Modeling global scene factors in attention

Antonio Torralba

Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 400 Technology Square, Cambridge, Massachusetts 02115

Received October 1, 2002; revised manuscript received February 12, 2003; accepted March 12, 2003

Models of visual attention have focused predominantly on bottom-up approaches that ignored structured contextual and scene information. I propose a model of contextual cueing for attention guidance based on the global scene configuration. It is shown that the statistics of low-level features across the whole image can be used to prime the presence or absence of objects in the scene and to predict their location, scale, and appearance before exploring the image. In this scheme, visual context information can become available early in the visual processing chain, which allows modulation of the saliency of image regions and provides an efficient shortcut for object detection and recognition. © 2003 Optical Society of America

OCIS codes: 330.0330, 330.4060, 100.5010.

1. INTRODUCTION

As illustrated in Fig. 1(a), contextual information allows one to unambiguously recognize an object, such as a pedestrian, even when the intrinsic local information is not sufficient for reliable recognition. Note that blurring of the image has reduced the available local information. Figure 1(b) illustrates that even when there is enough local information for object recognition, contextual information plays a major role in object search. The target object can be localized easily when it is clearly differentiated from the background as well as when the background context constrains the possible locations of the target.

A popular strategy for finding an object embedded in a complex background is to look for local features that differ the most from the rest of the image. Accordingly, most computational models of attention are based on low-level saliency maps that ignore contextual information provided by the correlation between objects and the scene.[1–5] A second class of models of attention include information about the appearance of the target in the search process.[5–7] This approach ignores image regions that have features incompatible with the target and enhances the saliency of regions that have features compatible with the target. But again, no contextual information is taken into account.

A number of studies have shown the importance of scene factors in object search and recognition. Yarbus[8] showed that the task changes the way observers look at a scene. Studies by Biederman *et al.*[9] and Palmer[10] highlight the importance of contextual information for object search and recognition. Rensink and co-workers[11,12] showed 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[13] reported results suggesting that the choice of these regions is governed not merely by their low-level saliency but also by scene semantics.[14] Chun and Jiang[15] showed that visual search is facilitated when there is correlation across different trials between the contextual configuration of the display and the target location. Oliva *et al.*[16] showed also that familiar scenes automatically activate visual search strategies that were successful in past experiences, without volitional control by the subject. All these results are in agreement with the idea that scene information can be processed fast and without relying on single objects.[17] Schyns and Oliva[17] showed that a coarse representation of the scene initiates semantic recognition before the identity of objects is processed. Several other studies support this idea that scene semantics can be available early in the chain of information processing[18–21] and suggest that scene recognition may not require object recognition as a first step.[22–24]

Notwithstanding the accumulating evidence for contextual effects on visual exploration, few models of visual search and attention proposed so far include the use of context.[25–30] In this paper a statistical framework for incorporating contextual information in the search task is proposed.

2. COMPUTATIONAL MODELS OF ATTENTION

In this section, first, saliency-based models of attention are introduced. Then a probabilistic framework is introduced in which a model that incorporates both low-level factors (saliency) and higher-level factors (scene/context) for directing the focus of attention is proposed.

A. Saliency-Based Models of Attention

In the feature-integration theory,[3] attention is driven by low-level features, and the search for objects is believed to require slow serial scanning; low-level features are integrated into single objects when attention focuses on them. Computational models of visual attention (saliency maps) have been inspired by this approach, as it allows a simple and direct implementation of bottom-up attentional mechanisms.[1,28,31,32]

Saliency maps provide a measure of the saliency of each location in the image based on low-level features such as contrast, color, orientation, texture, and motion

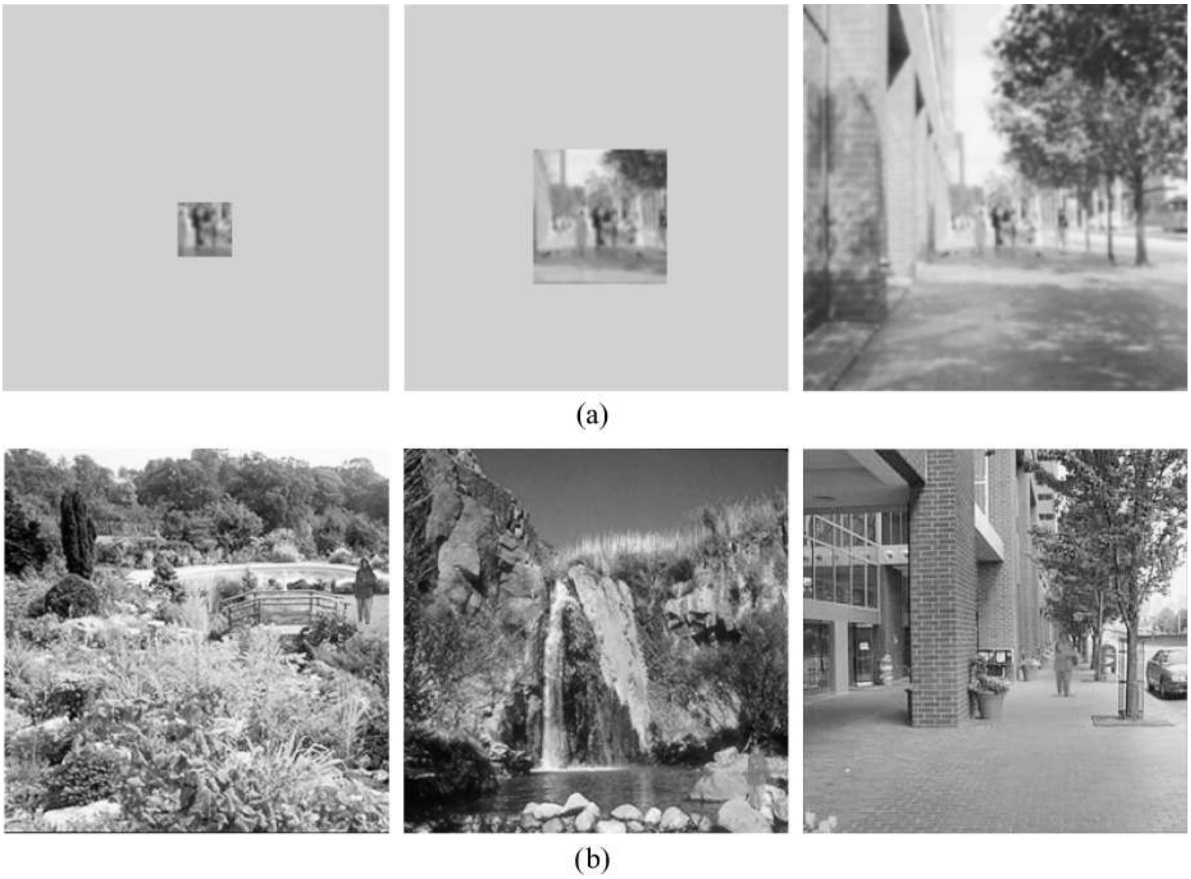

Fig. 1. Examples of scene/context influences in object recognition and search. (a) Examples with increasing contextual information but where the local target information remains constant. Observer recognition improves drastically as background information is added. (b) Scene information affects the efficiency of search for a target (a person). The context acts in two competing ways: (1) by introducing distractors but also (2) by offering more constraints on target location. Most previous models of attention focus on modeling masking effects (saliency maps).

(see Ref. 33 for a review). In the saliency-map model, regions with properties different from those of their neighborhoods are considered more informative and therefore attract attention. When one is looking for an object, a possible strategy for exploring the image is analyzing first the salient regions in the scene. This approach is very efficient when the target object is distinct from the background in terms of simple low-level image features.

