

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



The role of context in object recognition

Aude Oliva¹ and Antonio Torralba²

¹Brain and Cognitive Sciences Department, MIT 46-4068, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

²Computer Science and Artificial Intelligence Laboratory, MIT 32-D462, 32 Vassar Street, Cambridge, MA 02139, USA

In the real world, objects never occur in isolation; they co-vary with other objects and particular environments, providing a rich source of contextual associations to be exploited by the visual system. A natural way of representing the context of an object is in terms of its relationship to other objects. Alternately, recent work has shown that a statistical summary of the scene provides a complementary and effective source of information for contextual inference, which enables humans to quickly guide their attention and eyes to regions of interest in natural scenes. A better understanding of how humans build such scene representations, and of the mechanisms of contextual analysis, will lead to a new generation of computer vision systems.

Introduction

The ability of humans to recognize thousands of object categories in cluttered scenes, despite variability in pose, changes in illumination and occlusions, is one of the most surprising capabilities of visual perception, still unmatched by computer vision algorithms. Object recognition is generally posed as the problem of matching a representation of the target object with the available image features, while rejecting the background features. In typical visual-search experiments, the context of a target is a random collection of distractors that serve only to make the detection process as hard as possible. However, in the real world, the other objects in a scene are a rich source of information that can serve to help rather than hinder the recognition and detection of objects. In this article, we review work on visual context and its role on object recognition.

Contextual influences on object recognition

In the real world, objects tend to co-vary with other objects and particular environments, providing a rich collection of contextual associations to be exploited by the visual system. A large body of evidence in the literature on visual cognition [1–8], computer vision [9–11] and cognitive neuroscience [12–16] has shown that contextual information affects the efficiency of the search and recognition of objects. There is a general consensus that objects appearing in a consistent or familiar background are detected more accurately and processed more quickly than objects appearing in an inconsistent scene.

The structure of many real-world scenes is governed by strong configural rules similar to those that apply to a single object. This is illustrated in [Figure 1](#). By averaging hundreds of images aligned on frontal faces, a common pattern of intensities emerges, showing a rigid organization of facial parts shared by all the members of the ‘face’ category. Average images aligned on a single object can reveal additional regions beyond the boundaries of the object that have a meaningful structure. For instance, a monitor and table emerge in the background of the average keyboard, despite the fact that the images were not constrained to contain those objects. The background of the fire hydrant is less distinct, but, because it must be supported on the ground, the average image reveals a ground plane. The presence of a particular object constrains the identity and location of nearby objects, and this property is probably used by the visual system.

Contextual influences on object recognition become evident if the local features are insufficient because the object is small, occluded or camouflaged. In the example shown in [Figure 2](#), the blobs corresponding to the car and pedestrian are ambiguous in isolation. However, the context of scene is so generous that it provides a distinct identity to each basic shape.

The effects of context

Early studies have shown that context has effects at multiple levels: semantic (e.g. a table and chair are probably present in the same images, whereas an elephant and a bed are not), spatial configuration (e.g. a keyboard is expected to be below a monitor), and pose (e.g. chairs are oriented towards the table, a pen should have a particular pose relative to the paper to be useful for writing and a car will be oriented along the driving directions of a street).

Hock *et al.* [17] and Biederman and collaborators [2] observed that both semantic (object presence, position and size) and physical (consistent support and interposition with other objects) object–scene relationships have an impact on the detection of a target object within the temporal window of a glance (<200 ms). These contextual relationships can have different strengths: a plate is expected to be on top of the table, but other locations are also possible, such as being on a shelf or wall; and a fire hydrant will always be on top of the sidewalk, not below the ground plane or floating in the air. Object recognition should be more accurate if the relationship between the context and the object is strong ([Figure 2](#)) and decrease as the strength of the object–scene relationship decreases. In

Corresponding authors: Oliva, A. (oliva@mit.edu); Torralba, A. (torralba@csail.mit.edu).



Figure 1. The structure of objects and their backgrounds. In this illustration, each image has been created by averaging hundreds of pictures containing a particular object in the center (a face, keyboard and fire hydrant) at a fixed scale and pose. Images come from the LabelMe dataset [65]. Before averaging, each picture is translated and scaled so that the target object is in the center. No other transformations are applied. The averages reveal the regularities existing in the intensity patterns across all the images. The background of many objects does not average to a uniform field, showing that an object extends its influence beyond its own boundaries, and this property is heavily used by the visual system.

recent work, these rules have been integrated into a common framework of contextual influences [11,18–20], in which context provides a robust estimate of the probability of an object's presence, position and scale.

The most documented effect of context on object recognition is the scene consistency–inconsistency effect [1,5,7,8]. Palmer [7] found that observers' accuracy at an object-categorization task was facilitated if the target (e.g. a loaf of bread) was presented after an appropriate scene (e.g. a kitchen counter) and impaired if the scene–object

pairing was inappropriate (e.g. a kitchen counter and bass drum). In a recent study, Davenport and Potter [3] observed that consistency information influences perception of both the object and the scene background if a scene is presented briefly (80 ms), which suggests a recurrent processing framework, in which objects and their settings influence each other mutually [21].

