TOP-DOWN CONTROL OF VISUAL ATTENTION IN OBJECT DETECTION

Aude Oliva
Cognitive Science, MSU
East Lansing
MI 48824

Antonio Torralba
AI-Lab, MIT
Cambridge
MA 02139

Monica S.Castelhano
Cognitive Science, MSU
East Lansing
MI 48824

John M.Henderson
Cognitive Science, MSU
East Lansing
MI 48824

ABSTRACT

Current computational models of visual attention focus on bottom-up information and ignore scene context. However, studies in visual cognition show that humans use context to facilitate object detection in natural scenes by directing their attention or eyes to diagnostic regions. Here we propose a model of attention guidance based on global scene configuration. We show that the statistics of low-level features across the scene image determine where a specific object (e.g. a person) should be located. Human eye movements show that regions chosen by the top-down model agree with regions scrutinized by human observers performing a visual search task for people. The results validate the proposition that top-down information from visual context modulates the saliency of image regions during the task of object detection. Contextual information provides a shortcut for efficient object detection systems.

1. INTRODUCTION

While looking for a specific object in a complex and cluttered scene, human observers use visual context information to facilitate the search, by directing their attention or eyes to relevant regions in the image (e.g. in the street when searching for cars, on a table searching for a plate). This strategy is not considered by current computational models of visual attention [3, 7], which focus on the saliency zones of the image, independently of the meaning of the scene.

In this paper, we describe a computational model of attention guidance that takes into account the visual context (e.g. the scene) in which objects are embedded [10, 11]. We show that the statistics of low-level features across a natural scene is strongly correlated with the location of a specific object. In the current study, the scheme is tested with the task of locating probable locations of people in scenes, and these selected regions are compared to human eye movement scan patterns.

Models that integrate attention mechanisms are relevant for computer vision as they can suggest strategies for finding shortcuts for object detection and recognition. These shortcuts can be used to select a set of possible candidate locations of target objects within an image. Then, computationally more expensive object recognition procedures can be applied in those regions [5]. In this paper, we propose a simple attentional mechanism that does not use specific information about the appearance of the target. Instead we use a simple model of image saliency based on the distribution of local features in the image and a model of contextual priors (that learns the relationship between context features and the location of the target during past experience) in order to select interesting regions of the image. The paper shows that there could exist pre-attentive heuristics based on the context within which an object is embedded, that would provide a low-cost object detection shortcut by pre-selecting relevant image regions.

2. SALIENCY AND OBJECT DETECTION

For bottom-up models of attention allocation, regions with different properties from their neighboring regions are considered more informative and are supposed to attract attention. Those models provide a measure of the 'saliency' of each location in the image across various low-level features (contrast, color, orientation, texture, motion, [3, 13]). Saliency measures are interesting in the framework of object detection because, when looking for a target object, frequent features in the image are more likely to belong to the background and, therefore, are poor predictors of the presence of the target.

In saliency models, a saliency map is computed using a hardwire scheme (e.g., [3]): the local image features are processed by center-surround inhibition and then a winner take all strategy is used to select the most salient regions. The image features most commonly used for describing local image structure (orientation, scale and texture) are the outputs of multiscale oriented band-pass filters. Here, we decompose each color subband using a steerable pyramide [9] with 4 scales and 4 orientations (fig. 1). Each location has a features vector $\mathbf{v}_l(\mathbf{x}) = \{v_l(\mathbf{x}, k)\}_{k=1,48}$ with 48 dimensions (fig. 1).