The image features most commonly used for describing local image structure (orientation, scale, and texture) are the outputs of multiscale oriented bandpass filters. Gabor-like filters have interesting properties for the encoding of natural images.[34,35] These features have also been shown to be relevant for the tasks of object detection[6,36] and scene recognition.[24,37–39] These image features are obtained as the convolution

$$v_k(\mathbf{x}) = \sum_{\mathbf{x}'} i(\mathbf{x}')g_k(\mathbf{x} - \mathbf{x}'), \tag{1}$$

where $i(\mathbf{x})$ is the input image, $u_k(\mathbf{x})$ is the output image, and $g_k(\mathbf{x})$ is an oriented bandpass Gabor filter defined by $g_k(\mathbf{x}) = \exp(-\|\mathbf{x}\|^2/\sigma_k^2)\exp(2\pi j\langle\mathbf{f_k}, \mathbf{x}\rangle)$. The real and imaginary parts of g_k are Gabor filters in quadrature. The variable k indexes filters tuned to different orientations and scales.

For each feature a saliency map is computed by using a hardwire scheme[1]: the output of each filter is processed by center–surround mechanisms to enhance regions with outputs that differ from the neighborhood. The results are then amplitude normalized and combined in order to produce a unique saliency map that also combines information coming from other image features such as contrast and color.[1] The scanning of attention is then modeled as exploring the image in succession following regions of decreasing saliency. Also, the set of low-level features can be enhanced to account for other image properties that also attract attention such as edge length and curvature.[4]

Most of the computations for the saliency map rely only on local measurements, and global information is considered only in the normalization step. Saliency maps are easy to compute and can be hard-wired into the design of a system, thus minimizing the need for learning. However, their reliance on local measures forces bottom-up models to treat the background as a collection of distracting elements that complicate rather than facilitate the search for specific objects. Most of the time, the target object is camouflaged in a cluttered background. In such situations, saliency alone may not provide an efficient way of finding an object.

Low-level saliency maps compute the saliency of a local image region by looking at the distribution of features in the neighborhood (context) of each location. This use of context treats the background as a collection of distractors instead of exploiting the correlations that exist between target properties and background properties. In the next sections a term will be introduced into the saliency of a target that depends on the correlation between object and scene properties. In the model presented in this paper, contextual features refer to image features correlated with a higher-level scene description.

B. Model for Contextual Modulation of Target Saliency

In contrast to traditional models of search for which background scene information is more of a hindrance than a help, the statistical correlations that exist between global scene structure and object properties will be used to facilitate search in complex scenes. This approach is motivated by the observation that the structure of many real-world scenes is governed by strong configurational rules (Fig. 2). These statistical regularities can provide estimates of the likelihood of finding an object in a given scene and can also indicate the most likely position and scales at which an object might appear. Next to be described is how these intuitions have been formalized into an operational scheme for object search in complex scenes. A statistical framework will be used for the model, as it provides a simple way of accounting for different factors in evaluating the target saliency.

The object in a scene is described here by means of a set of parameters $O = \{o, \mathbf{x}, \mathbf{t}\}$, where o denotes the category of the object, $\mathbf{x} = (x, y)$ is its spatial location, and $\mathbf{t}$ are object appearance parameters such as the object's scale in the image and its pose. In a statistical framework, object search requires the evaluation of the probability density function[7,36] (PDF), $P(O|\mathbf{v})$. This function is the probability of the presence of an object O in a scene given a set of image features $\mathbf{v}$. Here $\mathbf{v}$ represents all the features obtained from the input image. Therefore $\mathbf{v}$ is very high dimensional, making the evaluation of the function $P(O|\mathbf{v})$ impractical. Furthermore, writing the object probability as $P(O|\mathbf{v})$ does not reveal how scene or object features might influence the search, because it does not differentiate between local and contextual features. Therefore we consider two sets of image features: (1) local features, $\mathbf{v}_L(\mathbf{x})$, which are the set of features obtained in a neighborhood of the location $\mathbf{x}$, and (2) contextual features, $\mathbf{v}_C$, which encode structural properties of the scene/background. Here it is proposed that object detection requires the evaluation of the probability function (target saliency function), $P(O|\mathbf{v}_L, \mathbf{v}_C)$, which provides the probability of the presence of the object O given a set of local and contextual measurements.[29,30,40]

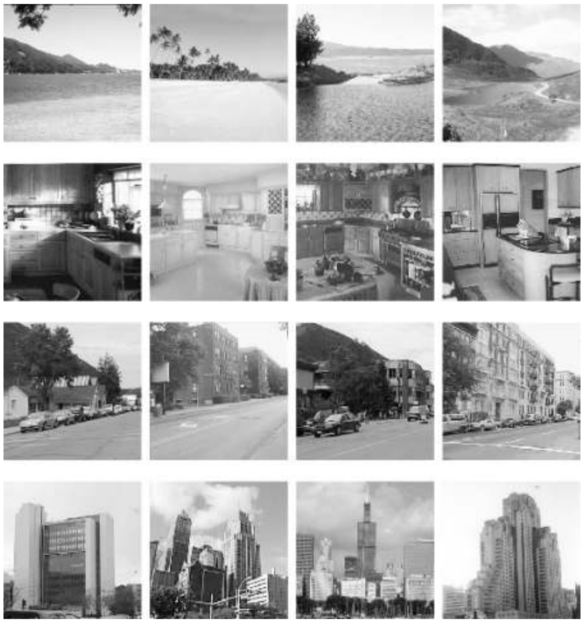

Fig. 2. Images that are similar in terms of global spatial properties have a tendency to be composed of similar objects with similar spatial arrangement.[17,24] Since scene semantics may be available early in the visual processing, these regularities suggest that an efficient procedure for object search in a new scene is to see how objects were organized in similar environments.

The object probability function can be decomposed by applying Bayes's rule as

$$P(O|\mathbf{v}_L, \mathbf{v}_C) = \frac{1}{P(\mathbf{v}_L|\mathbf{v}_C)} P(\mathbf{v}_L|O, \mathbf{v}_C) P(O|\mathbf{v}_C). \quad (2)$$

Those three factors provide a simplified framework for representing three levels of attention guidance.

1. *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 $S(\mathbf{x}) = 1/P(\mathbf{v}_L(\mathbf{x})|\mathbf{v}_C)$. Saliency is large for unlikely features in a scene.

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 key features[40] for the detection of (interesting) objects.

2. *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. Regions of the image with features unlikely to belong to the target object are vetoed, and regions with attended features are enhanced.[33,41] Note that when the object properties O fully constrain the object appearance, then it is possible to approximate $P(\mathbf{v}_L|O, \mathbf{v}_C) \simeq P(\mathbf{v}_L|O)$. This is a good approximation, because O does not just include the definition of the object category (e.g., a car) but also specifies information about the appearance of the target (location, scale, pose, etc.). This approximation allows dissociation of the contribution of local image features and contextual image features.

3. *Contextual Priors*

The third factor, the PDF $P(O|\mathbf{v}_C)$, provides context-based priors on object class, location, and scale.[29,30] It is of critical importance for ensuring reliable inferences in situations where the local image measurements $\mathbf{v}_L$ produce ambiguous interpretations.[40] This factor does not depend on local measurements and target models. Therefore the term $P(O|\mathbf{v}_C)$ modulates the saliency of local image properties in the search for an object of the class o.

With an object in a scene defined as $O = \{o, \mathbf{x}, \mathbf{t}\}$, contextual influences become more evident if we apply Bayes's rule successively in order to split the PDF $P(O|\mathbf{v}_C)$ into three factors that model three kinds of context priming on object search:

$$P(O|\mathbf{v}_C) = P(\mathbf{t}|\mathbf{x}, \mathbf{v}_C, o)P(\mathbf{x}|\mathbf{v}_C, o)P(o|\mathbf{v}_C). \quad (3)$$

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

1. Object-class priming: $P(o|\mathbf{v}_C)$. This PDF 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 (e.g., we do not need to look for cars in a living room).

2. Contextual control of focus of attention: $P(\mathbf{x}|o, \mathbf{v}_C)$. This PDF gives the most likely locations for the presence of object o given context information, and it allocates computational resources into relevant scene regions.