A natural way of representing the context of an object is in terms of its relationship to other objects (Figure 3). Learning statistical contingencies between objects can cause the perception of one object or scene to generate strong expectations about the probable presence and location of other objects (see Box 1 for a description of current theories about the mechanisms involved in contextual inference). Chun and Jiang [22] showed that people can learn the contingencies between novel objects, predicting the presence of one object on the basis of another, over the course of only 30 min. Contextual interactions between objects can be sensitive to subtle visual aspects. Green and Hummel [23] found that mechanisms of object perception are sensitive to the relative pose of pairs of objects. In a priming design, observers were presented with a prime object (e.g. a pitcher) for 50 ms, followed by a target image (e.g. a glass) for 50 ms. Crucially, the accuracy of target recognition was significantly higher if the prime object was oriented to interact with the target object in a consistent manner (e.g. a pitcher facing a glass) than if the pair interacted in an inconsistent manner (e.g. a pitcher oriented away from a glass).

In most of these studies, the participants did not learn any new contextual rules. All the experiments were designed to prove the effects of contextual rules previously learnt by the observers in the real world, and the tests were performed on realistic images or line drawings depicting real scenes. In the next section, we review recent work that shows that human observers have a remarkable ability to learn contextual associations. Observers do not need to be explicitly aware of contextual associations to benefit from them.

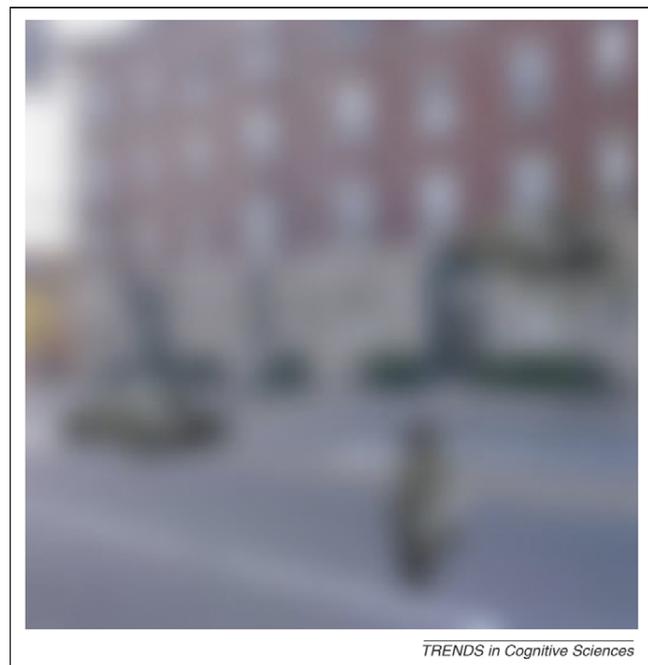


Figure 2. The strength of context. The visual system makes assumptions regarding object identities according to their size and location in the scene. In this picture, observers describe the scene as containing a car and pedestrian in the street. However, the pedestrian is in fact the same shape as the car, except for a 90° rotation. The atypicality of this orientation for a car within the context defined by the street scene causes the car to be recognized as a pedestrian.

