Esaliency (Extended Saliency): Meaningful Attention Using Stochastic Image Modeling

Tamar Avraham and Michael Lindenbaum, *Member*, *IEEE*

Abstract—Computer vision attention processes assign variable-hypothesized importance to different parts of the visual input and direct the allocation of computational resources. This nonuniform allocation might help accelerate the image analysis process. This paper proposes a new bottom-up attention mechanism. Rather than taking the traditional approach, which tries to model human attention, we propose a validated stochastic model to estimate the probability that an image part is of interest. We refer to this probability as saliency and thus specify saliency in a mathematically well-defined sense. The model quantifies several intuitive observations, such as the greater likelihood of correspondence between visually similar image regions and the likelihood that only a few of interesting objects will be present in the scene. The latter observation, which implies that such objects are (relaxed) global exceptions, replaces the traditional preference for local contrast. The algorithm starts with a rough preattentive segmentation and then uses a graphical model approximation to efficiently reveal which segments are more likely to be of interest. Experiments on natural scenes containing a variety of objects demonstrate the proposed method and show its advantages over previous approaches.

Index Terms—Computer vision, scene analysis, similarity measures, performance evaluation of algorithms and systems, object recognition, visual search, attention.

1 Introduction

Image analysis processes often scan the image exhaustively, looking for familiar objects of different location and size. This process can be made much more efficient if an attention mechanism assigns priorities to different image parts and thus directs the analysis process to examine more interesting locations first. The highly effective attention mechanisms of the human visual system have been studied extensively from the psychophysical and physiological points of view. Interestingly, it was found that effective attention is possible even in the absence of any information about the sought for entities.

Neisser [33] suggested that human visual processing is divided into preattentive and attentive stages. The first consists of parallel processes that simultaneously operate on large portions of the visual field and form the units to which attention may then be directed. The second stage consists of a focused processing effort applied to a limited portion of the visual field. Triesman and Gelade [52] suggested the *feature integration theory* (FIT). Consistent with physiology findings is their suggestion that, in the preattentive stage, the visual input is represented by separate retinotopic maps for each of the basic visual attributes. According to their hypothesis, the binding of the features (such as color and shape) requires focal attention. Initially, they suggested that there is a dichotomy between visual search tasks in which the target is located immediately (*pop-out*) and search tasks that required scanning. Much evidence (e.g., [31], [15], or the survey [62]) was found, however, against this dichotomy, which is nowadays considered outdated, and the FIT was updated accordingly [51]. Many computational models (e.g., [28], [61], [53], [26]) followed other principles of the FIT and suggested methods for grading the conspicuousness of each spatial location in a viewed scene using the saliency map concept. Itti et al. [26], for instance, following Koch and Ullman [28], have suggested such a bottom-up computational model. Given an input image, separate feature maps of color, intensity, and orientation in different scales are extracted using linear filtering. Local spatial contrast is estimated for each feature at each location, providing a separate conspicuity map for each feature. These are combined to form one saliency map that guides the attention focus. Using *winner-take-all* and *inhibition-of-return* mechanisms, attention is drawn to different locations in descending priority (saliency) order.

Human attentional preferences are also directed by top-down information, which might include prior world knowledge or task-driven information [63]. Combining top-down information, when available, in attention models improves their predictive abilities (e.g., [61], [32], [22]). Surprisingly, however, it was shown that bottom-up information alone might lead to fixation paths that are highly correlated with those of humans (e.g., [35], [9]).

Note that the locations to which attention is directed may be specified, in principle, at every spatial point of the retinotopic saliency map. Methods that do so are also, therefore, referred to as *space-based* or *feature-based*. For any salient location, attention may be focused on a region around it, as suggested by the *spot light* [39] or the *zoom-lens* [21] metaphors. Note that this region may contain only one object, a part of an object, or several objects.

• *The authors are with the Computer Science Department, Technion—Israel Institute of Technology, Haifa 32000, Israel.*
E-mail: {tammya, mic}@cs.technion.ac.il.

Manuscript received 17 July 2007; revised 19 Jan. 2009; accepted 20 Feb. 2009; published online 2 Mar. 2009.
Recommended for acceptance by R. Zabih.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-07-0430.
Digital Object Identifier no. 10.1109/TPAMI/2009.53.

An alternative approach suggests that human visual attention is *object-based*. That is, the spatial region of attention is not of fixed shape but rather adapted to the perceived object [18], [44]. This approach implies that the decision as to where to place one's attention follows perceptual grouping processes. The term "object" should be specified with care in this context. It is unlikely, for example, that an almost perfect grouping process, in which the perceived scene is divided into semantically meaningful objects such as cars or people, precedes the attention process. Several psychological studies (e.g., [58]) provide evidence that "objects" are specified by simple grouping processes related to basic Gestalt laws [59] such as proximity, similarity, and uniform connectedness [36]. To avoid confusion between a model that relies on semantically meaningful segmentation and one that relies on relatively weak grouping cues, we sometimes refer to the latter as *region-based*.

In computer vision, Itti et al.'s model is currently considered the dominant bottom-up saliency method and many variations of it have been proposed; see, e.g., [17], where a more consistent use of scale is provided. Bottom-up attention mechanisms were found useful in computer vision for recognition (e.g., [46], [23], [42]) and for learning [42].

As described above, the vast majority of attention models identify saliency with local exception. That is, the saliency value at each location is essentially the local contrast in one feature or more. These models do not check uniqueness in the context of the whole scene. Many instances from the same category appearing in one image can be considered salient if each contrasts with the background, while a single instance of another category can be considered less salient if it contrasts less with the background. Therefore, while this approach may be biologically plausible, it is suboptimal for computer vision. We propose an alternative saliency method, based on a (validated) stochastic model. The proposed *extended saliency* (or *Esaliency*) algorithm differs significantly from previous methods in its motivation, its methodology, and its end result. Rather than trying to build a model explaining human attention, we propose to use a validated stochastic model to estimate the probability that an image part is of interest. We refer to this probability as saliency, and thus, specify saliency in a mathematically well-defined sense. Usually saliency values have only a relative meaning. That is, a high saliency value implies that the corresponding location is more likely to be of interest than another location associated with a lower saliency value. Using our approach, the (E)saliency value is meaningful in and of itself. Also, note that, with this saliency, the common practice of scanning the candidates for attention with decreasing saliency order is not only a very reasonable heuristic, but also an optimal strategy for minimizing the expected scanning time until a object of interest is found.

The proposed *Esaliency* method differs from most previous methods in several other ways. First, it is *region-based*, as it begins with a rough preattentive grouping process. The uniform regions that are suggested by the grouping process are used as initial candidates for attention. Second, the judgment as to whether some part of the image is salient is context-sensitive and global. If, for example, a locally salient object appears many times in the image, its saliency is reduced. Finally, the (local) uniqueness assumption is replaced by a preference for a small number of similar or dissimilar salient regions that may be located near to or far away from one another. This is compatible with real scenes, which may contain multiple interesting objects (of one or more categories).

