Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2004-020
AIM-2004-009

April 14, 2004

Contextual Influences on Saliency

Antonio Torralba

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

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 short cut 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. Scene information can then be used to modulate the saliency of image regions early during the visual processing in order to provide an efficient short cut for object detection and recognition. [1].

[1] This work was sponsored in part by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement

1 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 (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) showed that visual search is facilitated when there exists a correlation 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.).

Here it is described a scheme 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 short cut 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.

2 The scene context

In figure 1.a, 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 degrees 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, the context of the target object is considered as

a) effects of scene context on object recognition.

b) effects of scene context on object search.

Fig. 1. a) when object intrinsic information is reduced, then context plays a mayor 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.

a collection of distractors. Fig. 1.b(left), shows a display with a salient target, where context (distractors) is not affecting target processing (Treisman and Gelade, 1983). In the central image, a person is embedded in the background. The person is masked by the context and it is difficult to find. In these two examples, context is non-informative and its only effect on the search is due to the ability of the background to mask the target. But context can also provide information about the presence of the target (fig 2.b-right). In fig 2.b-right the context, instead of masking the person, it provides priors about what are the expected locations and scales in which we can find the target. In the canyon scene, a person can be almost in any location. However, in the street scene, the environment imposes strong constraints about what are the typical locations in which people is expected to be. This use of context is the one we are interested in modeling here.

3 The representation of scenes

One 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. 2, the analysis of the image is perform using two parallel pathways: a local (e.g., objects) and a global pathway (e.g., scenes). Here we describe the features that can be used in both pathways.

3.1 Local features

Most models of attention and object recognition rely on the definition of sets of local features. In the local pathway, each location is represented by a vector of features that describe 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. 2, each local feature vector is a jet of filter responses: $\mathbf{v}_l(x) = \{g_1(x), g_2(x), ..., g_N(x)\}$. 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.

3.2 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 shown in Fig. 2, the global feature shown responds to a combination of the output of oriented filters at different image locations (Oliva and Torralba, 2001): $\mathbf{v}_c = \{\sum_x \sum_k g_k(x)\phi_m(x,k);\ m = 1, M\}$, 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 shown in Fig. 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 would correspond to the typical structure of a street scene). The global image representation is build by a collection of such kind of features.

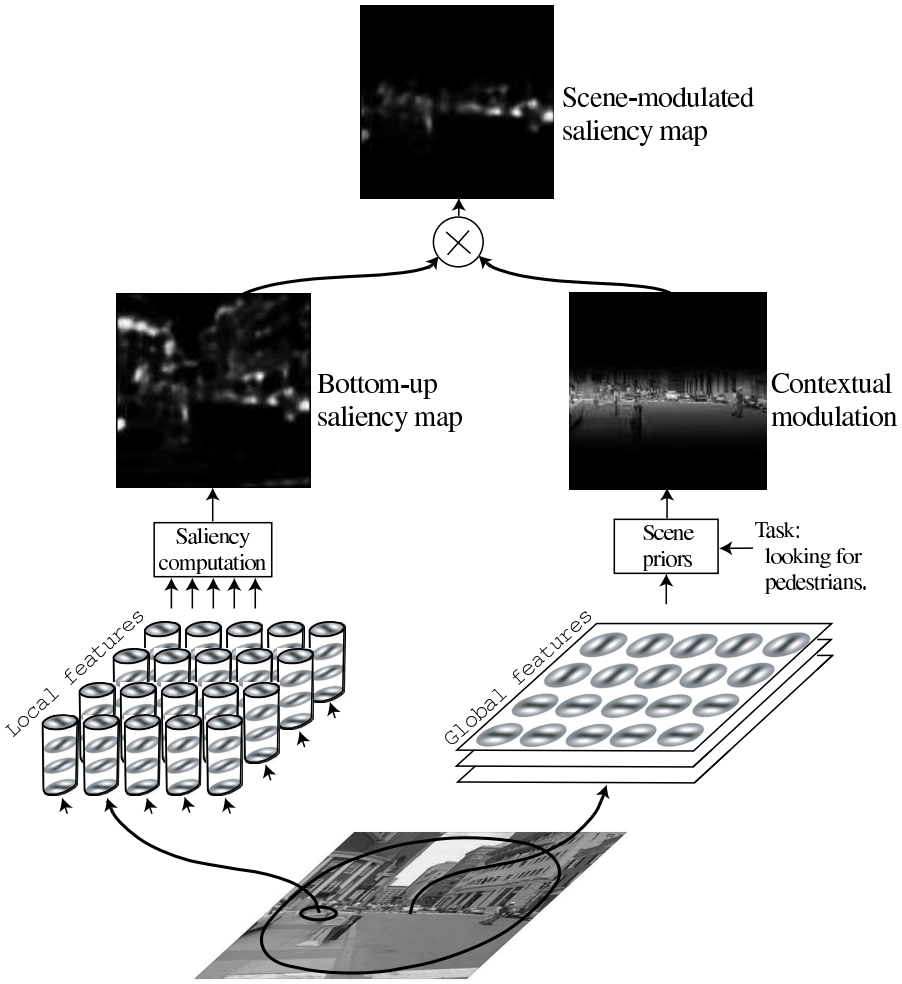

Fig. 2. Contextual and local pathways for pre-attentive object search. This scheme incorporates contextual information to modulate image saliency. The scheme consists in two parallel pathways, the first one processes local image information, the second one, encodes globally the pattern of activation of the feature maps. When looking for a person in the image, the saliency map, which is task independent, 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). The final attentional map, obtained as the product of both maps, will select the salient locations in side the image region relevant for the task.