3. Contextual selection of local target appearance: $P(\mathbf{t}|\mathbf{x}, \mathbf{v}_C, o)$. This PDF gives the likely (prototypical) shapes (point of views, size, aspect ratio, object aspect) of the object o in the context $\mathbf{v}_C$. For instance, $\mathbf{t} = \{\sigma, p, \ldots\}$, σ being scale and p being aspect ratio. Other parameters describing the appearance of an object in an image can be added.

Computational models of object recognition have focused on modeling the probability function $P(O|\mathbf{v}_L)$, ignoring contextual priors.[7,36,42–46]

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 carry all the information needed to identify the scene. The scene is identified at a glance, without the need for eye movements.[18,19,22] Eye movements are required for a detailed analysis of regions of the image that are relevant for a task (e.g., to find somebody).

Scene factors in attentional deployment are effective

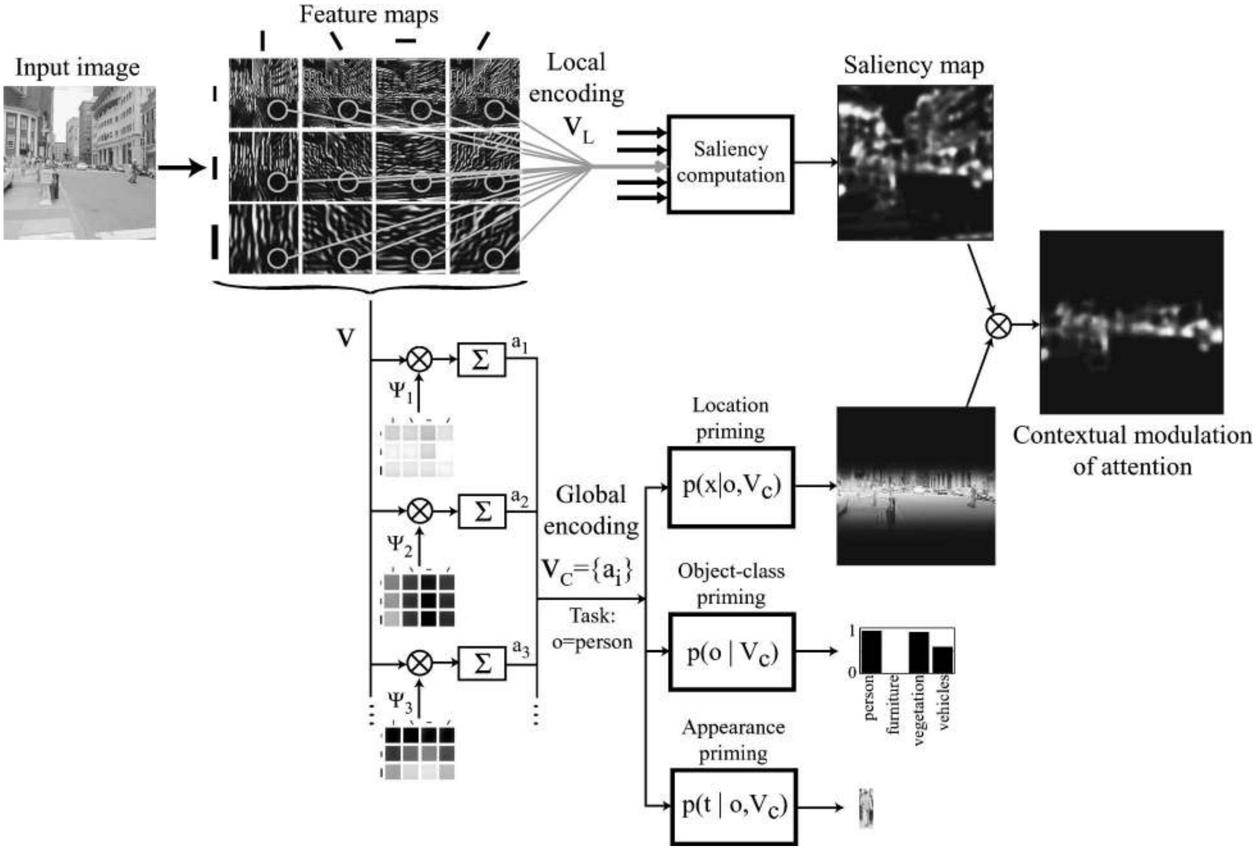

Fig. 3. Scheme that incorporates contextual information to select candidate target locations. The scheme consists of two parallel pathways: The first processes local image information, and the second encodes globally the pattern of activation of the feature maps. Contextual information is obtained by projecting the feature maps into the (holistic) principal components. In the task of looking for a person in the image, the saliency map, which is task independent, will select image regions that are salient in terms of local orientations and spatial frequencies. However, the contextual priming (task dependent) will drive attention to the image regions that can contain the target object (sidewalks for pedestrian). Combining context and saliency gives better candidates for the location of the target.

only after the system has accumulated enough experience. That is when the system knows about the regularities of the visual world: how objects and scenes are related and which visual search strategies are successful in finding objects of interest, given a visual context. Therefore the likelihood function $P(O|\mathbf{v}_C)$ contains the information about how the scene features $\mathbf{v}_C$ were related to the target properties O (image location, scale, position) during previous experience.

As shown in Eq. (3), the contextual priming can be decomposed into three factors: object-class priming, contextual guidance of focus of attention, and contextual selection of object appearance. In the next sections it will be shown how to compute global scene features[24,47] and how these features can be used in the object search task.

C. Global Image Features and Context Representation

As discussed in Subsection 2.A, the image features most commonly used for describing local image structure are the outputs of multiscale oriented bandpass filters. Therefore the local image representation at the spatial location $(\mathbf{x})$ is given by the vector $\mathbf{v}_L(\mathbf{x}) = \{v_k(\mathbf{x})\}_{k=1,N}$, where $v_k(\mathbf{x})$ is given by Eq. (1) and N is the number of Gabor filters. In such a representation, $v(\mathbf{x}, k)$ is the output magnitude at the location $\mathbf{x}$ of a complex Gabor filter tuned to the spatial frequency $\mathbf{f}_k$. The variable k indexes filters tuned to different spatial frequencies and orientations.

Contextual features have to describe the structure of the whole image. It has been shown that a holistic low-dimensional encoding of the image features conveys relevant information for a semantic categorization of the scene/context[24,47] and can be used for contextual priming in object-recognition tasks.[29,30] This definition of context does not require the identification of other objects in the scene.

Such a representation can be achieved by decomposing the image features into the basis functions provided by principal components analysis:

$$a_n = \sum_{\mathbf{x}} \sum_{k} |v(\mathbf{x}, k)| \, \psi_n(\mathbf{x}, k). \tag{4}$$

I propose to use the decomposition coefficients $\mathbf{v}_C = \{a_n\}_{n=1,N}$ as context features. The functions ψ_n are the eigenfunctions of the covariance matrix defined by the image features $v(\mathbf{x}, k)$. Figure 3 illustrates how the features a_n are obtained from the magnitude output of the Gabor bank. Each feature a_i is obtained as a linear combination of the magnitude output of all the Gabor filters used in the image decomposition.[24] By using only a reduced set of components ($N = 60$ for the rest of the paper; we use a filter bank with six orientations and four scales), the coefficients $\{a_n\}_{n=1,N}$ encode the main spectral characteristics of the scene with a coarse description of their spatial arrangement. In essence, $\{a_n\}_{n=1,N}$ is a holistic representation, as all the regions of the image contribute to all the coefficients, and objects are not encoded individually.

In the next sections we discuss each of these three factors and show results using an annotated database of real-world images (see Appendix A).