Intuitively, we know that objects from the same category are likely to be visually similar. Therefore, two visually similar candidates are likely to both be associated with objects of interest or to both be associated with noninteresting objects. In a stochastic context, this intuition is quantified by the second order statistics (correlation) between the candidates' labels. The first order statistics is set using the common expectation for a relatively small number of interesting objects in a scene. A graphical model that approximately satisfies these constraints is constructed, allowing the hypotheses for salient locations with the highest likelihoods to be efficiently revealed. In our implementation, the algorithm relies on a simple method of candidate characterization and on a simple similarity measure, and is therefore fast. Yet, we show that using the fixation order specified by the descending saliency makes the whole process of object recognition, or object detection, much more efficient.

As was done in previous computer vision attention studies (e.g., [26], [55]), we focus here on a specific but wide context: images where the interesting occurrences take up only a small fraction of the image. In this context, the attention task is often denoted *visual search*, and the goal is to find the (small) image regions where the important objects lie. Attention processes can deal with other contexts where, for example, a single object of interest takes up a large fraction of the image. Then, however, the visual system task is not to find the object but rather to analyze it, and the attention is heavily knowledge-dependent. We do not consider the latter context here.

The computer vision community uses the term saliency, borrowed from cognitive psychology, for two different tasks. Attention-motivated saliency, considered, e.g., in [26], [9], [45] and here, aims to detect promising, high-priority image parts on which high-level processes focus their resources, thus achieving more efficient analysis. The term saliency is also used sometimes in the context of local invariant features (e.g., [27], [43], [24]). These may serve as a partial description of the image, or at least as an "adaptive coordinate system." Local feature detectors aim to find a relatively large number of points (or small regions), whose locations, relative to the objects in the scene, are stable under pose and illumination changes. The two tasks benefit from different properties of image exceptions. For attention-motivated saliency, exceptions turn out to be a reliable indication of significant scene objects. For local feature detection, local exceptions carry more information than other points, and more importantly, are usually stable under imaging changes. Unlike attention-motivated saliency, local feature detection is tuned to provide a large number (typically several hundreds) of detections in an image. Every detection corresponds to a small part of the object, and it is typical for many to correspond to the same object. Indeed, the local feature detection maps are

completely different from attention-motivated saliency maps. Therefore, we see local feature detections as only weakly relevant to the approach described here. The algorithm proposed here is very different from previous algorithms proposed for both tasks. It could also be used, in principle, for local feature detection. But, it is expected to function nonoptimally in that context because the weak global exceptions that we look for are not necessarily desirable. See [57], however, for a use of global exception to accelerate the correspondence process.

We are aware of only a few papers related to the region-based attention approach proposed here. Most attention mechanisms are *space-based*. A computer vision implementation of Duncan's [18] object-based attention model was suggested in [49]. To be effective, however, it requires high-quality hierarchial segmentation (and indeed uses in the experiments, human segmentations), and therefore, seems impractical for attention of complex natural scenes. A region-based approach relying on local saliency was suggested in [29].

Several space-based studies propose alternative methods for computing global exceptions. The methods in [50] and [9] estimate distributions corresponding to the image content and search for exceptional locations that correspond to content of low likelihood. In [9], for instance, the distribution of ICA coefficients across the image is estimated by a histogram. The coefficient's self-information is set higher when its value is less common. The saliency at an image point is specified as the sum of self-information values over all coefficients corresponding to that location. Experimentally, the authors show their model to be comparable with Itti et al.'s model for predicting human eye fixation paths. One disadvantage of this approach, in our opinion, is its insensitivity to the *degree of similarity* between candidates: Note, for example, that the self-information of a red element among pink distractors is either zero or similar to the self-information of the same red element among green distractors, depending on the histogram binning.

A related stochastic model was developed in [5] for a supervised search mechanism, optimizing the interaction between search and recognition. The problem considered in [5] is different from the *Esaliency* algorithm proposed here, which is purely bottom-up, and, like previous algorithms of this type (e.g., [26]), neither assumes the availability of an object recognition oracle nor changes the saliencies after they are set. The image (or video) analysis approach [8] considers a region noninteresting if it is similar to another image region. This principle is close to our approach. However, it differs from ours in several important ways: First, it does not rely on an explicit stochastic model of the objects' identities. Furthermore, it does not allow for two (or a few) similar items that together are globally unique to be considered salient. Finally, it relies on a relatively costly, part-based similarity evaluation. Therefore, while it gives very good results as an analysis tool, it does not seem useful as a quick prerecognition attention mechanism.

We present our basic assumptions and the stochastic model in Section 2, and then, describe the Esaliency algorithm in Section 3. In Section 4, we describe experiments that test the algorithm on synthetic data and on a few data sets of natural scenes. We then compare its results to those of a feature-based attention method [26], test a few possible extensions to the basic algorithm, compare Esaliency's results with human eye fixations, and test the benefits of using Esaliency for the task of pedestrian detection. Some future research directions are suggested in Section 5. A short version of this work, using a more complex, substantially different algorithm, was presented in [4].

2 Framework and Underlying Assumptions

The attention process proposed in this paper is region-based. That is, it builds on a preattentive grouping process, dividing the image into segments, which are the candidates for attention. The attention process associates each candidate with a quantitatively meaningful saliency value, which is an estimate of its likelihood to be a target.

The segmentation process need not be accurate and can actually be very rough, resulting in fragmented objects. In fact, even with the best available segmentation algorithms, setting the parameters for oversegmentation is the only way to ensure, with high probability, that most objects are not split between segments. A (segment) candidate is denoted as "target" if it corresponds to an object of interest or to a part of it. Since our model is bottom-up and no specific task is predefined, an "interesting object" is specified by common human knowledge. This is similar to the meaning suggested in [37], where the Internet users selected "the most interesting points in various scenes," without further guidance or goal. We consider realistic scenes where several objects of interest may be present. This, and the inevitable object fragmentation, imply that several candidates may be targets.

The dominant approaches to saliency (largely based on feature integration theory and the implied center-surround mechanisms) are based on the search for a local exception. This approach may fail in two ways: First, the presence of several targets in nearby locations might reduce the response to each and cause the attention mechanism to miss them. Moreover, a large set of similar objects (or even just regions), which are, by definition, nonexceptions, might be falsely detected as salient if they are associated with a high local contrast. Then, the true exceptions, associated with a lower contrast, will be missed. For instance, consider the simple synthetic case of 10 red disks and one yellow disk scattered over a white background. Although it is obvious that the yellow disk is the exception, locally salient searching models may suggest each of the red disks as more salient as they contrast more with the background.

Our approach, denoted as *Extended saliency* (Esaliency), seeks relaxed global exceptions; it prefers objects that belong to small groups of similar objects that are relatively dissimilar to the rest of the image. It would recognize the yellow exception in the colored disks example above. Moreover, if, for instance, there were three yellow disks, each of them would still be recognized as the most "visit-worthy" item in that display.

2.1 Stochastic Modeling of Target-Nontarget Labels

Our approach to the design of the saliency algorithm is to quantify the target-nontarget labels of the candidates in a probabilistic model, which would eventually identify the saliency of a candidate with its probability to be a target.