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

4 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) it is assigned a probability of containing the target: $p(o,\, x,\, \sigma,\, \alpha|\, \mathbf{v}_l,\, \mathbf{v}_c)$. t is a vector a parameters that 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:

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

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

4.1 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 $S(x) = 1/p(\mathbf{v}_l(x)\,|\,\mathbf{v}_c)$. 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 multivariate power-exponential function to the distribution of local features in the image.

4.2 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 (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. Note that when the object properties O fully constraint the object appearance, then it is possible to approximate $p(\mathbf{v}_l\,|\,O,\, \mathbf{v}_c) \simeq p(\mathbf{v}_l\,|\,O)$. This

approximation allows dissociating the contribution of local image features and global (contextual) image features.

4.3 Scene priors

The third factor, the PDF $p(O \mid \mathbf{v}_c)$, provides context-based priors on object class, location, scale and appearance. This term is of capital importance for insuring reliable inferences in situations where 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 $p(O \mid \mathbf{v}_c)$ into several factors that model different kinds of scene priors for object search:

$$p(O \mid \mathbf{v}_c) = p(\alpha \mid x, \mathbf{v}_c, o)p(\sigma \mid x, \mathbf{v}_c, o)p(x \mid \mathbf{v}_c, o)P(o \mid \mathbf{v}_C) \tag{2}$$

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

- *Object-class priming*: $P(o \mid \mathbf{v}_c)$ provides the probability of presence of the object class o in the scene. If $P(o \mid \mathbf{v}_c)$ is very small, then object search need not be initiated.
- *Contextual control of focus of attention*: $p(x \mid o, \mathbf{v}_c)$. This PDF gives the most likely locations for the presence of object o given context information.
- *Contextual selection of scale*: $p(\sigma \mid x, o, \mathbf{v}_c)$. 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*: $p(\alpha \mid x, o, \mathbf{v}_c)$. This gives the expected shapes (point of views, aspect ratio) of the object.

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

5 Results

Fig. 3.a shows the effect that contextual priors $p(O \mid \mathbf{v}_c)$ have on subject performances for recognition. First, subjects are asked to guess the identity of the objects behind the masks (Fig. 3.a-top). That experiment allows us to evaluate the distribution of objects that subjects are considering for each scene: $P(objects|x, \sigma, scene)$. 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 placed in consistent and inconsistent backgrounds. The results (Fig. 3.a-right) show that observer's performance on a recognition task is correlated with the strength of the priors (1).

Fig. 3.b 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 $p(O \mid \mathbf{v}_c)$ contain the information about how the scene features $\mathbf{v}_c$ were related to the target properties O (image location, scale, pose) during the past experience. The system is trained by first providing to the system a collection of images in which the target has been already located. The PDF is learnt using a mixture of gaussians and the EM algorithm (Torralba, 2003b). Once the system has learnt the relationship between scenes and objects, it can predict the expected locations for several objects in new scenes (fig. 3.b-right).

Fig. 3.b provides examples of the results of global contextual priming for: (b1) predicting the presence/absence of objects. Here we show the results for solving the task of animal present/absent using only scene priors, before scanning the image (Torralba and Oliva, 2003). The system has 80% correct prediction rate on this task. (b2) Focus of attention (e.g., expected locations of people and trees). Contextual priors for location reduce that area of the image that needs to be explored when looking for the target. And (b3) scale selection (e.g., expected size of face in the image).

Finally, Fig. 3.c compares the salient points and the region of interest predicted by a model using only bottom-up saliency maps (fig. 3.c-center) and when combining saliency and the scene priors for location (fig. 3.c-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