3. RESULTS

A. Object-Class Priming

Before attention is deployed across the different parts of the scene, the global configuration may be a strong indicator of the presence or absence of an object. If the scene has a layout in which, given previous experience, the target was rarely present, then the system can rapidly decide not to initiate the search. If we assume that the feature vector $\mathbf{v}_C$ conveys enough information about the identity of the context, then there should exist strong priors on object identities, at least at the superordinate level (people, furniture, vehicles, vegetation, etc.). For instance, contextual object priming should capture the fact that while we do not expect to find cars in a room, we do expect a high probability of finding furniture.

These intuitions are formalized by means of the PDF $P(o|\mathbf{v}_C)$ that gives the probability of presence of the object class o given contextual information $\mathbf{v}_C$. For instance, if for a scene we obtain $P(o|\mathbf{v}_C) \sim 1$, then we can be almost certain about the presence of the object class o in the scene even before exploring the image in detail. On the other hand, if $P(o|\mathbf{v}_C) \sim 0$, then we can decide that the object is absent and forego initiating search. The number of scenes in which the system may be able to make high-confidence decisions will depend on various factors such as the strength of the relationship between the target object and its context and the ability of the features $\mathbf{v}_C$ to characterize the context efficiently. The function $P(o|\mathbf{v}_C)$ is learned by using an annotated image database (see Appendix A, Subsection B).

Figure 4 shows some typical results from the priming model for four categories of objects (people, furniture, vegetation, and vehicles). For each category, high-confidence predictions were made in at least 50% of the tested scenes, and presence or absence was correctly predicted by the model on 95% of those images. When the model was forced to make binary decisions in all the images (by selecting an acceptance threshold of 0.5), the presence or absence of the objects was correctly predicted by the model on 81% of the scenes of the test set. Images in the test set were selected such that a random guess about the presence or absence of an object gives 50% correct predictions.

The results reveal the ability of the contextual features to distinguish between different environments. Object priming provides an efficient technique for reducing the set of possible objects that are likely to be found within the scene and for determining whether search needs to be initiated.

B. Contextual Guidance of Focus of Attention

The PDF $P(\mathbf{x}|o, \mathbf{v}_C)$ indicates the most likely locations for the presence of the object class o given context information. This PDF can be thought of as the input to an attentional system that directs computational resources (focus of attention) toward regions more likely to contain an object of interest. It also provides criteria for rejecting possible false detections that fall outside the primed region. When the target is small (a few pixels), the problem of detection using only object intrinsic (local) features is ill-posed. As illustrated in Fig. 1(a), in the absence of contextual information, local information might not be enough for reliable recognition, because when only local information is used, similar arrangements of pixel intensities can be found in other regions of the image. For instance, some of the pedestrians in Fig. 5 are so small as to be mere scratches on the image. Similar scratches can be found in other locations of the picture, but given the context information they are not considered potential targets because they fall outside the likely "pedestrian region." During the learning stage (see Appendix A, Subsec-

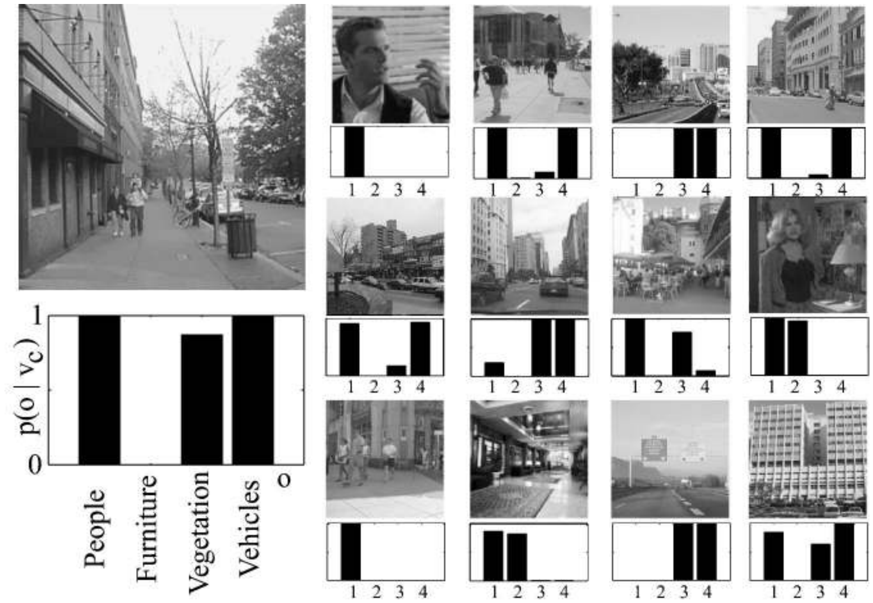

Fig. 4. Contextual priming of superordinate object categories (1, people; 2, furniture; 3, vegetation; 4, vehicles). The heights of the bars show the model's predictions of the likelihood $P(o|\mathbf{v}_C)$ of finding members of these four categories in each scene.

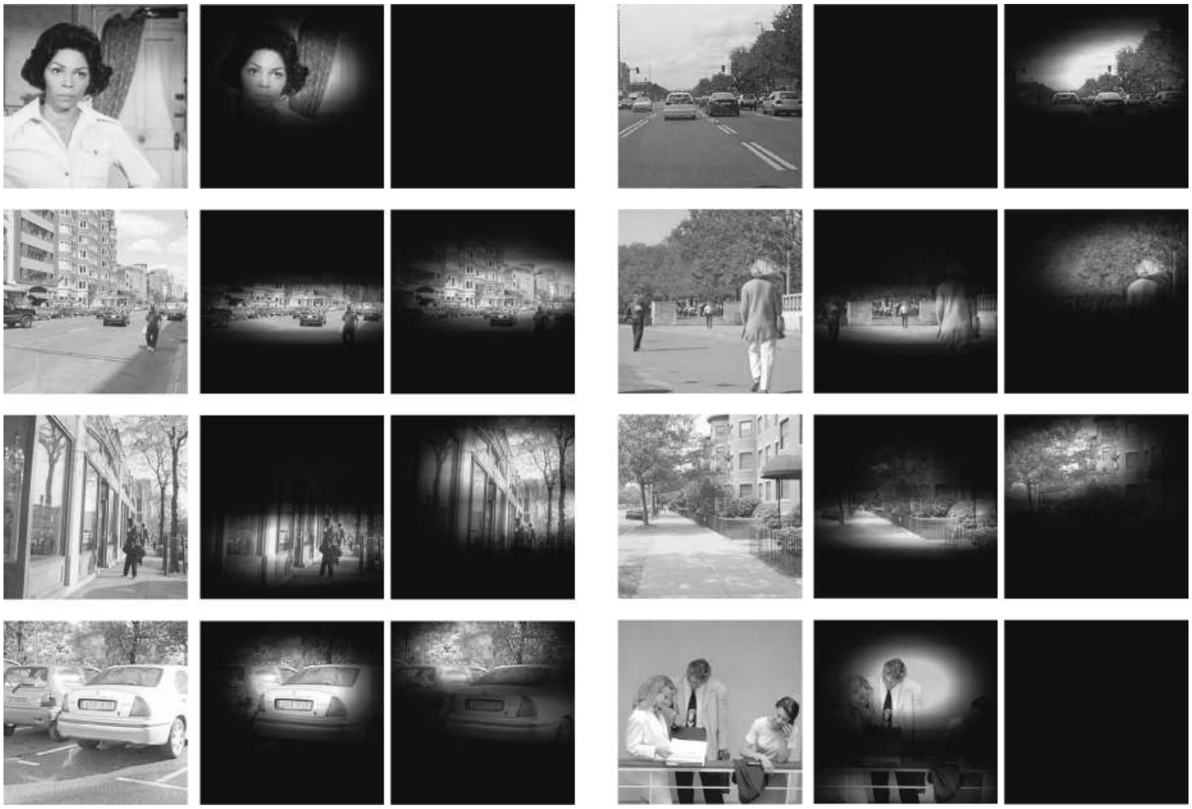

Fig. 5. Model results on context-driven focus of attention in the task of looking for faces (left) and vegetation (right). Examples of real-world scenes and the image regions with the largest likelihood $P(\mathbf{x}, o|\mathbf{v}_C) = P(\mathbf{x}, o|\mathbf{v}_C)P(o|\mathbf{v}_C)$. The two foci of attention for each image show how the task (o = faces or o = trees) changes the way attention is deployed in the image in considering scene/context information. The factor $P(o|\mathbf{v}_C)$ is included here to illustrate how attention is not driven to any image region when the target object o is inconsistent with the context (e.g., trees in an indoor scene or pedestrians on a highway).