Formally, let $(c_1, \ldots, c_n)$ denote the candidates for attention. Taking a stochastic approach, we consider the labels of the candidates $l_1, \ldots, l_n$ as binary random variables, which take value 1 if the candidate is a target and 0 if it is a nontarget. Estimating the probability $P(l_i) = p(l_i = 1)$ that the candidate is a target, is the goal of this paper. To estimate this probability, we take an indirect approach and start with the corresponding joint distribution. Let $\bar{l} = (l_1, \ldots, l_n)$ denote a vector of candidate labels, and $\mathcal{L} = \{\bar{l} = (l_1, \ldots, l_n); l_1, \ldots, l_n \in \{0, 1\}\}$ be the set of all 2^n label vectors. Let $p(\bar{l})$ be a probability distribution function on $\mathcal{L}$.

We shall now make some observations which constrain the distribution $p(\bar{l})$. Then, in Section 3, we propose a specific distribution approximating these constraints, and a computationally efficient way for estimating the probability of each candidate to be a target.

2.2 Underlying Observations

To construct the Esaliency process, we combine the following three related but different observations, which we elaborate on and quantify in the rest of this section:

Observation 1. The number of target candidates is usually small. The number of interesting objects in a scene, and the total area that they cover, is usually small.

Observation 2. There is a correlation between visual similarity and target-nontarget labels. Objects belonging to the same category are usually all targets or all nontargets. Candidates associated with these objects are often visually similar. Therefore, two visually similar candidates are likely to be both targets or both nontargets. Note that if the candidates are dissimilar, then, independently, every one of them may be a target or a nontarget.

Observation 3. Natural scenes are often composed of clustered structural units. We argue that natural images may often be partitioned into small parts clustered in some feature space. That is, the feature vectors characterizing the parts are not uniformly distributed in the feature space but rather concentrated in a few locations.

These observations do not hold for every scene, but they do for many of them. We found that using them as the basis for the proposed stochastic mechanism enabled it to direct the attention focus on interesting objects and reject most of the image background, even when it is highly textured.

Note that combining the two last observations implies that every cluster is associated with candidates of the same target/nontarget label. Also, note that adding the first one implies that the targets are in small cluster(s).

Using these observations directly would lead to a simple attention algorithm: Start by clustering and then choose the smallest resulting clusters. This approach, however, has its drawbacks: First, clustering requires some knowledge for setting the number of clusters or the maximal cluster diameter. Second, hard clustering methods provide a specific partition of the data that is sometimes just a little better than other possible partitions. Last, it is not clear how to assign continuous saliency values, or even just priorities, to the members of the selected small clusters and, in particular, how the distances within the clusters and between them influence this assignment. Our approach avoids these difficulties by constructing a distribution on the possible target/nontarget joint assignment, which may be viewed as a "soft" clustering.

In the rest of this section, we provide some empirical evidence for these (indirect) observations and quantify them in the context of the proposed stochastic model.

2.2.1 The Number of Targets Is Usually Small

Looking at many images, we observed that the number of interesting objects (denoted targets) in a scene and the total area that they cover are usually small. This observation is experimentally supported (see below) but may be also explained as follows: When the image contains only a few objects, they are important because they are usually the image content (the figure-ground case). In the case where the number of objects in the scene is large, many of them may belong to the same category (e.g., a group of trees, flowers, pebbles, or horses). Each of these objects is then nonspecial and noninteresting. (Note the human attention mechanism's ability to efficiently filter out homogenous distractors in an oddity search.) The number of objects that do not belong to large groups is usually small unless the image is unnaturally cluttered. These objects are therefore special and considered interesting.

As discussed in Section 1, this paper (as well as other attention studies in computer vision) focuses on the common context where the targets take up only a small fraction of the image. This is the case in many images of outdoor scenes. The 156 typical images of natural scenes taken from the University of Washington's Ground Truth database (UWGT database) [2] were, in our experiments, oversegmented on the average to 306 regions. The same images were presented to human observers, who were asked to mark "the interesting objects in each image." On average, 12.3 segments per image were associated with selected objects; this is about 4 percent of the segments.

Another example would be the MIT StreetScenes database [7], where the (arguably) most interesting objects, people, are indeed few and take up a small fraction of the image. For the 852 images containing people, there are, on average, 1.7 people (targets) per image, and the average fraction of the area covered by them is 2.14 percent. These examples support Observation 1.

In the stochastic context, this observation is simply expressed as a relatively low expected value μ_i for every random variable l_i. When no knowledge about the size, location, and the properties of targets, nor any general knowledge about the scene are available, the probability of a candidate to be a target is uniformly set as $\mu_i = \mu$. In our basic implementation of Esaliency, we set $\mu = 0.05$. We show in Section 4 that this setting is a reasonable description of natural scenes. Moreover, we found that the Esaliency algorithm is not sensitive to the exact value of μ, as long as it is relatively small.

The uniform setting of priors may be modified when some candidates are preferred, due to, say, high local saliency, or a preference for certain image locations; see Section 4.4 for related experiments.

2.2.2 A Correlation between Visual Similarity and Target-Nontarget Labels

According to observation 2, two visually similar candidates are likely to both be objects of interest, from the same

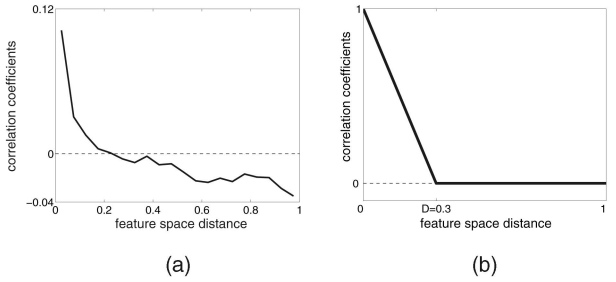

Fig. 1. Label correlation versus feature-space distance. (a) Empirically estimated using the UWGT database. (b) (Piecewise linear descending) function used in all experiments.

category, or likely to both be noninteresting objects. This observation is actually quite obvious and serves as the basis for all categorization algorithms that try to find a certain object using visual similarity.

In this paper, the similarity between candidates is inferred from their feature space distance. That is, every candidate is represented as a vector of, say, texture and color features. A short distance between the two vectors indicates that the corresponding candidates are visually similar. For complex object categories and different imaging conditions, the distance between different images from the same category is not necessarily small. Yet, we found that this assumption still holds most of the time. Moreover, we observed that different instances of the same category in a single image are more similar to each other than different instances of the same category in different images.

Let d_{ij} denote the feature-space distance between the two vectors associated with the ith and the jth candidates. A natural model of the dependency between the corresponding two labels l_i, l_j uses the correlation coefficients as a descending function of d_{ij}:

$$\rho(l_i, l_j) = \frac{\mathrm{cov}(l_i, l_j)}{\mathrm{var}(l_i)\mathrm{var}(l_j)} = \gamma(d_{ij}), \tag{1}$$

where $\gamma(d_{ij})$ is 1 for zero feature-space distance and approaches 0 for an increasing d_{ij}.