Here, we define the saliency in terms of the likelihood

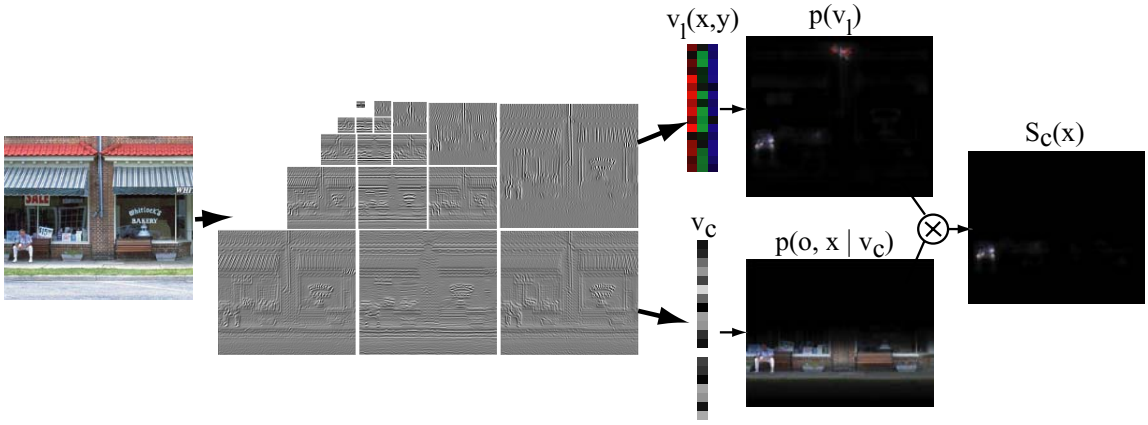

Fig. 1. Attentional system for object detection integrating local saliency and contextual priors about target location.

of finding a set of local features in the image. We use a probabilistic definition of saliency that more naturally fits with object detection and recognition formulations we later show :

$$S(\mathbf{x}) = p(\mathbf{v}_l)^{-1} \tag{1}$$

In this definition, the saliency of a location is large when the image features at that location are more unexpected in the image. We approximate this probability by fitting a gaussian to the distribution of local features in the image ([8]):

$$p(\mathbf{v}_l) = \frac{e^{-1/2\,(\mathbf{v}_l-\mu)^T X^{-1}(\mathbf{v}_l-\mu)}}{(2\pi)^{N/2}|X|^{1/2}} \tag{2}$$

Although a mixture of gaussians produces a better fit of the distribution, it did not significantly change the selected salient points. As discussed later (fig. 2), the accuracy of this model in predicting the fixated points by human subjects did not differ with the performance of a more complex model of saliency maps [3].

3. CONTEXTUAL OBJECT PRIMING

However, when looking for an object, the use of saliency $S(\mathbf{x})$ as defined in eq. (1) is insufficient for explaining human performance or for building interesting object detection procedures. During the first glance at a scene (or 200 msec), the attention of the observer is driven towards a region in the image and the first saccade is programmed. This process is task-dependent. When subjects are asked to search for a specific target object, that object is fixated (and so located) faster when it is consistent with the scene context than when it is inconsistent [2]. Human observers are clearly using a top-down mechanism to find regions of interest where an object should be located, independent of the presence of the physical features of the object [2, 1].

3.1. Contextual modulation of saliency

The role of the visual context is to provide information about past search experiences in similar environments and strategies that were successful in finding the target. When using a statistical framework, object detection is formulated as the evaluation of the probability function $p(o|\,\mathbf{v}_l)$. This is the probability of the presence of the object o given a set of local measurements. As suggested in [11] a more robust approach should include contextual information. We can write the probability of presence of object o at the location $\mathbf{x}$ as:

$$p(o, \mathbf{x} \,|\, \mathbf{v}_l, \mathbf{v}_c) = \frac{p(\mathbf{v}_l \,|\, o, \mathbf{x}, \mathbf{v}_c)}{p(\mathbf{v}_l \,|\, \mathbf{v}_c)} p(o, \mathbf{x} \,|\, \mathbf{v}_c) \tag{3}$$

where $\mathbf{v}_c$ is the vector of contextual features (see next section). Using Bayes rule, the probability can be decomposed into three factors[11]: the object likelihood, ($p(\mathbf{v}_l \,|\, o, \mathbf{x}, \mathbf{v}_c)$), the local saliency $p(\mathbf{v}_l \,|\, \mathbf{v}_c)$ and the contextual priors $p(o, \mathbf{x} \,|\, \mathbf{v}_c)$. We are interested in the terms that do not require knowledge of the appearance of the target:

$$S_c(\mathbf{x}) = \frac{p(o, \mathbf{x} \,|\, \mathbf{v}_c)}{p(\mathbf{v}_l \,|\, \mathbf{v}_c)} = S(\mathbf{x}) p(o, \mathbf{x} \,|\, \mathbf{v}_c) \tag{4}$$

The interest of this term is that it avoids using a specific model of the appearance of the target. The term $S_c(\mathbf{x})$ does not incorporate any information about the distribution of features that belong to the target. Therefore, it can be computed efficiently, and does not depend on the featural complexity of the target. The term $S(\mathbf{x}) = p(\mathbf{v}_l \,|\, \mathbf{v}_c)^{-1}$ (local saliency) provides a measure of how unlikely it is to find a set of local measurements within the context $\mathbf{v}_c$. It can be approximated by the distribution of local features within the image as in eq. (2). A term of saliency, eq. (1), appears naturally in a statistical framework for object detection.