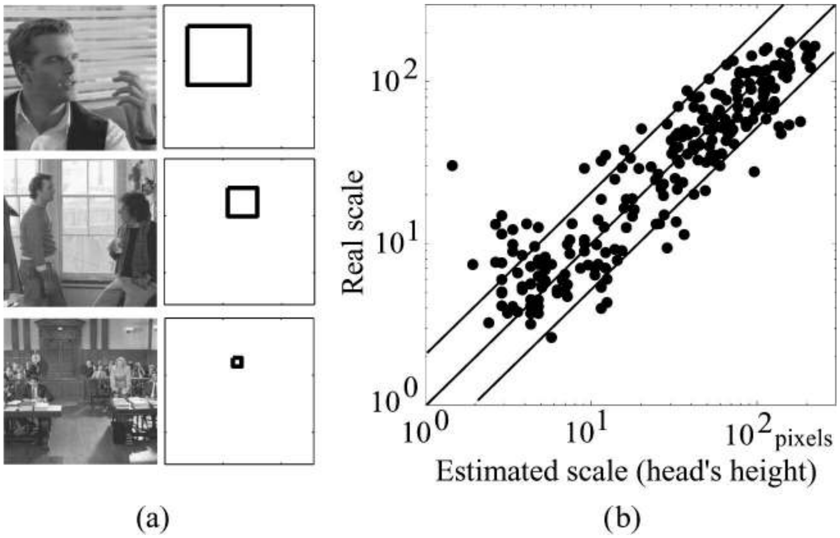

Fig. 6. Scale priming from a familiar context. (a) Examples of scenes and the model's estimate of the size of a face at the center of focus of attention. (b) Scale estimation results plotted against ground truth.

tion B for a description of the training database and the learning procedure), the system associates the locations of objects with the features of the context in which they are embedded. Then, given a new image, the PDF $P(\mathbf{x}|o,\ \mathbf{v}_C)$ can be used to predict the most likely locations of the target.

Figure 5 shows some results of the focus-of-attention system when the task is to look for heads and vegetation in real-world scenes. For each image, we show the PDF $P(\mathbf{x}|o,\ \mathbf{v}_C)$ superimposed on the original image to better show the selected regions. The dark regions indicate low values of the PDF and therefore image locations with low probability of containing the target object. Starting the search in the regions that are selected by the contextual priming mechanism greatly reduces the need for exhaustive search. Note also that task constraints (looking for faces or for vegetation) changes the way attention is deployed in the image when one is integrating contextual information. Note that in some of the examples of Fig. 5, attention is directed to the region that is most likely to contain the target even when the target object is not present in the scene. This illustrates the point that, at this stage, attention is driven only by global contextual features and not by the presence of local features that may belong to the target object. In the examples shown in Fig. 5, no target model is included in the selection procedure. The selected locations are chosen only as a function of task and context.

It is worth contrasting these results to those from bottom-up models in which focus of attention is mediated by low-level feature-saliency maps (see Subsection 3.D). Common to saliency models is the use of features in a local-type framework, ignoring high-level context information that is available in a global-type framework. Our model's use of the PDF $P(\mathbf{x}|o,\ \mathbf{v}_C)$ provides information that is both task-driven (looking for object o) and context-driven (given holistic context features). Figure 3 is an example of region selection using both saliency maps and context-driven focus of attention.

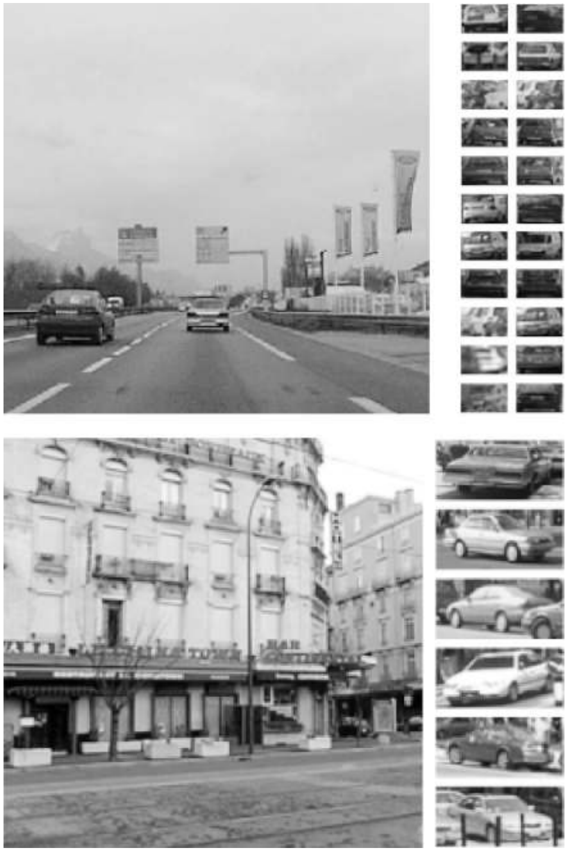

Fig. 7. Selection of prototypical object appearances based on contextual cues.

C. Contextual Selection of Object Appearance

One major problem encountered in computational approaches to object detection is the large variability in object appearance. The classical solution is to explore the space of possible shapes, looking for the best match. The main sources of variability in object appearance are size, pose (point of view), intraclass shape variability (deformations, style, etc.), and illumination effects. Including contextual information can reduce the possible appearances of the target object that are compatible with the rest of the scene. For instance, the expected size of people in an image differs greatly between an indoor environment and a perspective view of a street. The two environments produce different patterns of contextual features.[24]

Automatic scale selection is a fundamental problem in computational vision. If scale information could be estimated by an efficient preprocessing stage, then subsequent stages of object detection and recognition would be greatly simplified by focusing the processing onto only the relevant scales. As in the problem of focus of attention, existing approaches in computational vision for automatic scale selection use a low-level approach that does not rely on contextual information.[2] Here it is shown that prior knowledge about context contained in the PDF $P(\sigma|o, \mathbf{v}_C)$ provides a strong constraint on scale selection for object detection.[47] See Appendix A for a description of the training database and the learning procedure of the PDF.

Figure 6 shows several results of preattentive scale selection obtained by using the contextual priming model when it is instructed to look for heads. For each scene the mean scale of a head within the scene was estimated by computing the expected value: $\hat{\sigma} = \int \sigma P(\sigma|o, \mathbf{v}_C) \mathrm{d}\sigma$. For 84% of the images tested, the estimated mean head scale was in the interval $[h_m/2, h_m 2]$, with h_m being the actual mean scale of the heads in the picture, and for 41% of the images the estimated mean head scale was in the interval $[h_m/1.25, h_m 1.25]$.

The same procedure can be used to estimate other object parameters. For instance, context introduces strong constraints on the three-dimensional orientation (pose) of cars.

Once these two aspect parameters (pose and scale) have been estimated, we can propose a prototypical model of the target object for a given context. In the case of a view-based object representation, the model of the object will consist of a collection of templates that correspond to the possible aspects of the target. As illustrated in Fig. 7, the model provides samples of the expected appearance of the object when it is embedded in a scene with similar contextual features. These views correspond to the distribution of local image features (here $\mathbf{v}_L$ correspond to pixel intensities), $P(\mathbf{v}_L|O, \mathbf{v}_C)$.