This dependency characterization was first proposed and verified in [5]. In a typical (new) verification experiment, we took a set of images (156 images from the UWGT database [2]) and asked naive observers to mark the significant objects in the scene (people, cars, etc.). We used the segmentation, features, and the similarity measure described in Section 4.1. The correlation coefficients between the target-nontarget labels of such pairs as a function of the feature-space distance is described in Fig. 1a. Note that, as expected, the correlation is higher for similar candidates (low feature-space distance) and decreases for decreasing similarity. Experimenting with several natural scene images, we found that a piecewise linear descending function, as shown in Fig. 1b, is a good approximation of the measured correlation-coefficient behavior for several choices of feature-space and metrics. This dependency model is used in all of the experiments.

2.2.3 Natural Scenes Are Often Composed of Clustered Structural Units

We argue that segment candidates from one image are clustered. That is, their feature vectors can be divided into subsets so that all vectors in the same subset are close. This is actually intuitive: Typical scenes are associated with a limited palette of colors or textures, related to the type of scene (e.g., urban or landscape), the season, the type of trees, flowers, or animals in it, etc. (e.g., [54], [56]).

To demonstrate the clustering observation, we carried out a simple experiment comparing the clustering within specific images to the clustering within a set of different natural images. For each image, we applied the same segmentation and feature extraction process as in the Esaliency algorithm implementation. This yielded a seven-dimension feature vector for each candidate of each image (see Section 4.1 for details).

We estimated the inclination of a given set of feature vectors to cluster as follows: The data were clustered to a mixture of multivariate Gaussians using the EM algorithm [16], following the method suggested in [12]. The number of clusters k was specified in the range of 1-10. The EM algorithm was repeated 10 times for each k value, with different initializations, and, for each k, the clustering associated with the maximum likelihood was retained. For each image, the minimum description length (MDL [41]) of the best clustering associated with each k was calculated, and the clustering associated with the lowest MDL was chosen.

For this experiment, we again used the same image set from the UWGT database [2]. The clusterings were done first for sets of feature vectors from the same image, and then, for similarly sized sets of feature vectors randomly sampled from different images in the full image set. Fig. 2 describes the histograms of the best (lowest) MDL for the two cases, as well as the histograms of the associated number k and the log of the likelihoods corresponding to the selected best clusterings. When the data are taken from one image, the number of clusters is approximately uniformly distributed across the range $[4, 10]$. The likelihood is high and the MDL is low (compared to the second scenario). This indicates that the mixture of a few narrow Gaussians is a good representation of the data, i.e., the data are clustered. Data drawn randomly from different images are described best, on the other hand, by two to five very wide Gaussians (usually by two), associated with a much lower likelihood and a much higher MDL. The choice of a relatively low number of Gaussians by the MDL analysis means that using more Gaussians does not better explain the data. Therefore, this clearly indicates that the data are scattered in the feature space.[1]

This observation is not quantified but is used for choosing the joint target/nontarget distribution; see Section 3.

2.3 Stochastic Modeling Discussion

The stochastic modeling suggested here is related to recent work on statistics of natural images (see, e.g., [48]). However, it differs from most image statistical modeling, where all locations in the image contribute evenly to the obtained distributions, in that it depends heavily on a target/nontarget ground truth labeling. In a sense, it is more related to the work on grouping cue statistics [30],

1. Some of the description lengths are negative. This is acceptable because only relative length matters; see [41] for a comment about distributions with continuous density.

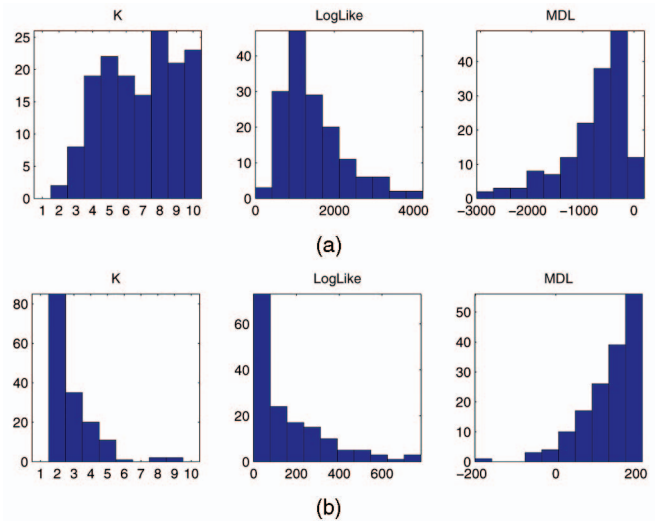

Fig. 2. Clustering analysis: Candidates from one image are more clustered than candidates randomly chosen from various images. Results on 156 images from the UWGT database. (a) MDL mixture of Gaussians for sets of candidates from the same image. (b) MDL mixture of Gaussians for sets of candidates randomly selected from different images. From left to right: The histograms of the number of Gaussians (k) that provide the lowest MDL, the histograms for the corresponding log-likelihood, and the histogram for the corresponding MDL. The analysis indicated that segments from the same image are clustered and segments from different images are scattered in the feature space.

where the statistics obtained depended on a ground truth (human) segmentation. There are differences, though, due to the difference in the desired labeling. Consider, for example, two objects of the same category present in the image as two noncontiguous segments. These segments should get the same label for attention, while for grouping, they should get different labels.

3 The Esaliency Algorithm

The proposed saliency algorithm is based on the stochastic model and the observations presented in the previous section. The algorithm is essentially a method for estimating the probability that a candidate is a target. The *Esaliency* algorithm is summarized in Fig. 3. The segmentation and the feature selection (Steps 1 and 3) are parts of the algorithm, but need not be the specific procedures we used in our experiments. The features chosen should be informative in the sense that they provide some rough but not necessarily robust distinction between the different types of objects in the scene. The distance (not necessarily euclidean) between every two vectors $d_{ij}, i, j = 1, 2, \ldots, n$, is calculated and used to obtain $\rho(l_i, l_j)$ by (1). Given the correlation matrix, the algorithm proceeds by specifying a distribution on the joint candidate labels using a Bayesian network. Then, the individual probability of every candidate to be a target is estimated using several most probable scene interpretations. These two components are described in detail below.

3.1 Specifying a Joint Label Distribution as a Bayesian Tree

The pairwise correlations, calculated from the image similarities, together with the priors $p_0(l_i)$ suggested in the previous section, can now be used to specify a joint probability distribution $p(\bar{l})$ on a binary hypothesis vector $\bar{l}$.

There are, in principle, many distributions satisfying any given set of correlations (provided the covariance matrix is positive definite). We use a simple, tree-based, Bayesian network which takes only the strongest correlations into account. This choice follows the third observation that the feature vectors associated with the candidates are clustered, rather than distributed uniformly. Together with the second observation, this means that, for strongly clustered data, all the labels within the cluster are strongly correlated and labels in different clusters are independent. Therefore, the dependence of a label of some candidate on all the other labels may be replaced by its dependence on another label in the same cluster.