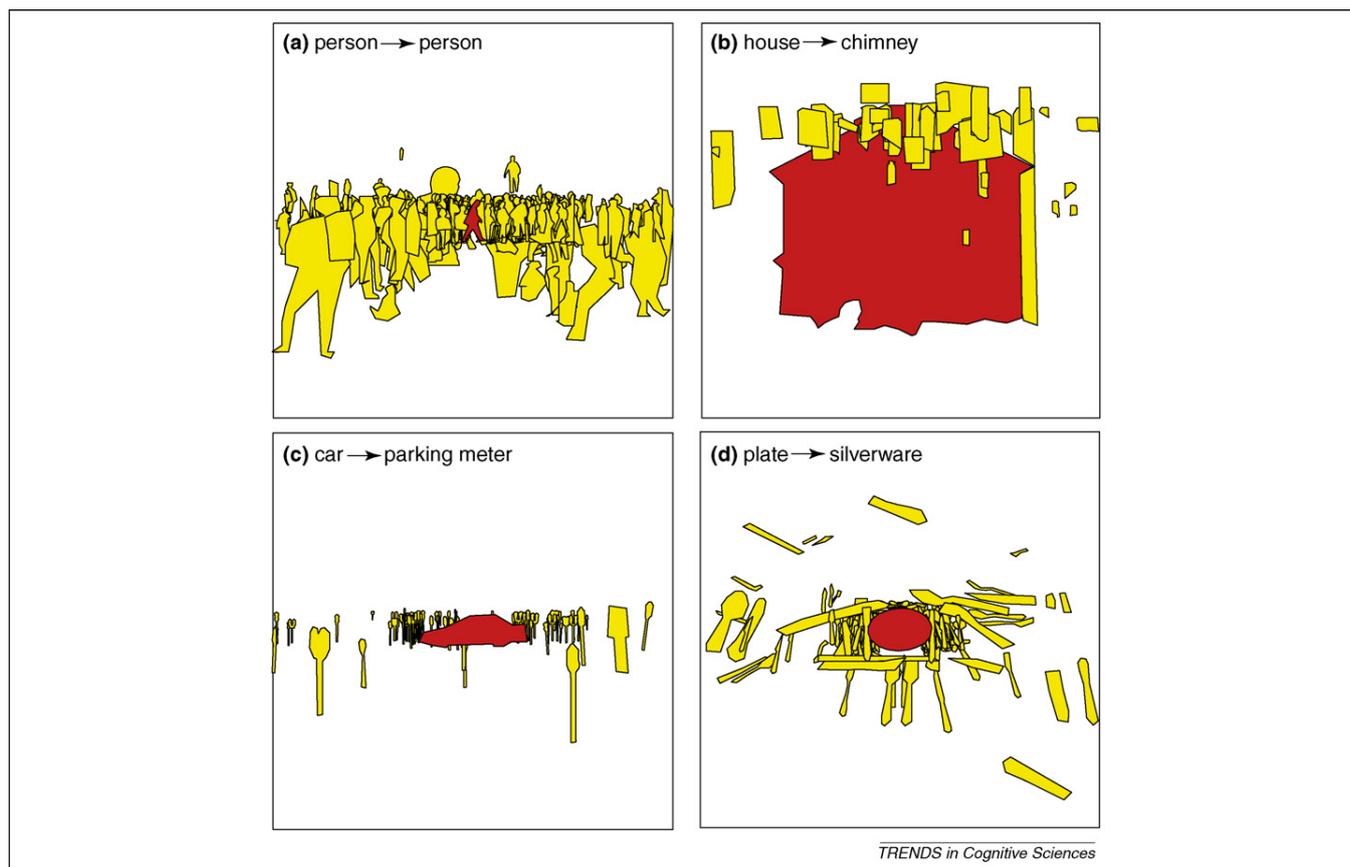


Figure 3. Measuring dependencies among objects. This figure illustrates the dependencies between objects, by plotting the distribution of locations, sizes and shapes of a target object (shown in yellow) conditional on the presence of a reference object (shown in red). To illustrate the form of the conditional distribution between target and reference objects, LabelMe [65], a large database of annotated objects, is used to search for all images containing the reference object in the pose specified. Then, all the target objects are plotted, preserving their relative position and scale with respect to the reference object. If in one image the location of a reference object is known, then the locations of other target objects are strongly constrained. The reference and target objects can belong to the same object category (a) or different object classes (b–d). The reference objects are (a) a person, (b) a house, (c) a car and (d) a plate and the target object is (a) another person in the same image, (b) a chimney, (c) a parking meter and (d) silverware.

Implicit learning of contextual cues

Fiser and Aslin [24,25] have shown that humans are good at routinely extracting temporal and spatial statistical regularities between objects and do so from an early age. Seminal work by Chun and Jiang [26] revealed that human observers can implicitly learn the contingencies that exist between arbitrary configurations of distractor objects (e.g. a set of the letter L) and the location of a target object (e.g. a

letter T), a form of learning called ‘contextual cueing’ (reviewed in Ref. [27]).

By showing that simple displays can elicit contextual learning, the paradigm of contextual cueing offers an interesting framework to determine which properties of the background provide informative context. By measuring how contextual cueing transfers to displays that differ slightly from the originally learned displays, this paradigm

Box 1. Mechanisms of contextual influences

At which level in the processing stream does contextual information have a role? In the literature, researchers often differentiate between two complementary mechanisms for explaining the effects of context on object recognition:

- 1) Contextual effects are mediated by memory representations by preactivating stored representations of objects related to the context [4,7]. Recognizing a scene context (e.g. a farm) enables an informed guess about the presence of certain classes of objects (e.g. a tractor versus a squid), in addition to their location and size.
- 2) Context changes the perceptual analysis of objects. Contextual effects occur early, at the perceptual level, facilitating the integration of local features to form objects [1–3,12]. These models suggest that even the way in which the image is analyzed is affected by context. The results of Auckland and collaborators [1], in addition to Davenport and Potter [3,66], support the existence of such a mechanism.

However, a definite answer to this issue remains a question for future work, and it is probable that multiple mechanisms have roles. In fact, contextual analysis involves a large network of different brain areas (reviewed in Ref. [12]) devoted to analysis of scene layout [67], object analysis (reviewed in Ref. [68]), and spatial and nonspatial associations [13,16].

The claim that object recognition is affected by the scene context is not without controversy. Hollingworth and Henderson [8] suggested the functional isolation hypothesis. In this model, object perception is isolated from information about the scene context and context does not interfere with the perceptual analysis of the target object. Their experimental results suggested that previous observations of contextual influences were owing to a response bias (for a discussion of those results, see Refs [1,66]).

can be used to identify statistically relevant features of the visual display that humans are sensitive to and can exploit for contextual cues. For instance, transfer between learned and novel displays is effective for geometric transformations that preserve the relative ordering of items across the whole display, such as stretching [28]. But transfer is significantly impaired by changes in viewpoint [29] or scene identity [30].

At a more local level, recent work has demonstrated similar magnitudes of contextual cueing when only two items surrounding the target are repeated or when the entire display is repeated [31,32]. A simple model quantifying this effect, by Brady and Chun [31], suggests that learning only the relationships between the locations of the local distractors and the target location is sufficient to demonstrate many of the major properties of contextual cueing: a small but robust effect of set size [26], the ability to recombine displays that cue the same location [28], and strong cueing from only the local configuration [33]. However, they also found a lack of transfer if the local configuration was moved, demonstrating that it was the location of the local layout within the global configuration that mattered for object detection.

A topic of current debate is the extent that context affects the speed at which attention is deployed towards the target [26,34], alters target analysis [35] or biases response selection [36]. In the real world, the following factors should be considered: co-occurrence might happen at a global (e.g. a kitchen will predict the presence of a stove) or local (e.g. a nightstand will predict the presence of an alarm clock) level; contextual associations can be definite or probabilistic; and observers might act on an object in a consistent manner, or not, [37] and might choose to rely on memory search, instead of visual search, when looking for objects in familiar scenes [38]. The respective roles of all these factors in explaining contextual influences constitute a challenging area for future investigation.