D. Contextual Modulation of Local Image Saliency

In this subsection we illustrate how the model for contextual priming based on global scene features can be used to modulate local image saliency. In the framework presented in Subsection 2.B (Eq. 2), saliency is defined as $S(\mathbf{x}) = P(\mathbf{v}_L|\mathbf{v}_C)^{-1}$. That is, saliency is large for local features that are unusual in the image.

When the target object is indeed salient in the image, then saliency maps provide an efficient shortcut for object detection. However, in general, the object of interest will not be the most salient object in the image. The inclusion of contextual priming provides an efficient mechanism for concentrating fixation points only in the image region that is relevant for the task. This is very important when the target object is not the most salient element of the scene. Context provides a way of shadowing salient image regions that are not relevant for the task. As described in Eqs. (2) and (3), contextual information modulates local image saliency as

$$S_c(\mathbf{x}) = S(\mathbf{x}) P(\mathbf{x}|o, \mathbf{v}_C) P(o|\mathbf{v}_C), \qquad (5)$$

where $S_c(\mathbf{x})$ is the local image saliency modulated by context and task demands. Note that if the object is inconsistent with the scene ($P(o|\mathbf{v}_C) \simeq 0$), then the system does not need to search for the object.

Figure 8 shows an image (a) and the local saliency (b). Figure 8(c) shows the image region that is relevant for the task of looking for pedestrians. In the saliency model, the image is explored according to the most salient features. When task and context information are included,

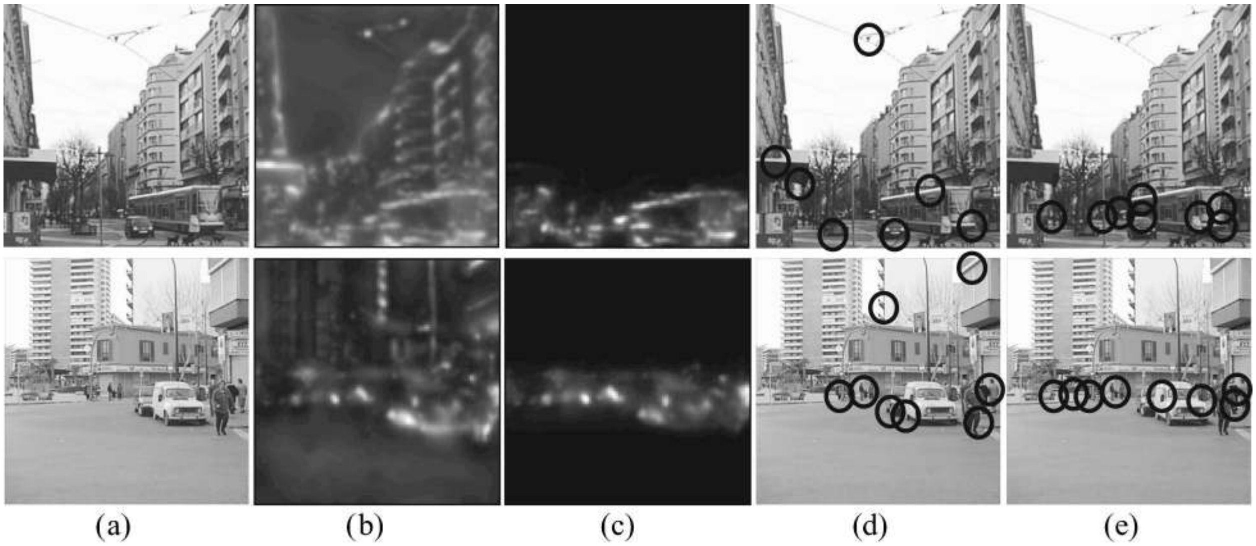

Fig. 8. (a) Input image (color is not taken into account). The task is to look for pedestrians. (b) Bottom-up saliency map, $S(\mathbf{x})$. (c) Context-driven focus of attention, $S_c(\mathbf{x})$. The image region in the shadow is not relevant for the task, and saliency is suppressed. (d) Points that correspond to the largest salience $S(\mathbf{x})$. (e) Image regions with the largest salience, including contextual priming, $S_c(\mathbf{x})$.

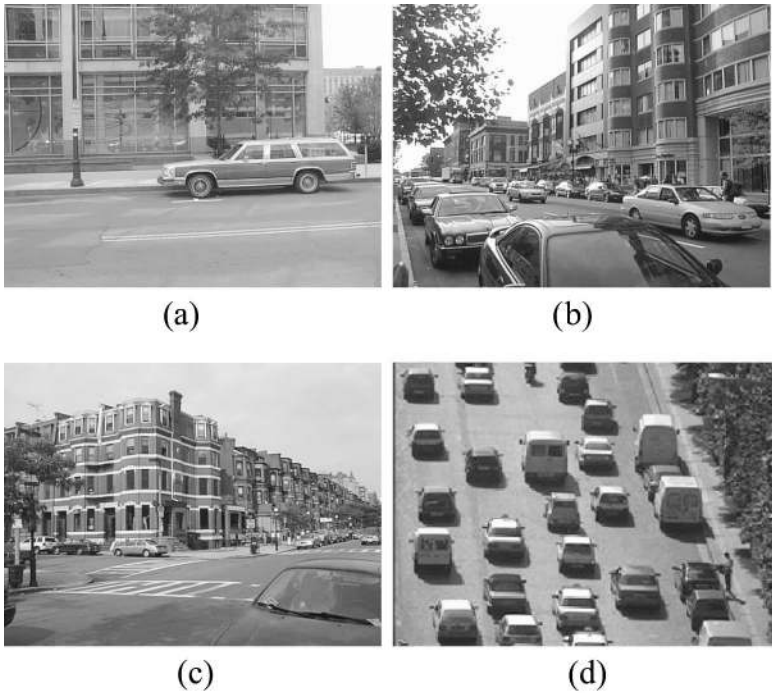

Fig. 9. The strength of contextual features in providing priors for objects depends on two factors: (1) how well the contextual features differentiate between different scenes and (2) how strong the relationship is between the object of interest and the scene.

saliency is shadowed outside the selected image region. Figures 8(d) and 8(e) show the selected image regions that correspond to maximum points of image saliency. Including contextual information allows the concentration of attention in the image regions that are relevant for the task (finding pedestrians).

4. DISCUSSION

The dominant framework in computational models of object recognition consists in modeling the probability of presence of an object using only $P(O|\mathbf{v}_L)$ and ignoring contextual priors.[7,36,42–46] In object-detection approaches based on local features, the background is a distracting element. The best performance in detection is obtained when the object is against an unstructured background.[48] However, when background is structured and there is correlation between object and context, a system that makes use of context could have higher performance and be more robust to image degradation (see Fig. 1).

In this paper the saliency of a target in an image is given by the more complete object probability function[29,40] $P(O|\mathbf{v}_L, \mathbf{v}_C)$. This allows for use of the correlations that exist between the background and the objects within the scene. There are a few studies that propose computational models of contextual influences in object recognition.[26,27] Common to these models is the use of object-centered representations in which the context is described as a collection of objects and a model of the joint distribution of objects in a reduced world. This approach requires object-centered mechanisms that provide candidate objects that are transformed into recognizable objects through analysis of the consistency with the other candidate objects in the scene. There is no attempt at recognizing the scene/context as a first stage. In this paper we have focused on encoding global scene configuration in order to introduce contextual information to drive the search for objects. This dispenses with the need to identify individual objects or regions within a scene.[22,23]

We use the global output distribution of a Gabor filter bank encoded at low resolution, in both the spatial and the spectral domains, as a representation of the scene/context. But there are many other global features that are relevant for encoding scene properties without encoding specific objects: texture histograms,[36] color histograms,[49] high-order statistics,[47] and blob arrangements.[23]

The strength of contextual features in providing priors for objects depends on two main factors: first, how well the contextual features differentiate between different scenes and second, how strong the relationship is between the object of interest and the scene. Some objects in reduced environments may be very poorly constrained by context. The strength of the contextual priors is a function of both the object property that is relevant for the task (e.g., the location, the scale) and the configurational regularities of the scene. For instance, in the case of face detection, the orientation of faces is very poorly related to the scene. However, other object properties such as scale or location have a strong relationship with the scene. Figure 9 illustrates how the strength of contextual priors for one object (e.g., cars) changes as a function of the scene: (a) scale, pose, and location are constrained; (b) pose and location are constrained but scale is unconstrained; (c) only location is constrained; (d) both scale and pose are constrained but location is unconstrained.