Other distribution choices seem to have disadvantages. The general joint Gaussian distribution seems attractive as it is the maximum entropy distribution for the second order statistics. However, it does not model binary random variables and does not offer an efficient method for finding the most likely binary assignments, which we need. Estimating more complex distributions usually requires higher order joint statistics between the candidates' labels (e.g., [13], [47], [20]), which we do not have. Finally, but no less important, this choice has a significant computational advantage: One possible way for revealing joint assignments with high likelihood is using sampling methods such as simulated annealing or MCMC. However, they converge too slowly for an attention process. As we shall see, an efficient computation is possible with the tree-based Bayesian network.

Let G be a graph with n nodes representing the random candidate labels and edges representing the pairwise dependency between the random variables. All of the label pairs associated with nonzero correlation are thus connected. The joint distribution of all labels may be written as a function of all joint distributions associated with cliques in this graph

The Esaliency Algorithm

1) Select candidates using some segmentation process.
2) Use the preference for a small number of expected targets (and possibly other preferences) to set the initial (prior) probability for each candidate to be a target.
3) Measure visual similarity between every two candidates and infer the correlations between the corresponding labels.
4) Represent the label dependencies using a Bayesian network.
5) Find the N most likely joint assignments.
6) Deduce the saliency of each candidate by marginalization.

Fig. 3. The Esaliency algorithm.

(see, e.g., [20]). Deleting edges so that the resulting graph becomes a tree leads to a simplified distribution that depends only on second order statistics. To get the best approximation, the edges of G are weighted by the mutual information between the corresponding nodes, and the maximum weighted spanning tree of G is selected. The resulting tree describes a distribution that is closest (by the Kullback-Leibler divergence) to the original distribution (relative to all approximations by trees) [14], [38].

Calculating the mutual information,

$$I(l_i, l_j) = \sum_{l_i=0,1} \sum_{l_j=0,1} p(l_i, l_j) \log \frac{p(l_i, l_j)}{p(l_i)p(l_j)},$$

for each pair of candidates requires the probabilities of the joint events. Recall that every label l_i is a binary r.v. with expected value μ_i and variance $\mu_i(1-\mu_i)$. Then, a straightforward calculation, relying on (1), leads to the following joint probabilities:

$$\begin{aligned}
p(l_i = 1, l_p = 1) &= \gamma(d_{ij})\sqrt{\mu_i(1-\mu_i)\mu_j(1-\mu_j)} + \mu_i\mu_j, \\
p(l_i = 1, l_j = 0) &= p(l_i = 1) - p(l_i = 1, l_j = 1) \\
&= \mu_i - p(l_i = 1, l_j = 1), \\
p(l_i = 0, l_j = 1) &= \mu_j - p(l_i = 1, l_j = 1), \\
p(l_i = 0, l_j = 0) &= 1 - \mu_i - \mu_j + p(l_i = 1, l_j = 1).
\end{aligned}$$

Given $I(l_i, l_j)$ as the weights, the maximum weighted spanning tree is found by the PRIM algorithm [40]. Choosing some node of this tree as a root r makes it a directed tree (with no effect on the resulting distribution). The directed tree is converted into a Bayesian network as follows: For the root, $p(l_r = 1) = E[l_r] = \mu_r$. For each of the other nodes in the tree, two conditional probabilities should be set: $p(l_i = 1 | l_p = 0)$ and $p(l_i = 1 | l_p = 1)$, where i is the index of the node and p is the index of its parent:

$$\begin{aligned}
p(l_i = 1 | l_p = 0) &= \frac{p(l_i = 1, l_p = 0)}{p(l_p = 0)} = \frac{p(l_i = 1, l_p = 0)}{1 - \mu_p}, \\
p(l_i = 1 | l_p = 1) &= \frac{p(l_i = 1, l_p = 1)}{p(l_p = 1)} = \frac{p(l_i = 1, l_p = 1)}{\mu_p}.
\end{aligned}$$

Finally, given a vector of labels $\bar{l} = (l_1, \ldots, l_n)$, we may calculate its probability by

$$p(\bar{l}) = p(l_r) \prod_{i=1,\ldots,n; i \neq r} p\big(l_i | l_{\mathrm{par}(i)}\big), \tag{2}$$

where $\mathrm{par}(i)$ is the parent node of node i in the tree [38].

While the tree is indeed only an approximation of the true distribution, we found that it works well for the relatively clustered feature vectors in one image. In particular, the joint assignments that are associated with the "1" value assigned to the members of small tight clusters are those with the highest probabilities.

3.2 Estimating Esaliency by Marginalization over the Most Probable Scene Interpretations

We are interested in the most probable scene interpretations, as expressed by the most likely joint label vectors. Estimating saliency only from the single most likely assignment would lead to binary saliency, which is probably not the best method for locating the targets correctly. Therefore, we propose to marginalize over several (N) most likely assignments, providing high saliency values to members of small tight clusters.

With the tree-based graphical model, the N assignments associated with the highest likelihood $\mathcal{L}_{\mathrm{best}} = \{\bar{l^1}, \bar{l^2}, \ldots, \bar{l^N}\}$ are found using Nilsson's algorithm [34]. This algorithm uses exact inference to find the top N configurations and their likelihoods by a sequence of maximum propagations. For a general Bayesian network, this algorithm's efficiency depends on the number of cliques in the network, multiplied by an exponent of the cliques' size. However, for a tree-based Bayesian network, the complexity is $O(Nn\log(Nn))$ (where n is the number of nodes in the tree or, in our case, the number of candidates).

Considering only the N most likely assignments as valid interpretations, the distribution on the joint assignment vectors is

$$p'(\bar{l}) = \frac{p(\bar{l})}{\sum_{j=1}^{N} p(\bar{l^j})}. \tag{3}$$

The saliencies are then,

$$p_T(c_i) = \sum_{j=1}^{N} p'(\bar{l^j}) \cdot l_i^j. \tag{4}$$

We found that, for scenes containing 100-500 candidates, finding the 100 first most probable assignments was informative enough for directing attention to salient locations.

3.3 Some Simple Variations on the Proposed Saliency Mechanism

The proposed probabilistic model may be used to create several simple variations on the basic attention algorithm described above.

3.3.1 Global, Nonextended Saliency

One possible approach to saliency might be to look for global exceptions, that is, a single candidate that is globally unique. Note that, with this approach, we need to consider only n different label vectors for which exactly one candidate is a target ("1") and the rest are nontargets ("0"). Thus, the algorithm does not look for the N most likely hypotheses, but simply evaluates the probability for the n one-target vectors using the Bayesian network and (2). While this algorithm seems reasonable, it turns out that it is not as good as the Esaliency approach. (See Section 4.3 for experiments.)

3.3.2 Esaliency Using Learned Expected Value

A natural extension would be to use a nonfixed expected value parameter μ. We considered two versions. In the first version, μ is uniform but is set adaptively to a specific context. In the second, μ is space varying and is either estimated from training data or is just set higher in the image center according to some heuristic. See further details in Section 4.4. We found that using the value of the uniform μ learned from a training set has almost no effect on the results, suggesting that the Esaliency algorithm is not sensitive to the value of μ as long as it is small. The learned preference and the preference for

the center, however, improve the algorithm's performance in many cases.