Perception of sets and summary statistics

A representation of context on the basis of object-to-object associations treats objects as the atomic elements of perception. It is an object-centered view of scene understanding. Here and in the next section, we will review work suggesting a cruder but extremely effective representation of contextual information, providing a complementary rather than an alternative source of information for contextual inference. In the same way that the representation of an object can be mediated by features that do not correspond to nameable parts [39], the representation of the scene context can also be built on elements that do not correspond to objects. Several recent pieces of work [40–44] have proposed that humans encode statistical properties from a display instead of encoding the individual elements that compose a display. The statistical properties currently under active investigation are the mean size and variance of a set of objects [40,42–44], the center of mass [41], texture descriptors [45] and also more complex structural information, such as the amount of clutter in an image [46], in addition to the mean depth and degree of perspective of a natural scene [47].

In a crucial paper, Ariely [40] found that, after presenting observers with a set of circular spots of various sizes for 500 ms, observers could judge the average size of the spots better than the sizes of individuals in the set. Importantly, the estimation of the mean size was unaffected by the number of objects in the display. Ariely's findings suggest that, from a quick look at a scene, observers know the mean size of a collection of homogeneous objects quite accurately but retain little information about the size of the individual objects. In a series of elegant studies, Chong and Treisman [42–44] demonstrated that this ability to compute the mean size was automatic, unaffected by the density of the display and generalized between displays that have different statistical distributions, confirming that human observers were indeed extracting the mean size from a brief glance at a display.

In a recent study, Alvarez and Oliva [41] observed that participants extracted another summary statistic, the center of mass of distractor elements. Moreover, this was done outside of the focus of attention. In a paradigm of multiple-object tracking, observers tracked objects moving continuously in a field of similar moving distractors. At a random moment during the trial, the distractors were deleted from the display. When asked to report the location of the distractors, observers reported the mean position of a group of distractors more accurately than the location of any individual distractor. In addition to Ref. [43], this result suggests that a statistical summary of ensemble features is computed automatically and outside of the focus of attention.

To what extent a statistical summary of an image influences local object processing is one of the outstanding questions in the field. Summary statistics are important because they provide an efficient and compact representation of the image that can be used to inform about scene properties, in addition to being used to prime local object features. The next section reviews recent work in computer vision that has shown the efficiency of such summary representations for encoding the structure and meaning of natural images (reviewed in Ref. [48]).

Global context: insights from computer vision

In computer vision, the most common approach to localizing objects in images is to slide a window across all locations and scales in the image and classify each local window as containing either the target or background. This approach has been successfully used to detect objects such as faces, cars and pedestrians (reviewed in Ref. [49]). However, contextual information can be used in conjunction with local approaches to improve performance, efficiency and tolerance to image degradation. One of the main problems that computational recognition approaches face by including contextual information is the lack of simple representations of context and efficient algorithms for the extraction of such information from the visual input.

Recent work in computer vision has shown that the identity of real-world scenes might be inferred from aggregated statistics of low-level features (Box 2) and has highlighted the importance of global scene representations as sources of contextual information [11,18,50]. These global

Box 2. Computing global features

There are two major families of global context representations: first, texture-based methods [50,52,69] or 'bag-of-words' models (a term borrowed from the literature on text analysis). A set of features are detected in the image and, once a decision has been taken about the presence or absence of a feature, the location from which it comes is not encoded in the representation. The scene descriptor is given by a vector in which each element encodes the number of times that each kind of feature appears in the image. Randomizing the spatial location of the features in the image would create an image with the same scene descriptor (Figure 1c). Despite their simplistic assumptions, these methods perform surprisingly well and can provide an initial guess of the scene identity. The second class of models encodes spatial layout [50,53]: the image is first divided into regions, and then each region is treated as a bag of words. The scene descriptor is a vector in which each element contains the number of times each type of feature appeared in each region. The final representation preserves some coarse spatial information. Randomizing the location of the edges within each region will produce a scene with the same descriptor (Figure 1d). However, moving features from one region to another will result in a different representation. This representation provides a significant increase in performance over bag-of-words models.

In the scene representation proposed in Ref. [50], the image is first decomposed by a bank of multiscale-oriented filters (tuned to six orientations and four scales). Then, the output magnitude of each filter is averaged over 16 nonoverlapping windows arranged on a 4×4 grid. The resulting image representation is a $4 \times 8 \times 16 = 512$ dimensional vector. The final feature vector, used to represent the entire image, is obtained by projecting the binned filter outputs onto the first 80 principal components computed on a large dataset of natural images. Other techniques involve computing histograms of complex features such as textons [69] or vector-quantized SIFT features (SIFT descriptors encode a local image patch by dividing the patch into 4×4 regions and computing the histogram of local image gradients within each region) [52,53,55]. Those features encode complicated patterns, such as grouping of edges. See Ref. [70] for a review of image representations used in applications for image indexing. Building more robust global scene representations will have a major impact on future object-detection systems.