Here, we are ignoring the term $p(\mathbf{v}_l \,|\, o, \mathbf{x}, \mathbf{v}_c)$. However, experimental work [13] has shown that simple features

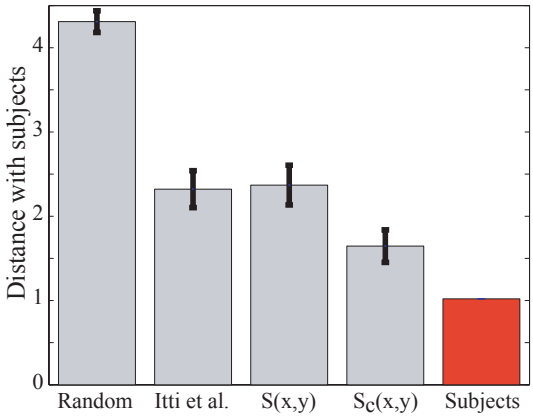

Fig. 2. Mean distance values of scan pattern similarity between human subjects and each of the different conditions.

of the target are also used for guiding attention (for instance when looking for a red spot, attention can be directed to red regions in the image).

3.2. Computing contextual image features

Contextual features have to describe the structure of the whole image [6]. The representation has to be low dimensional so that the PDF $p(o, \mathbf{x} \mid \mathbf{v}_c)$ can be learnt efficiently and has to keep relevant information about the scene so that the PDF can provide strong priors about the location of the target. There are many possible representations of contextual information such as collecting global image statistics, color histograms, wavelet histograms, etc. Following [6], here we represent the context by reducing the dimensionality of the local features $v_l(\mathbf{x}, k)$. First we take the absolute value to remove variability due to contrast. Then we subsample each subband by a factor M:

$$v(\mathbf{x}, k) = \left\{ |v_l(\mathbf{x}, k)|^2 \downarrow M \right\} \tag{5}$$

We further reduce the dimensionality by decomposing the image features $v(\mathbf{x}, k)$ into the basis functions provided by the principal component analysis:

$$a_n = \sum_{\mathbf{x}} \sum_{k} |v(\mathbf{x}, k)| \; \psi_n(\mathbf{x}, k) \tag{6}$$

We propose to use the decomposition coefficients $\mathbf{v}_C = \{a_n\}_{n=1,L}$, with $L = 60$, as context features. The functions ψ_n are the eigenfunctions of the covariance matrix defined by the image features $v(\mathbf{x}, k)$. The resolution reduction of eq. (5) allows for the PCA to be computed more efficiently. We perform the PCA on more than 3000 natural images.

3.3. Learning the location of people

The role of the visual context factor in modulating attention is to provide information about past search experience in similar environments and the strategies that were successful in finding the target. The learning is performed by training the PDF $p(o, \mathbf{x} \mid \mathbf{v}_c)$ using a database of images for which the location of the target is known. For the results shown in this paper we train the PDF to predict the location of people. For each image, the features $\mathbf{v}_c$ and the location $\mathbf{x}$ of the target ($o = people$) are known. We model the PDF using a mixture of gaussians and the learning is performed using the EM algorithm (see [11] for details). Fig. 1 shows an example of PDF $p(o, \mathbf{x} \mid \mathbf{v}_c)$.

4. HUMAN EYE MOVEMENTS

In this section we study how the system explores a set of 36 real-world scenes for which eye movements have been recorded for 8 subjects. In order to model the human eye scan paths, we comparedhuman fixation patterns to patterns derivedfrom a purely bottom-up approach (saliency) and patterns that included top-down information (contextual priming). None of the images used in the experiments were used during the training. First we describe the procedure for recording eye movements.

4.1. Apparatus and Procedure

The right eye was tracked using a Generation 5.5 Fourward Technologies Dual Purkinje Image Eyetracker. Digitized full-color photographs of 36 real-world scenes, taken from various sources, were displayed on a NEC Multisync P750 monitor (refresh rate = 143 Hz) at a viewing distance of 1.13 m, subtending 15.8 deg. x 11.9 deg. of visual angle. There were 14 scenes with no people and 22 scenes containing anywhere from 1 to 6 people. After the subject centered their fixation, a scene appeared and observers were instructed to count the number of people present. A scene was displayed until the subject's response or for a maximum of 10 s. The eyetracker was used to record the position and duration of fixations during the search.

