (Fig. 96.3A, top). That experiment allows us to evaluate the distribution of objects that subjects are considering for each scene: $P(\text{objects} | x, \sigma, \text{scene})$. Then, we can sort the scenes according to the strength of the priors (by measuring the entropy of the distributions). In a second experiment, we show how the strength of these priors affect recognition. We ask subjects to recognize blurred objects when they are placed in consistent and inconsistent backgrounds. The results (Fig. 96.3A, right) show that observer's performance on a recognition task is correlated with the strength of the priors (Bar, 2003; Chapter 25).

Figure 96.3B summarizes the results of the contextual model. The role of the contextual priors in modulating attention is to provide information about past search experience in similar environments and the strategies that were successful in finding the target. In this model, we assume that the contextual features \mathbf{v}_c already carries all the information needed to identify the scene and that the scene is identified at a glance, without requiring eye movements. The eye movements are only required in order to analyze in detail regions of the image that are relevant for a task (i.e., to find somebody). The contextual priors $p(O | \mathbf{v}_c)$ contain the information about how the scene features \mathbf{v}_c were related to the target properties O (image location, scale, and pose) during the past experience. The system is trained by our first providing the system a collection of images in which the target has been already located. The PDF is learned using a mixture of gaussians and the EM algorithm (Torralla, 2003a). Once the system has learned the relationship between scenes and objects, it can predict the expected locations for several objects in new scenes (Fig. 96.3B, right).

Figure 96.3B provides examples of the results of global contextual priming for:

- (b1) Predicting the presence or absence of objects. Here we show the results for solving the task of animal present/absent using only scene priors, before scanning the image (Torralla and Oliva, 2003). The system has an 80% correct prediction rate on this task.
- (b2) Focus of attention (e.g., expected locations of people and trees). Contextual priors for location reduce the area of the image that needs to be explored when looking for the target.

- (b3) Scale selection (e.g., expected size of face in the image).

Finally, Fig. 96.3C compares the salient points and the region of interest predicted by a model using only bottom-up saliency maps (center) and when combining saliency and the scene priors for location (right). When including scene priors, the candidate locations are only within the image region that has a high probability of containing the target. Experiments show that including scene priors provides better predictions of human eye movements than saliency alone (Oliva et al., 2003).

VI. CONCLUSION

The model proposed includes scene priors for object search early in the visual processing chain. Therefore, the scene priors constitute an effective shortcut for object detection as it provides priors for the object's presence/absence before scanning the image.

From an algorithmic point of view, contextual control of the focus of attention is important because it avoids expending computational resources in spatial locations with a low probability of containing the target based on prior experience. It also provides criteria for rejecting possible false detections or salient features that fall outside the primed region.

References

- Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* **15**, 600–609.
- Biederman, I., Mezzanotte, R. J., and Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cogn. Psychol.* **14**, 143–177.
- Chun, M. M., and Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cogn. Psychol.* **36**, 28–71.
- Henderson, J. M., and Hollingworth, A. (1999). High level scene perception. *Annu. Rev. Psychol.* **50**, 243–271.
- Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Vis.* **20**, 1254.
- Kersten, D., and Yuille, A. (2003). Bayesian models of object perception. *Curr. Opin. Neurobiol.* **13**, 150–158.
- Oliva, A., and Torralla, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comp. Vis.* **42**, 145–175.
- Oliva, A., Torralla, A., Castelano, M. S., and Henderson, J. M. (2003). Top-down control of visual attention in object detection. *IEEE Proceedings of the International Conference on Image Processing*, Barcelona (Spain).
- Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory and Cogn.* **3**, 519–526.

- 5 Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., and Ballard, D. H. (1996). Modeling saccadic targeting in visual search. "NIPS'95." MIT Press, Cambridge, MA.
- Rensink, R. A., O'Regan, J. K., and Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychol. Sci.* **8**, 368–373.
- Schyns, P. G., and Oliva, A. (1994). From blobs to boundary edges: evidence for time and spatial scale dependent scene recognition. *Psychol. Sci.* **5**, 195–200.
- Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* **381**, 520–522.
- Torralba, A. (2003a). Contextual priming for object detection. *Int. J. Comp. Vis.* **53**, 169–191.
- Torralba, A. (2003b). Modeling global scene factors in attention. *J. Opt. Soc. Am. A.* **20**, 1407–1418.
- Torralba, A., and Oliva, A. (2003). Statistics of natural image categories. *Network: Computation Neural Syst.* **14**, 391–412.
- Treisman, A., and Gelade, G. (1980). A feature integration theory of attention. *Cogn. Psychol.* **12**, 97–136.
- Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bull. Rev.* **1**, 202–228.

AUTHOR QUERY FORM

Dear Author,

During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof.

Many thanks for your assistance.

Query References	Query	Remarks
1	Au: Treisman & Gelade (1983) Not in Refs	
2	Au: Please confirm Fig. 96.1B (right) is correct	
3	Au: I don't see t vector in these formulas. Please advise.	
4	Au: Please provide editor, pages, publisher, city of pub	
5	Au: Please provide editor, pages, full title	