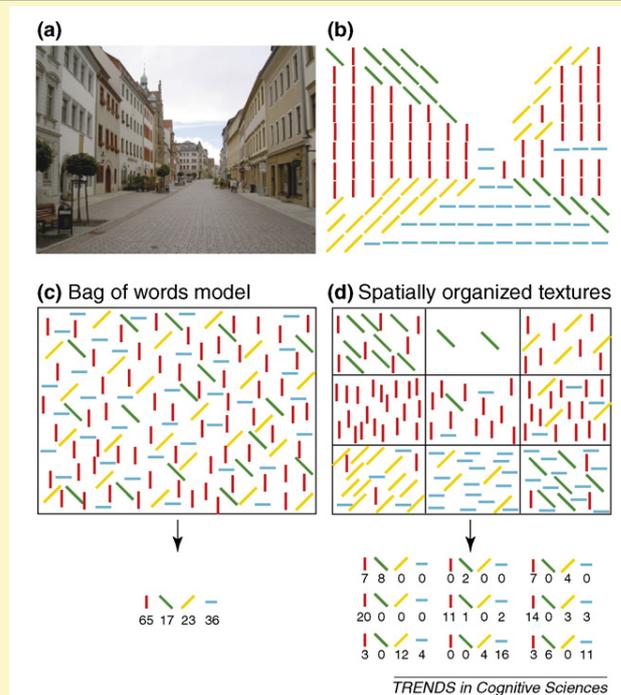


Figure 1. Computing global features. This illustration shows the general scheme underlying many current global scene representations [50,52,53,55,69]. (a) Input image. (b) A set of features is detected in the image. In this schematic example, the features are edges grouped into four different orientations at each location. (c,d) Summary of two scene representations. (c) A bag-of-words model in which location information is not explicitly stored (randomizing the spatial locations of the features results in the same representation). (d) Spatially organized textures; the image is partitioned into several regions. Each region is encoded as if it was a stationary texture, in which location is irrelevant. The final vector descriptor contains the number of times each feature is present at each region; therefore, spatial information is preserved at a coarse resolution.

image representations were developed within the framework of scene recognition (e.g. classifying an image as being a beach scene, street or living room [48]). The main characteristic of global image representations is that the scene is represented as a whole, rather than splitting it into its constituent objects. Such models correspond to the state of the art in scene recognition and context-based object recognition.

Box 2 summarizes the general framework used to compute global scene representations. These representations are derived from computing statistics of low-level features (similar to representations available in early visual areas, such as oriented edges and vector-quantized image patches) over fixed image regions. Despite the low dimensionality of the representation, global features preserve most of the relevant information needed for categorizing scenes into superordinate categories (e.g. nature, urban or indoor), which can be used to provide strong contextual priors. Because object information is not explicitly represented in the global features, they provide a complementary source of information for scene understanding, which can be used to improve object recognition. For instance, global features have been used to classify images into those that contain a particular object and those that do not [18,21,51], and this decision is taken without localizing

the object within the image. These representations are reminiscent of visual-cognition work on summary statistics, the perception of sets and contextual cueing.

Although they might not be the only mechanisms for scene recognition, global representations have been surprisingly effective at the scene-recognition task [50,52–56]. In tasks that require finer scene-category discrimination (living room versus dining room rather than city versus beach), recognition of specific objects will undoubtedly have a major role. Nevertheless, robust global scene representations will have a major impact in future object-detection systems.

Contextual effects on eye movements

When exploring a scene for an object, an ideal observer will fixate the image locations that have the highest posterior probability of containing the target object according to the available image information [57]. Attention can be driven by global scene properties (e.g. when exploring a street scene for a parking meter, attention is directed to regions near the ground plane) and salient objects contextually related to the target (e.g. when looking for a computer mouse, the region near a computer screen is explored first).

Most scenes can be recognized by just a glance, even before any eye movements can be initiated and without