4.2. Eye Movement Data Analysis

In order to analyze the eye movements data, we computed the squared difference between corresponding fixation points in two sets of fixations (see [4]): the eye movements pattern made for each image and subject, was compared with the eye movements pattern made by Itti et al. [3] saliency model, $S(\mathbf{x})$ saliency model, and $S_c(\mathbf{x})$ context control model (cf. Figs. 2 and 3). In addition, we compared the distance between subjects and a random pattern of fixations. Like in [4], the study is restricted to the first seven fixations. Figure 2 summarizes the results. Distance between patterns of fixations was normalized so that average distance within subjects was 1 (each subject compared with other subjects on the same image, for all images). The graph shows that

(a) Subject (b) Itti et al. (c) S(x) (d) S_c(x)

Fig. 3. Pattern of 7 first eye movements performed by (a) a human observer (b) Itti et al. saliency model, (c) $S(\mathbf{x})$ saliency model, (d) and saliency with inclusion of top down information, $S_c(\mathbf{x})$.

the pattern of fixations suggested by the contextual model resembled human eye movements the closest. Pure saliency models performed worse, but were still more similar to human data than a purely random process. There is no statistical difference between performances of Itti's model or the probabilistic definition of saliency (eq. 1).

5. SUMMARY AND CONCLUSIONS

We presented a computational model of attention guidance that integrates context information with image saliency to determine regions of interest. By comparing scan patterns of different models to those of human observers, we validate the proposition that top-down information from visual context modulates the saliency of regions during the task of object detection. Contextual information provides an essential shortcut for efficient object detection systems.

6. ACKNOWLEDGMENTS

This research was supported by the Nat. Sci. Foundation (BCS-0094433 and KDI award ECS-9873531), NIMH (1R03MH068322-1)and the Army Research Office (DAAD19-00-1-0519; the opinions expressed in this article are those of the authors and do not necessarily represent the views of the Department of the Army or any other governmental organization. Reference to or citations of trade or corporate names do not constitute explicit or implied endorsement of those entities or their products by the author or the Department of the Army). Authors may be contacted by emails. A.O: aoliva@msu.edu, A.T: torralba@ai.mit.edu, M.S.C: monica@eyelab.msu.edu, J.M.H: john@eyelab.msu.edu.

7. REFERENCES

[1] M.S. Castelhano, and J.M. Henderson, Flashing scenes and moving windows: An effect of initial scene gist on eye movements, *3rd Annual Meeting of the Vision Sciences Society*, Sarasota, Florida, 2003.

[2] J.M. Henderson, P.A. Weeks, and A. Hollingworth, Effects of semantic consistency on eye movements during scene viewing, *JEP: HPP*, 25, 210, 1999.

[3] L. Itti, C. Koch, and E. Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans. PAMI*, 20(11):1254, 1998.

[4] S. Mannan, K.H. Ruddock, and D.S. Wooding, Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images, *Spatial Vis.*, 9(3):363,1998.

[5] F. Miau, L. Itti, A Neural Model Combining Attentional Orienting to Object Recognition: Preliminary Explorations on the Interplay Between Where and What, *Proc. IEEE Engineering in Medicine and Biology Society (EMBS)*, Istanbul, Turkey, Oct 2001.

[6] A. Oliva, and A. Torralba. Modeling the shape of the scene, *IJCV*, 42(3), pp. 145–175, 2001.

[7] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention, *Vis. Res.*, 42, pp. 107–123, 2002.

[8] R. Rosenholtz, A simple saliency model predicts a number of motion popout phenomena, *Vis. Res.*, 39, pp. 3157–3163, 1999.

[9] E. P. Simoncelli and W. T. Freeman, The steerable pyramid: a flexible architecture for multi-scale derivative computation, In *2nd Annual Intl. Conf. on Im. Proc.*, Washington, DC, 1995.

[10] T.M. Strat, and M.A. Fischler, Context-based vision: recognizing objects using information from both 2-D and 3-D imagery, *IEEE trans. on PAMI*, 13(10): 1050-1065, 1991.

[11] A. Torralba, Contextual priming for object detection, *IJCV*, vol. 53, pp. 153167, 2003.

[12] A. Torralba, Modeling global scene factors in attention, *JOSA - A*, vol. 20, 7, 2003.

[13] J.M. Wolfe, Guided search 2.0. A revised model of visual search. *Psych. Bull. and Rev.*, 1:202-228, 1994.