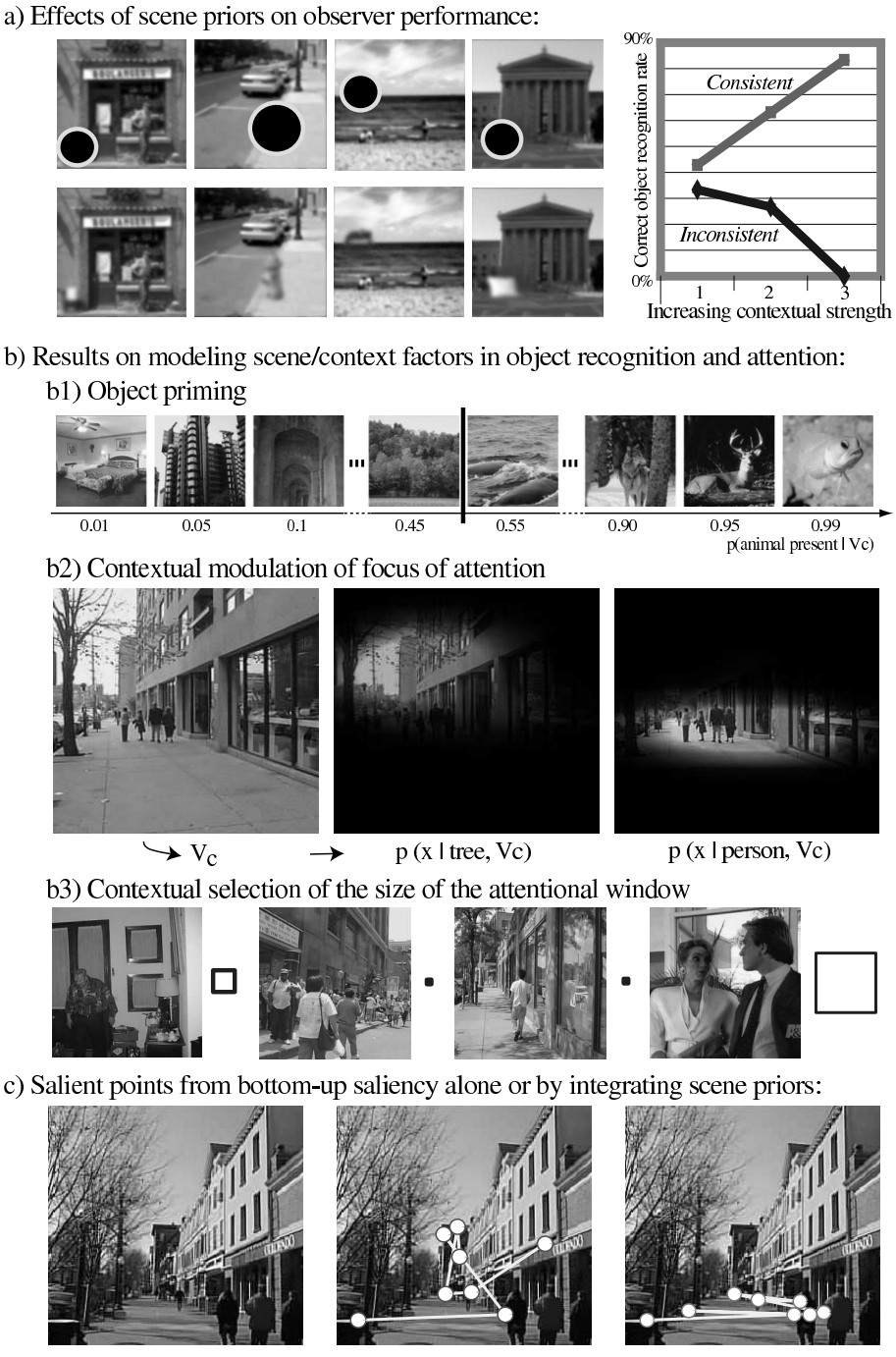

Fig. 3. a) Role of contextual priors on object recognition by subjects. b) Scene priors obtained from global image features. c) Examples of salient locations obtained from saliency alone (center) and combining both context and saliency (right). Including scene priors provides better candidates for the location of the target.

including scene priors provide better predictions of human eye movements than saliency alone (Oliva et al., 2003).

6 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 presence/absence before scanning the image.

From an algorithmic point of view, contextual control of the focus of attention is important as it avoids expending computational resources in spatial locations with 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

[1] Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, 15, 600-609.

[2] Biederman, I., Mezzanotte, R.J., and Rabinowitz, J.C. 1982. Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14:143–177.

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

[4] Henderson, J.M., and Hollingworth, A. 1999. High level scene perception. *Annual Review of Psychology*, 50, 243-271.

[5] 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):1254.

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

[7] Oliva, A., and Torralba, A. Modeling the shape of the scene: a holistic representation of the spatial Envelope. *Int. Journal of Computer Vision*, 42(3), pp. 145–175, 2001.

[8] Oliva, A., Torralba, A., Castelhano, M. S., Henderson, J. M. 2003. Top-Down Control of Visual Attention in Object Detection *IEEE proc. International Conference on Image Processing*, Vol. I, pp. 253–256. September 14-17, in Barcelona, Spain.

[9] Palmer, S. E. 1975. The effects of contextual scenes on the identification of objects. *Memory and Cognition* , 3:519–526.

[10] 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.

[11] 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. *Psychological Science*, 8:368-373

[12] 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.

[13] Torralba, A. 2003a. Modeling global scene factors in attention. *Journal of Optical Society of America A*, 20(7): 1407-1418.

[14] Torralba, A. 2003b. Contextual Priming for Object Detection. *Int. Journal of Computer Vision*, 53(2): 169-191.

[15] Torralba, A., Oliva, A. 2003. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14: 391-412.

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

[17] Treisman, A., and Gelade, G. 1980. A feature integration theory of attention. *Cognitive Psychology*, Vol. 12:97–136.

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