The results provided in this paper have focused on contextual priors $P(O|\mathbf{v}_C)$ and on saliency $P(\mathbf{v}_L|\mathbf{v}_C)^{-1}$. Both of these factors are interesting, because they provide information about the location of the target without using any information about the expected appearance of the target (e.g., cars have dominant horizontal structure, and pedestrians are vertical structures). We have ignored the factor $P(\mathbf{v}_L|O, \mathbf{v}_C)$ from Eq. (2), which is also known to have a large influence on object search.[6,33]

5. CONCLUSION

I have shown how simple holistic contextual features provide strong constraints for the identity, locations, and scales of objects in a scene. The proposed architecture for attention guidance consists of three parallel modules extracting different information: (1) bottom-up saliency, (2) object-centered features, and (3) contextual modulation of attention. The focus has been on showing how to introduce global scene factors in order to model the contextual modulation of local saliency. The proposed system learns the relationship between global scene features and local object properties (identity, location, and image scale).

The inclusion of scene factors in models of attention and also in computational approaches for object detection and recognition is essential for building reliable and efficient systems. Context information offers a way of cutting down the need for exhaustive search, thus providing a shortcut to object detection.

APPENDIX A: PROBABILITY DENSITY FUNCTION MODELING AND LEARNING

A. Image Saliency

In this model, saliency is computed from the distribution of features within the image.[50] We model the PDF $P(\mathbf{v}_L|\mathbf{v}_C)$ by using a mixture of Gaussians with N clusters,

$$P(\mathbf{v}_L|\mathbf{v}_C) = \sum_{i=1}^{N} b_i G(\mathbf{v}_L;\, \mu_i,\, \mathbf{X}_i), \tag{A1}$$

with

$$G(\mathbf{v}_L;\, \mu,\, \mathbf{X}) = \frac{\exp[-1/2(\mathbf{v}_L - \mu)^T \mathbf{X}^{-1}(\mathbf{v}_L - \mu)]}{(2\pi)^{N/2}|\mathbf{X}|^{1/2}}. \tag{A2}$$

For simplicity, we approximate the distribution $P(\mathbf{v}_L|\mathbf{v}_C)$ as the distribution of local features $\mathbf{v}_L$ in the input image (this approximation assumes that images with similar contextual features have similar distributions of local features). The parameters of the mixture of Gaussians (b_i, μ_i, and $\mathbf{X}$) are obtained by using the EM algorithm.[51–53] Given a set of N_t training samples (in this case these samples correspond to the set of all local feature vectors computed from one image), the EM algorithm is an iterative procedure:

E step

$$h_i^k(t) = \frac{b_i^k G(\mathbf{v}_t;\, \mu_i^k,\, \mathbf{X}_i^k)}{\sum_{i=1}^{L} b_i^k G(\mathbf{v}_t;\, \mu_i^k,\, \mathbf{X}_i^k)}, \tag{A3}$$

M step

$$b_i^{k+1} = \frac{\sum_{t=1}^{N_t} h_i^k(t)}{\sum_{i=1}^{L}\sum_{t=1}^{N_t} h_i^k(t)}, \tag{A4}$$

$$\mu_i^{k+1} = \frac{\sum_{t=1}^{N_t} h_i^k(t)\mathbf{v}_t}{\sum_{t=1}^{N_t} h_i^k(t)}, \tag{A5}$$

$$\mathbf{X}_i^{k+1} = \frac{\sum_{t=1}^{N_t} h_i^k(t)(\mathbf{v}_t - \mu_i^{k+1})(\mathbf{v}_t - \mu_i^{k+1})^T}{\sum_{t=1}^{N_t} h_i^k(t)}. \tag{A6}$$

Once the learning is completed (usually there is no improvement after ten iterations), we can evaluate Eq. (A1) at each location.

B. Contextual Priors

For the experiments presented in this paper, a database of 2700 annotated images was used. Pictures were 256×256 pixels in size. Images were transformed into gray scale, as color was not included in this study. Each image was annotated by indicating the categories of the objects present in the scene and their locations and sizes in pixels.

The contextual prior model requires the learning of the PDFs: $P(o|\mathbf{v}_C)$, $P(\mathbf{x}|o,\, \mathbf{v}_C)$, and $P(\mathbf{t}|\mathbf{x},\, \mathbf{v}_C,\, o)$. Again a mixture of Gaussians is used to model each PDF. Half of the database was used for the learning stage and the other half for the test.

The learning of the $P(o|\mathbf{v}_C) = P(\mathbf{v}_C|o)P(o)/P(\mathbf{v}_C)$ with $P(\mathbf{v}_C) = P(\mathbf{v}_C|o)P(o) + P(\mathbf{v}_C|\neg o)P(\neg o)$, where $\neg o$ denotes object absent, is done by approximating the in-class and out-of-class PDFs by a mixture of Gaussians. For the in-class PDF we use

$$P(\mathbf{v}_C|o) = \sum_{i=1}^{N} b_i G(\mathbf{v}_C;\, \mathbf{a}_i,\, \mathbf{A}_i), \tag{A7}$$

where $G(\mathbf{v}_C;\, \mathbf{a}_i,\, \mathbf{A}_i)$ is a multivariate Gaussian function of $\mathbf{v}_C$ with center $\mathbf{a}_i$ and covariance matrix $\mathbf{A}_i$; N is the number of Gaussians used for the approximation. The model parameters are obtained with the EM algorithm.[30,53] The same scheme holds for $P(\mathbf{v}_C|\neg o)$. The probability of the object presence $P(o)$ is approximated by the frequency presence of the object class. In our database we use $P(o) = 0.5$ for evaluating model performances. In our experiments we found that the learning requires the use of a few Gaussian clusters for modeling the PDFs (the results summarized in Subsection 3.A were obtained with $N = 2$). The learning is performed by using the half of the database; the remaining half is used for the testing stage.

The learning of the PDF $P(\mathbf{x}|o,\, \mathbf{v}_C)$ provides the relationship between the context and the more typical locations of the objects belonging to one class. The images used for the training of the PDF $P(\mathbf{x},\, \mathbf{v}_C|o)$ are a random selection of pictures among the ones that contain the object o. For each image we know the location of the object of interest, and we also compute the contextual features $\mathbf{v}_C$. The PDF learns the joint distribution between contextual features and the location of the target. For modeling the PDF we use a mixture of Gaussians:

$$P(\mathbf{x},\, \mathbf{v}_C|o) = \sum_{i=1}^{N} b_i G(\mathbf{x};\, \mathbf{x}_i,\, \mathbf{X}_i) G(\mathbf{v}_C;\, \mathbf{v}_i,\, \mathbf{V}_i). \tag{A8}$$

The joint PDF is modeled as a sum of N Gaussian clusters. Each cluster is decomposed into the product of two Gaussians. The first Gaussian models the distribution of object locations, and the second Gaussian models the distribution of contextual features for each cluster. The center of the Gaussian distribution of object locations is written as having a linear dependency with respect to the contextual features for each cluster[52]: $\mathbf{x}_i = \mathbf{a}_i + \mathbf{A}_i(\mathbf{v}_C - \mathbf{v}_i)$. The learning is performed with the EM algorithm.[52] The performances shown in Subsection 3.B (Fig. 5) were obtained with $N = 4$ clusters. Learning for scale and pose priming follows a similar strategy.