4 Experimental Evaluation of Esaliency

This section describes a comprehensive set of experiments that illustrate the Esaliency algorithm and test it, quantitatively, on several data sets, and with respect to competing algorithms and human attention. We first discuss some implementation issues and illustrate the idea behind Esaliency with a simple synthetic example. Then, we test the algorithm on four image sets of natural outdoor scenes. We compare Esaliency's performance to that of the feature-based approach described in [26] using the iLab implementation [1]. Some variations of the Esaliency algorithm are tested as well. The computational savings from Esaliency in a complex detection task (pedestrian detection) are estimated in a separate experiment. Finally, we make a preliminary attempt to relate the Esaliency algorithm's fixation pattern to that of the human visual attention mechanism.

4.1 Implementation

All of the experiments were carried out with the same attention implementation and with the same default parameters, unless otherwise stated. The candidates are specified by a simple, fast, multiscale segmentation process [10]. We used its openCV implementation. Segments with bounding boxes whose widths and heights are between 2 and 20 percent of the image height are specified as candidates. Better segmentation algorithms may reduce search time, but could be computationally expensive.

Each candidate (segment) is characterized by a short feature vector describing some simple properties: average color (R, G, and B), dimensions of its bounding box, and its area relative to the area of its bounding box. Every feature is separately normalized to the $[0,1]$ range, and the color features are weighted by a factor of 10. This, of course, makes them dominant. The distance d_{ij} between two candidates was calculated as the weighted euclidean distance between the feature vectors, normalized so that the mean distance between pairs of feature vectors (from the candidates of the processed image) is 0.5, and clipped to the $[0,1]$ range. The correlations were calculated using these distances and the γ function described in Fig. 1b (with $D = 0.3$). The expected values μ_i were uniformly set to $\mu_i = 0.05$ in all experiments, excluding those that specifically test their influence (Section 4.4). N was set to 100 in all experiments. In all of the experiments, we evaluate the algorithm by considering the fixation path specified by the descending Esaliency values.

With the current implementation, calculating Esaliency for a 512×384 image takes, on average, about 250 ms (on a Pentium 4, with a 3 GHz processor and 1 GB memory). This is fast enough for most complex applications and is now being further improved. The application is available for download at http://cis.cs.technion.ac.il/.

4.2 A Synthetic Illustration

The first example is nonrealistic and is brought here as an illustration of the proposed attention algorithm. In this example, we consider the Esaliency assigned to the objects in an image containing 21 painted disks: 10 red, 5 blue, 3 green, 2 yellow, and 1 pink (Fig. 4a). The dependencies' maximum spanning tree, which is built out of the correlations, is shown in Fig. 4b. The Esaliency map and the attention fixation path are shown in Figs. 4c and 4d. The results are as expected—the item that appears once gets the maximal saliency and is attended to first. The saliency is a bit lower for the items that appear twice, followed by the saliency of those repeated three times, and so on.

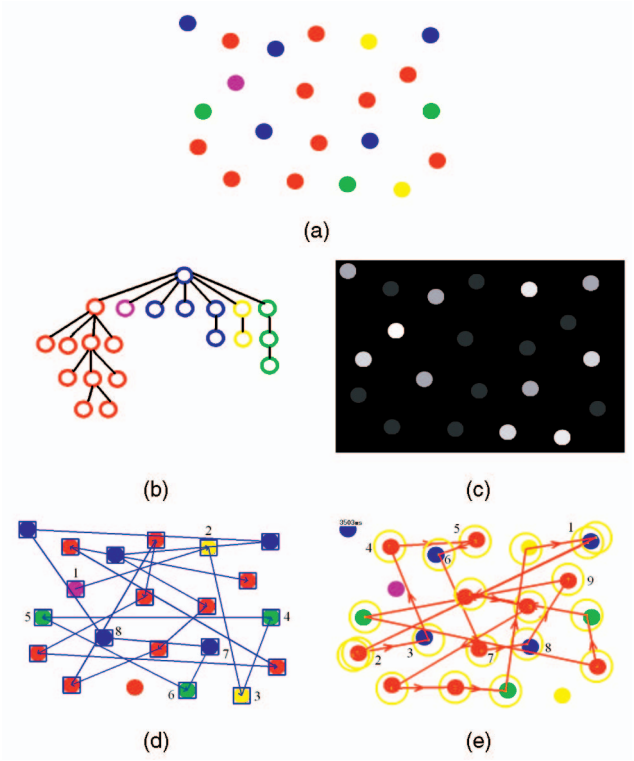

Fig. 4. Demonstrating the Esaliency algorithm on a synthetic image. (a) The input image. (b) The dependencies' spanning tree calculated by the algorithm (each node is colored according to the corresponding candidates). (c) The Esaliency map. (d) The first 20 fixations suggested by Esaliency. (e) The first 20 fixations proposed by local saliency (iLab's toolkit).) This figure should be viewed in color.

The red and the blue candidates appear several times and thus are definitely nonexceptions. Yet, their higher contrast with the white background makes them locally more salient. Indeed, running iLab's toolkit (with default parameters), we see that the yellow and pink candidates are not attended to early in the search; see Fig. 4e. In fact, due to the local approach and the inhibition of return time constants, the pink target, which is the only global exception in the image, is never attended.

4.3 Testing Esaliency on Natural Scenes

The first set of natural scenes was selected from the UWGT database [2]. This database includes many 512×768 outdoor scene images, and an annotation file describing the content of each. The 206 images containing objects such as people, cars, houses, animals, bags, signs, and boats were selected. Images containing only background items such as sky, grass, trees, clouds, streets, and rocks were not used. The attention experiments were carried out using 50 of these images. (The other 156 images were used to validate

Fig. 5. Esaliency versus iLab's toolkit for images from the UWGT database. (a) The input images. (b) The objects marked as interesting by human subjects. (c) The resulting fixation order of Esaliency marked on the segmented image, where candidates intersecting with marked targets are marked in yellow. (d) iLab's toolkit results. For both algorithms, the first 20 fixations (or a smaller number if all targets are detected earlier) are plotted. Best viewed on a color computer screen.

our basic observations in Section 2, and as a training set for setting nonuniform expected values in Section 4.4).

Two subjects, unaware of the research goal, were asked "to mark the interesting objects in each scene." These markings served as ground truth for the targets in this experiment. The average number of marked objects (targets) in an image in the test set was about 3.5. Some of the images (13) contained only one target, while the others contained between 2 and 12 targets. Some of the multiple target images included targets from the same category and some from different categories; see the two leftmost columns in Fig. 5 for a sample of the images and the corresponding marked targets.

We ran Esaliency with the default parameters. The candidates (segments) were scanned in descending order of saliency. Let m be the number of candidates scanned until all targets are detected. A target was considered as detected when an attended candidate segment intersected with the corresponding marked region. Fig. 5c shows the scan path associated with the first $\min(m, 20)$-scanned candidates.