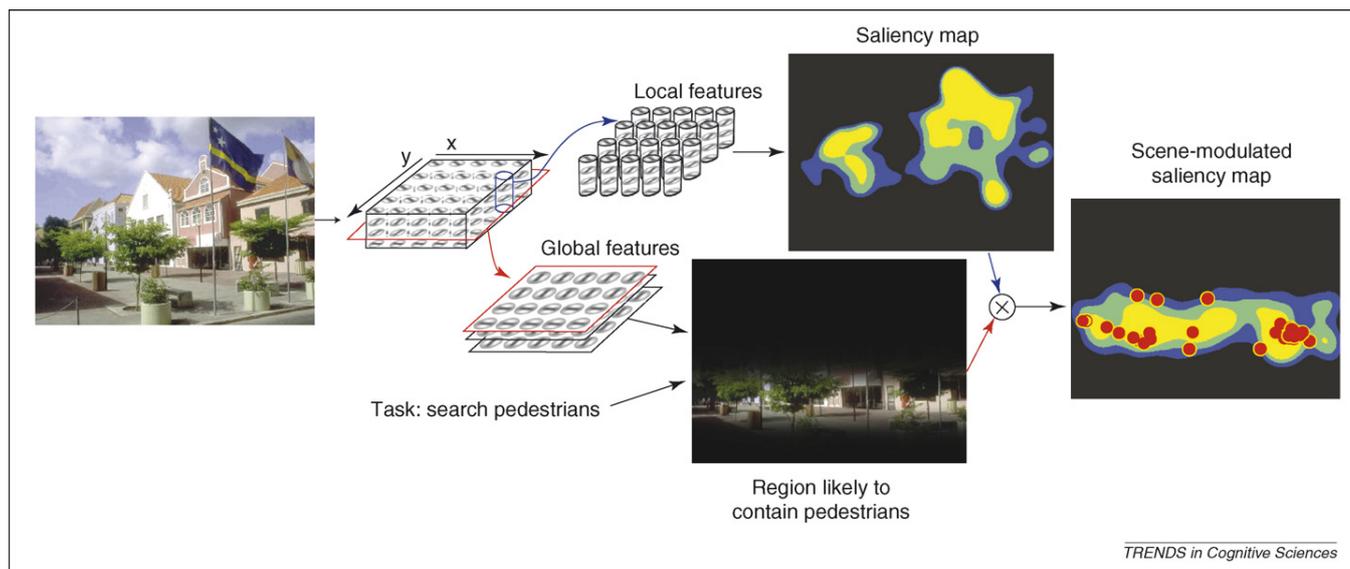


Figure 4. Global context effects in attention deployment. When looking for a target, eye movements make use of contextual information to direct attention to the most probable target location, and this happens from the first fixation. In the people-search task, saliency and contextual guidance are combined, resulting in a saliency map modulated by the task. The red dots show the first two locations fixated by eight participants. Participants scanned first the expected locations of the target according to the scene context and did not consider the most salient location (the flag). Adapted from Ref. [9].

requiring foveated vision to scrutinize the image [58,59]. Therefore, the scene content will have an immediate effect in the planning of subsequent eye movements [9,19,60,61], overriding salient regions that would otherwise attract attention. The role of context goes beyond focusing limited computational resources on the most relevant image regions. Instead, for contextually defined objects, the scene context will direct attention to the only image regions for which the target will have the appropriate role (e.g. if there is a parking meter on the roof of a house, it is certainly not the parking meter being looked for). Even for objects that preserve their identity in a large collection of different contexts (e.g. a mug or bicycle), contextual information will provide additional cues in cases of image degradation, such as noise, heavy occlusions or poor resolution, resulting in an increase in detection [19].

Computational models of attention [62] provide predictions about which regions are likely to attract observers' attention. Despite having no notion of the task or context, saliency models perform significantly better than chance in predicting image regions that will be fixated by participants. These models work best in situations in which the image itself provides little semantic information and no specific task is driving the observer's exploration. Saliency models can be enhanced by introducing task constraints and a context model [9,10,19,20,60,63]. In Ref. [9], a scene is analyzed by two parallel pathways (Figure 4). The local pathway represents each spatial location independently and is used to compute image saliency and perform object recognition on the basis of local appearance. The global pathway represents the entire image by extracting global statistics from the image (Box 2) and is used to provide information about the expected location of the target in the image. The contextual guidance model in Figure 4 predicts the image regions likely to be fixated by human observers

performing a natural object-search task (e.g. searching for a pedestrian, mug or painting).

Concluding remarks

A scene composed of contextually related objects is more than just the sum of the constituent objects. Objects presented in a familiar context are faster to localize and recognize. In the absence of enough local evidence about an object's identity, the scene structure and prior knowledge of world regularities might provide the additional information needed for recognizing and localizing an object. Even if objects can be identified by intrinsic information, context can simplify the object discrimination by decreasing the number of object categories, scales and positions that must be considered. How objects are remembered also depends on the scene context they are in [64].

But, how powerful are the real-world relationships between objects? To what extent is contextual information useful before the power of local features must be used to predict the identity of an object? Recent work in cognitive psychology and computer vision has shown that a statistical summary of the elements that comprise the scene can provide an extremely effective source of information for contextual inference.

Research on the mechanisms underlying contextual inference and scene recognition in humans [48], and its neural correlates [12], will begin to address these questions and also, in doing so, have far-reaching implications for computer vision, for which context-based object recognition is a fast growing area of research.

Acknowledgements

We thank George Alvarez, Moshe Bar, Timothy Brady, Michelle Greene, Barbara Hidalgo-Sotelo, Jeremy Wolfe and three anonymous reviewers for insightful comments on the manuscript. A.O. was funded by a National Science Foundation Career award (IIS 0546262) and a National

Science Foundation contract (IIS 0705677). A.T. was partly funded by the National Geospatial-Intelligence Agency NEGI-1582-04-0004.