ACKNOWLEDGMENTS

The author thanks Aude Oliva, Whitman Richards, and William T. Freeman for fruitful discussions and Pawan Sinha and Gadi Geiger for comments.

The author's e-mail address is torralba@ai.mit.edu.

REFERENCES

1. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1254–1259 (1998).
2. T. Lindeberg, "Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention," Int. J. Comput. Vision **11**, 283–318 (1993).
3. A. Treisman and G. Gelade, "A feature integration theory of attention," Cogn. Psychol. **12**, 97–136 (1980).
4. A. Shashua and S. Ullman, "Structural saliency: the detection of globally salient structures using a locally connected network," in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE Computer Society Press, Los Alamitos, Calif., 1988), pp. 321–327.
5. J. M. Wolfe, "Guided search 2.0. A revised model of visual search," Psychon. Bull. Rev. **1**, 202–228 (1994).
6. R. P. N. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard, "Modeling saccadic targeting in visual search," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds. (MIT Press, Cambridge, Mass., 1996), Vol. 8, pp. 830–836.
7. B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," IEEE Trans. Pattern Anal. Mach. Intell. **19**, 696–710 (1997).
8. A. L. Yarbus, *Eye Movements and Vision* (Plenum, New York, 1967).
9. I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: detecting and judging objects undergoing relational violations," Cogn. Psychol. **14**, 143–177 (1982).
10. S. E. Palmer, "The effects of contextual scenes on the identification of objects," Memory Cognit. **3**, 519–526 (1975).
11. R. A. Rensink, J. K. O'Regan, and J. J. Clark, "To see or not to see: the need for attention to perceive changes in scenes," Psychol. Sci. **8**, 368–373 (1997).
12. R. A. Rensink, "The dynamic representation of scenes," Visual Cogn. **7**, 17–42 (2000).
13. J. M. Henderson and A. Hollingworth, "High-level scene perception," Annu. Rev. Psychol. **50**, 243–271 (1999).
14. P. De Graef, D. Christiaens, and G. d'Ydewalle, "Perceptual effects of scene context on object identification," Psychol. Res. **52**, 317–329 (1990).
15. M. M. Chun and Y. Jiang, "Contextual cueing: implicit learning and memory of visual context guides spatial attention," Cogn. Psychol. **36**, 28–71 (1998).
16. H. Arsenio, A. Oliva, and J. M. Wolfe, "Exorcising 'ghosts' in repeated visual search," J. Vision **2**, 733a (2002).
17. P. G. Schyns and A. Oliva, "From blobs to boundary edges: evidence for time and spatial scale dependent scene recognition," Psychol. Sci. **5**, 195–200 (1994).
18. S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," Nature **381**, 520–522 (1996).
19. M. C. Potter and E. I. Levy, "Recognition memory for a rapid sequence of pictures," J. Exp. Psychol. **81**, 10–15 (1969).
20. M. C. Potter, "Meaning in visual search," Science **187**, 965–966 (1975).
21. T. Sanocki and W. Epstein, "Priming spatial layout of scenes," Psychol. Sci. **8**, 374–378 (1997).
22. A. Oliva and P. G. Schyns, "Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli," Cogn. Psychol. **34**, 72–107 (1997).
23. A. Oliva and P. G. Schyns, "Diagnostic color blobs mediate scene recognition," Cogn. Psychol. **41**, 176–210 (2000).
24. A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," Int. J. Comput. Vision **42**, 145–175 (2001).
25. D. Noton and L. W. Stark, "Scanpaths in eye movements during pattern perception," Science **171**, 308–311 (1971).
26. D. A. Chernyak and L. W. Stark, "Top-down guided eye movements," IEEE Trans. Syst. Man Cybern. **31**, 514–522 (2001).
27. T. M. Strat and M. A. Fischler, "Context-based vision: recognizing objects using information from both 2-D and 3-D imagery," IEEE Trans. Pattern Anal. Mach. Intell. **13**, 1050–1065 (1991).
28. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, "Modeling visual-attention via selective tuning," Artif. Intell. **78**, 507–545 (1995).
29. A. Torralba and P. Sinha, "Statistical context priming for object detection: scale selection and focus of attention," in *Proceedings of the International Conference on Computer Vision* (IEEE Computer Society Press, Los Alamitos, Calif., 2001), Vol. 1, pp. 763–770.
30. A. Torralba, "Contextual modulation of target saliency," in *Advances in Neural Information Processing Systems*, T. G. Dietterich, S. Becker, and Z. Ghahramani, eds. (MIT Press, Cambridge, Mass., 2002), Vol. 14, pp. 1303–1310.
31. C. Koch and S. Ullman, "Shifts in visual attention: towards the underlying circuitry," Hum. Neurobiol. **4**, 219–227 (1985).
32. D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," Vision Res. **42**, 107–123 (2002).
33. J. M. Wolfe, "Visual search," in *Attention*, H. Pashler, ed. (University College London Press, London, 1998).
34. D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," J. Opt. Soc. Am. A **4**, 2379–2394 (1987).
35. B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature **381**, 607–609 (1996).
36. B. Schiele and J. L. Crowley, "Recognition without correspondence using multidimensional receptive field histograms," Int. J. Comput. Vision **36**, 31–50 (2000).
37. C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Region-based image querying," in *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Li-*

braries (IEEE Computer Society Press, Los Alamitos, Calif., 1997), pp. 42–49.

38. M. M. Gorkani and R. W. Picard, "Texture orientation for sorting photos at a glance," in *Proceedings of the IEEE International Conference on Pattern Recognition* (IEEE Computer Society Press, Los Alamitos, Calif., 1994), Vol. 1, pp. 459–464.
39. M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *Proceedings of the IEEE International Workshop on Content-Based Access of Image and Video Databases* (IEEE Computer Society Press, Los Alamitos, Calif., 1998), pp. 42–51.
40. A. Jepson, W. Richards, and D. Knill, "Modal structures and reliable inference," in *Perception as Bayesian Inference*, D. Knill and W. Richards eds. (Cambridge U. Press, Cambridge, UK, 1996), pp. 63–92.
41. A. Treisman, "Properties, parts and objects," in *Handbook of Human Perception and Performance*, K. R. Boff, L. Kaufman, and J. P. Thomas, eds. (Wiley, New York, 1986), pp. 35.1–35.70.
42. B. Heisele, T. Serre, S. Mukherjee, and T. Poggio, "Feature reduction and hierarchy of classifiers for fast object detection in video images," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society Press, Los Alamitos, Calif., 2001), Vol. 2, pp. 18–24.
43. P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society Press, Los Alamitos, Calif., 2001), Vol. 1, pp. 511–518.
44. S. Ullman, M. Vidal-Naquet, and E. Sali, "Visual features of intermediate complexity and their use in classification," Nat. Neurosci. **5**, 682–687 (2002).
45. M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," Nat. Neurosci. **2**, 1019–1025 (1999).
46. S. Edelman, "Computational theories of object recognition," Trends Cogn. Sci. **1**, 296–304 (1997).
47. A. Torralba and A. Oliva, "Depth perception from familiar structure," IEEE Trans. Pattern Anal. Mach. Intell. **24**, 1226–1238 (2002).
48. M. P. Eckstein and J. S. Whiting, "Visual signal detection in structured backgrounds I. Effect of number of possible spatial locations and signal contrast," J. Opt. Soc. Am. **A13**, 1777–1787 (1996).
49. M. Swain and D. Ballard, "Color indexing," Int. J. Comput. Vision **7**, 11–32 (1991).
50. R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," Vision Res. **39**, 3157–3163 (1999).
51. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. R. Stat. Soc. Ser. B. **39**, 1–38 (1977).
52. N. Gershenfeld, *The Nature of Mathematical Modeling* (Cambridge U. Press, Cambridge, UK, 1999).
53. M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," Neural Comput. **6**, 181–214 (1994).