Some statistics of the search task results are summarized in the left column of Table 1. Note that the search mechanism is very efficient: Only a few candidates were falsely visited on the way to detecting the true targets. With our implementation, every image contained an average of 330 segments. This means that only a small fraction of the image was scanned.

We then applied the local saliency model to the same 50 images using iLab's toolkit (with its default parameters).

TABLE 1
Results for Esaliency and Nonextended Global Saliency (Section 3.3.1) on 50 Images from the UWGT Database [2]

	Esaliency	Non-extended global saliency
False detections before 1st target detected	$0.7 \pm 1.5; 0$	$1.8 \pm 3.7; 0$
False detections per target before 50% of targets detected	$0.88 \pm 1.5; 0$	$2.3 \pm 3.9; 0.6$
False detections per target before 75% of targets detected	$1.7 \pm 2.3; 0.67$	$3.1 \pm 3.6; 2$
False detections per target before all targets detected	$2.5 \pm 3.0; 1$	$3.7 \pm 3.7; 3$
False detections before 50% of targets detected	$1.6 \pm 2.7; 0$	$3.4 \pm 5.6; 1.5$
False detections before 75% of targets detected	$5.5 \pm 8.9; 2$	$8.4 \pm 10.1; 4.5$
False detections before all targets detected	$12.2 \pm 24.4; 3$	$14.1 \pm 18.8; 7.5$

Mean, standard deviations (std), and median are reported. There are 1-12 targets per image, with mean and std 3.5 ± 2.8.

Some of the resulting scan paths are demonstrated in Fig. 5d. Many targets are efficiently detected, but some problems are apparent. In the third image, for example, the sky patches between the trees are indeed locally salient and are selected, by the local saliency process, long before the pedestrians. The proposed algorithm, however, is able to take advantage of the larger number of sky patches, reducing their Esaliency and focusing attention on the pedestrians earlier.

We observed that about 30 percent of the targets were not detected by the local saliency algorithm even after a very large number of fixations. Therefore, comparing the results by average detection attempts over all targets (Table 1) was meaningless. We compared the algorithms by plotting the number of fixations required for each target to be found (Fig. 6). Note the advantage of the proposed

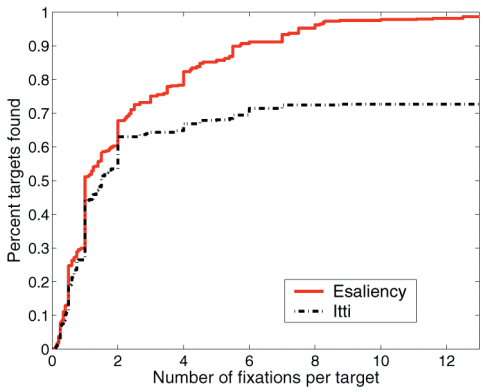

Fig. 6. Comparing Esaliency and iLab on images from the UWGT database: the target detection ratio as a function of the number of fixations per target.

algorithm both in finding all the targets and in finding more targets for the same number of fixations. It seems that some of the misses of the local saliency algorithm are due to worse priority specification (see Fig. 5), and others are due to the inhibition-of-return mechanism and its noninformative time constant (that is inevitable in the biology-motivated space-based design).

We also experimented with the global, nonextended saliency algorithm (Section 3.3.1). The results are summarized in Table 1 (right-hand column). Because of the strict requirement for uniqueness, which is not consistent with many images, the results are not as good as those of the Esaliency algorithm. Although global saliency also yields impressive results for many images, it fails when there are a few salient regions with similar appearance (associated with different targets or with the same target); see Fig. 7 for examples.

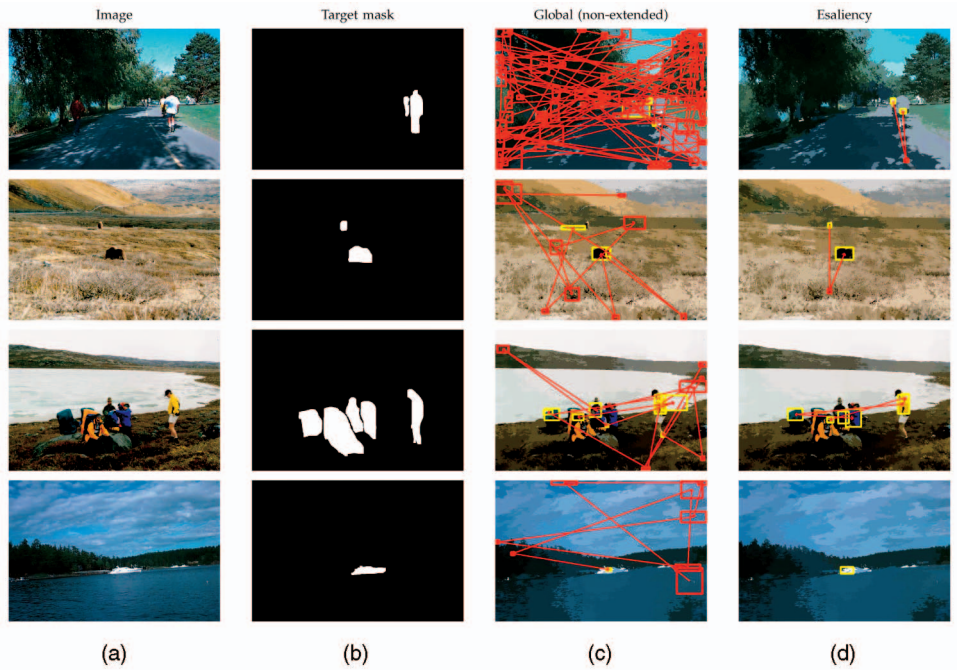

Fig. 7. Some additional examples for Esaliency compared with nonextended global saliency. (a) Input images. (b) The objects marked as interesting by human subjects. (c) The fixation set by nonextended global saliency. (d) The fixation order specified by Esaliency. Candidates intersecting with marked targets are marked in yellow, others in red. Best viewed on a color computer screen.

TABLE 2
Searching for Locally Salient Objects
(Red Can, Emergency Triangle, and Traffic Signs)

	False detection before targets found by Itti et al.	False detection before targets found by Esaliency
Red can	1.67 ± 2.01	$2.95 \pm 3.36; 2$
Emergency triangle	1.69 ± 2.28	$1.25 \pm 1.60; 1$
Traffic signs first	0.49 ± 1.06	$0.69 \pm 1.28; 0$
Traffic signs all	1.27 ± 2.12	$2.09 \pm 2.12; 1$

Comparing results of Esaliency and results reported in Itti et al. [25]. The mean, standard deviations, and median number of false detections before the targets are detected are reported. Some images from the traffic sign database include more than one sign (while the other databases always include one target per image).