References

- 1 Auckland, M.E. *et al.* (2007) Non-target objects can influence perceptual processes during object recognition. *Psychon. Bull. Rev.* 14, 332–337
- 2 Biederman, I. *et al.* (1982) Scene perception: detecting and judging objects undergoing relational violations. *Cognit. Psychol.* 14, 143–177
- 3 Davenport, J.L. and Potter, M.C. (2004) Scene consistency in object and background perception. *Psychol. Sci.* 15, 559–564
- 4 Friedman, A. (1979) Framing pictures: the role of knowledge in automatized encoding and memory of gist. *J. Exp. Psychol. Gen.* 108, 316–355
- 5 Gordon, R.D. (2004) Attentional allocation during the perception of scenes. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 760–777
- 6 Henderson, J.M. *et al.* (1999) Effects of semantic consistency on eye movements during scene viewing. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 210–228
- 7 Palmer, S.E. (1975) The effects of contextual scenes on the identification of objects. *Mem. Cognit.* 3, 519–526
- 8 Hollingworth, A. and Henderson, J.M. (1998) Does consistent scene context facilitate object detection. *J. Exp. Psychol. Gen.* 127, 398–415
- 9 Torralba, A. *et al.* (2006) Contextual guidance of attention in natural scenes: the role of global features on object search. *Psychol. Rev.* 113, 766–786
- 10 Torralba, A. (2003) Modeling global scene factors in attention. *J. Opt. Soc. Am. A* 20, 1407–1418
- 11 Hoiem, D. *et al.* (2006) Putting objects in perspective. *Proc. IEEE Comp. Vis. Pattern Recog.* 2, 2137–2144
- 12 Bar, M. (2004) Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629
- 13 Bar, M. and Aminoff, E. (2003) Cortical analysis of visual context. *Neuron* 38, 347–358
- 14 Goh, J.O.S. *et al.* (2004) Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. *J. Neurosci.* 24, 10223–10228
- 15 Aminoff, E. *et al.* (2007) The parahippocampal cortex mediates spatial and non-spatial associations. *Cereb. Cortex* 27, 1493–1503
- 16 Gronau, N. *et al.* (2007) Integrated contextual representation for objects' identities and their locations. *J. Cogn. Neurosci.* (in press)
- 17 Hock, H.S. *et al.* (1974) Contextual relations: the influence of familiarity, physical plausibility, and belongingness. *Percept. Psychophys.* 16, 4–8
- 18 Torralba, A. (2003) Contextual priming for object detection. *Int. J. Comput. Vis.* 53, 169–191
- 19 Eckstein, M.P. *et al.* (2006) Attentional cues in real scenes, saccadic targeting and Bayesian priors. *Psychol. Sci.* 17, 973–980
- 20 Peters, R.J. and Itti, L. (2006) Computational mechanisms for gaze direction in interactive visual environments. In *Proceedings of the 2006 Symposium on Eye Tracking Research and Applications (San Diego, California, March 27–29, 2006)*, pp. 27–32, ACM
- 21 Murphy, K.P. *et al.* (2003) Using the forest to see the trees: a graphical model relating features, objects and scenes. *Adv. in Neural Information Processing Systems* 16, 1499–1507
- 22 Chun, M.M. and Jiang, Y. (1999) Top-down attentional guidance based on implicit learning of visual covariation. *Psychol. Sci.* 10, 360–365
- 23 Green, C. and Hummel, J.E. (2006) Familiar interacting object pairs are perceptually grouped. *J. Exp. Psychol. Hum. Percept. Perform.* 32, 1107–1119
- 24 Fiser, J. and Aslin, R.N. (2001) Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol. Sci.* 12, 499–504
- 25 Fiser, J. and Aslin, R.N. (2005) Encoding multi-element scenes: statistical learning of visual feature hierarchies. *J. Exp. Psychol. Gen.* 134, 521–537
- 26 Chun, M.M. and Jiang, Y. (1998) Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognit. Psychol.* 36, 28–71
- 27 Jiang, Y. and Chun, M.M. (2003) Contextual cueing: reciprocal influences between attention and implicit learning. In *Attention and Implicit Learning* (Jimenez, L., ed.), pp. 277–296, John Benjamins Publishing Company
- 28 Jiang, Y. and Wagner, L.C. (2004) What is learned in spatial contextual cueing: configuration or individual locations? *Percept. Psychophys.* 66, 454–463
- 29 Chua, K.P. and Chun, M.M. (2003) Implicit scene learning is viewpoint-dependent. *Percept. Psychophys.* 65, 72–80
- 30 Brockmole, J.R. *et al.* (2006) Contextual cueing in naturalistic scenes: global and local contexts. *J. Exp. Psychol. Learn. Mem. Cogn.* 32, 699–706
- 31 Brady, T. and Chun, M.M. (2007) Spatial constraints on learning in visual search: modeling contextual cuing. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 798–815
- 32 Song, J.-H. and Jiang, Y. (2005) Connecting the past with the present: how do humans match an incoming visual display with visual memory? *J. Vis.* 5, 322–330
- 33 Olson, I.R. and Chun, M.M. (2002) Perceptual constraints on implicit learning of spatial context. *Vis. Cogn.* 9, 273–302
- 34 Peterson, M.S. and Kramer, A.F. (2001) Attentional guidance of the eyes by contextual information and abrupt onsets. *Percept. Psychophys.* 63, 1239–1249
- 35 Hidalgo-Sotelo, B. *et al.* (2005) Human learning of contextual priors for object search: where does the time go? In *Proceedings of the 3rd Workshop on Attention and Performance in Computer Vision at the Int. CVPR*, pp. 1063–1069, IEEE Computer Society
- 36 Kunar, M.A. *et al.* (2007) Does contextual cueing guide the deployment of attention? *J. Exp. Psychol. Hum. Percept. Perform.* 33, 816–828
- 37 Land, M.F. and Hayhoe, M.M. (2001) In what ways do eye movements contribute to everyday activities? *Vision Res.* 41, 3559–3565
- 38 Oliva, A. *et al.* (2004) Panoramic Search: the interaction of memory and vision in search through a familiar scene. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 1132–1146
- 39 Ullman, S. *et al.* (2002) Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.* 5, 682–687
- 40 Ariely, D. (2001) Seeing sets: Representation by statistical properties. *Psychol. Sci.* 12, 157–162
- 41 Alvarez, G. and Oliva, A. The representation of simple ensemble visual features outside the focus of attention. *Psychological Science* (in press)
- 42 Chong, S.C. and Treisman, A. (2003) Representation of statistical properties. *Vision Res.* 43, 393–404
- 43 Chong, S.C. and Treisman, A. (2005) Attentional spread in the statistical processing of visual displays. *Percept. Psychophys.* 67, 1–13
- 44 Chong, S.C. and Treisman, A. (2005) Statistical processing: computing the average size in perceptual groups. *Vision Res.* 45, 891–900
- 45 Chubb, C. *et al.* (2007) The three dimensions of human visual sensitivity to first-order contrast statistics. *Vision Res.* 47, 2237–2248
- 46 Rosenholtz, R. *et al.* (2007) Measuring Visual Clutter. *J. Vis.* 7, 1–22
- 47 Greene, M.R. and Oliva, A. (2006) Natural scene categorization from the conjunction of ecological global properties. In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*. (Vol. 1), pp. 291–296, Cognitive Science Society
- 48 Oliva, A. and Torralba, A. (2005) Building the gist of a scene: the role of global image features in recognition. *Prog. Brain Res.* 155, 23–36
- 49 Forsyth, D.A. and Ponce, J. (2003) *Computer Vision. A Modern Approach*. Prentice Hall
- 50 Oliva, A. and Torralba, A. (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175
- 51 Torralba, A. and Oliva, A. (2003) Statistics of natural images categories. *Network Comput. Neural Syst.* 14, 391–412
- 52 Fei-Fei, L. and Perona, P. (2005) A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 2), pp. 524–531, IEEE Computer Society
- 53 Lazebnik, S. *et al.* (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *Proceeding of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. 2), pp. 2169–2178, IEEE Computer Society
- 54 Vailaya, A. *et al.* (1998) On image classification: city images vs. landscapes. *Pattern Recognit.* 31, 1921–1935
- 55 Bosch, A. *et al.* (2006) Scene classification via pLSA. *Lecture Notes in Computer Science* 3954, 517–530
- 56 Vogel, J. and Schiele, B. (2006) Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. *Int. J. Comput. Vis.* 72, 133–157
- 57 Najemnik, J. and Geisler, W.S. (2005) Optimal eye movement strategies in visual search. *Nature* 434, 387–391

- 58 Thorpe, S. *et al.* (1996) Speed of processing in the human visual system. *Nature* 381, 520–522
- 59 Potter, M.C. (1975) Meaning in visual scenes. *Science* 187, 965–966
- 60 Neider, M.B. and Zelinski, G.J. (2006) Scene context guides eye movements during visual search. *Vision Res.* 46, 614–621
- 61 Over, E.A.B. *et al.* (2007) Coarse-to-fine eye movement strategy in visual search. *Vision Res.* 47, 2272–2280
- 62 Itti, L. and Koch, C. (2001) Computational Modeling of Visual Attention. *Nat. Rev. Neurosci* 2, 194–203
- 63 Tsotsos, J.K. *et al.* (1995) Modeling visual-attention via selective tuning. *Artif. Intell.* 78, 507–545
- 64 Hollingworth, A. and Henderson, J.M. (2003) Testing a conceptual locus for the inconsistent object change detection advantage in real-world scenes. *Mem. Cognit.* 31, 930–940
- 65 Russell, B. *et al.* (2007) LabelMe: a database and web-based tool for image annotation. *MIT AI Lab Memo AIM-2005-025*
- 66 Davenport, J.L. (2007) Consistency effects between objects in scenes. *Mem. Cognit.* 35, 393–401
- 67 Epstein, R. and Kanwisher, N. (1998) A Cortical Representation of the Local Visual Environment. *Nature* 392, 598–601
- 68 Kanwisher, N. (2003) The ventral visual object pathway in humans: evidence from fMRI. In *The Visual Neurosciences* (Chalupa, J. and Werner, J.S., eds), pp. 1179–1189, MIT Press
- 69 Walker Renninger, L. and Malik, J. (2004) When is scene identification just texture recognition? *Vision Res.* 44, 2301–2311
- 70 Smeulders, A.W.M. *et al.* (2000) Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intel.* 22, 1349–1380

Five things you might not know about Elsevier

1.

Elsevier is a founder member of the WHO's HINARI and AGORA initiatives, which enable the world's poorest countries to gain free access to scientific literature. More than 1000 journals, including the *Trends* and *Current Opinion* collections and *Drug Discovery Today*, are now available free of charge or at significantly reduced prices.

2.

The online archive of Elsevier's premier Cell Press journal collection became freely available in January 2005. Free access to the recent archive, including *Cell*, *Neuron*, *Immunity* and *Current Biology*, is available on ScienceDirect and the Cell Press journal sites 12 months after articles are first published.

3.

Have you contributed to an Elsevier journal, book or series? Did you know that all our authors are entitled to a 30% discount on books and stand-alone CDs when ordered directly from us? For more information, call our sales offices:

+1 800 782 4927 (USA) or +1 800 460 3110 (Canada, South and Central America)
or +44 (0)1865 474 010 (all other countries)

4.

Elsevier has a long tradition of liberal copyright policies and for many years has permitted both the posting of preprints on public servers and the posting of final articles on internal servers. Now, Elsevier has extended its author posting policy to allow authors to post the final text version of their articles free of charge on their personal websites and institutional repositories or websites.

5.

The Elsevier Foundation is a knowledge-centered foundation that makes grants and contributions throughout the world. A reflection of our culturally rich global organization, the Foundation has, for example, funded the setting up of a video library to educate for children in Philadelphia, provided storybooks to children in Cape Town, sponsored the creation of the Stanley L. Robbins Visiting Professorship at Brigham and Women's Hospital, and given funding to the 3rd International Conference on Children's Health and the Environment.