The proposed Esaliency algorithm is not necessarily better than local saliency for all tasks. We considered the tasks reported in [25], where (variations of the) iLab algorithm were tested in detecting red cans, traffic signs, and emergency triangles. Table 2 compares the results for the three tasks using the default-normalized iLab algorithm and the proposed Esaliency algorithm. The images in the red can and triangle data sets are 640×480 pixels, while the traffic sign images are 512×384 pixels. As in [25], an image fixation is considered successful if some part of the target object is inside the circle centered at the chosen fixation point (center of selected segment) with radius 80 for the two first data sets and radius 64 for the third data set. (Note that, in other experiments, an image fixation was considered successful only if the selected segment intersected with a marked target.)

The Esaliency algorithm performed somewhat better for the emergency triangle task, and somewhat worse for the red cans and traffic signs. Note that these objects are designed to be *locally* salient, either for safety or for commercial reasons. In the latter data set, Esaliency fails mostly on the roadside light reflectors, which are discriminable by their dominant oblique orientation. We found that the image fixations that preceded fixation on the target light reflectors were always on other interesting objects in the scene.

4.4 Esaliency on Natural Scenes with Nonfixed Expected Value

We further tested how other preferences derived from human behavior or soft learning affect Esaliency. First, we tested whether setting the expected value parameter μ according to the context makes a difference. To specify μ, we used training sets of natural images and corresponding (manually obtained) binary maps of target locations. μ was uniformly set as the average fraction of target pixels. The μ values were 0.024, 0.015, 0.013, and 0.037 for the UWGT database, the red cans, the triangles, and the traffic signs, respectively. See Table 3 (second column). Clearly, the trained μ has almost no effect for all four data sets, suggesting that the Esaliency algorithm is not sensitive to the value of μ as long as it is small.

We then tested the effect of a nonuniform expected value. Using the known eccentricity effect in human vision [11], [60], we specified μ_i for the ith candidate by following an exponentially decreasing function which is maximal in the image center and is lower by a factor of e^{-1} in all corners. This function was normalized so that its average value is the default value $\mu = 0.05$; see Fig. 8a. The prior μ_i for the ith candidate is set to be the maximum value of the map in it. The performance was improved for all data sets; see Table 3, third column. We verified that the improvement cannot be explained simply by ordering the candidates according to the varying expected values; see the rightmost column of Table 3.

We also experimented with space varying, context-dependent expected values. A priority map was created by averaging the binary maps of target locations, separately for each training set, and then, smoothing them. See Figs. 8b, 8c, 8d, and 8e. All the priority maps show a preference for the center, reflecting the tendency of people to center their photography objects. Here as well, μ_i is set to the maximum value of the map inside the corresponding candidate region. Note that, while the median number of fixations never increases, their average number may be higher in some cases; see Table 3, fourth column. This is probably due to candidates being in atypical locations not represented in the training set.

TABLE 3
The Performance Associated with Several Versions of Esaliency, Described in Section 4.4, as Well as Some Reference Figures
(Two Bottom Rows)

	Esaliency default	Esaliency with trained μ	Esaliency with center preference	Esaliency with trained locations	Candidates in random order	Cand. ordered only by center pref.
Washington first	$0.7 \pm 1.5; 0$	$0.78 \pm 1.9; 0$	$0.36 \pm 1.0; 0$	$0.48 \pm 1.4; 0$	$8.7 \pm 10.1; 5$	$14.1 \pm 27.4; 0$
Washington all	$12.2 \pm 24.4; 3$	$12.4 \pm 24.9; 3$	$8.6 \pm 13.5; 2$	$12.6 \pm 24.6; 3$	$50.6 \pm 58.9; 27.5$	$62.7 \pm 59.6; 50.5$
Red cans	$2.9 \pm 3.4; 2$	$2.9 \pm 3.2; 2$	$1.2 \pm 2.6; 1$	$1.1 \pm 2.5; 1$	$6.8 \pm 9.1; 4$	$2.5 \pm 7.9; 0$
Triangles	$1.2 \pm 1.6; 1$	$1.2 \pm 1.6; 1$	$0.59 \pm 0.93; 0$	$0.75 \pm 1.2; 0$	$8.9 \pm 9.3; 5.5$	$8.2 \pm 15.4; 1$
Traffic signs first	$0.69 \pm 1.3; 0$	$0.67 \pm 1.3; 0$	$0.62 \pm 1.2; 0$	$0.82 \pm 1.7; 0$	$4.1 \pm 5.4; 2$	$11 \pm 18.7; 0$
Traffic signs all	$2.1 \pm 2.1; 1$	$2.1 \pm 2.3; 1$	$1.2 \pm 1.8; 0$	$1.7 \pm 2.3; 1$	$9.2 \pm 9; 6$	$29.9 \pm 42.5; 15$

The mean, standard deviations, and median number of false detections before the targets are detected are reported.

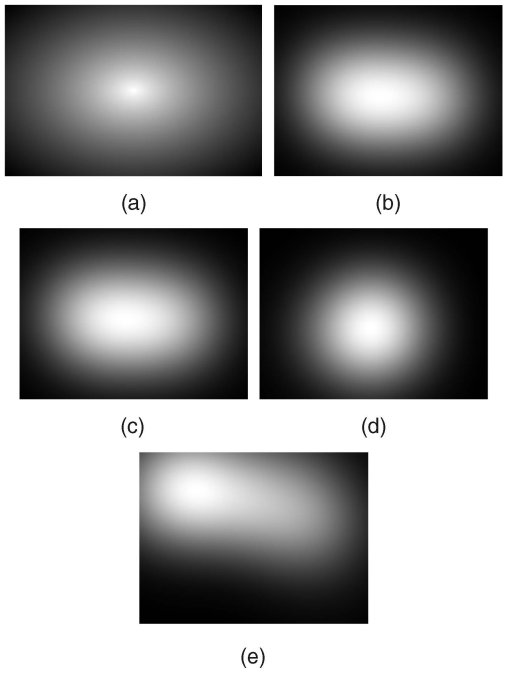

Fig. 8. Target location priority maps. The map (a) specifies a preference for objects that are closer to the center. The other four maps are learned from training sets associated with the four data sets described in Section 4.4. (b) UWGT. (c) Red cans. (d) Triangle. (e) Traffic signs.

4.5 Using Esaliency to Accelerate Pedestrian Detection

We next turned to test the Esaliency algorithm in the context of pedestrian detection. We used the MIT StreetScenes database [7] containing 3,547 images of urban scenes. The locations of objects from nine categories—cars, pedestrians, bicycles, buildings, trees, roads, sky, sidewalks, and stores—are annotated in the images. We focused on the task of detecting pedestrians. Esaliency (using default parameters) was applied to all (852) images that contained marked pedestrians. (The images were downsampled from $1{,}280 \times 960$ to 640×480). The mean number of false fixations before locating the first pedestrian was 27. The median was 12. The mean and the median numbers of false fixations before all pedestrians were located were 41.55 and 21, respectively; see Fig. 9.

To estimate the computational savings, we considered a simple model of the search process. Without attention, a common detection mechanism (e.g., [55]) evaluates subimages. Suppose that the detection process starts from the upper left corner of an $h \times w$ image, and scans the image in raster scan with a 20×30 window, jumping in steps of two pixels. This is repeated for, say, six scales (to detect pedestrians of different sizes), each time for an image 1.5 times smaller in both dimensions. The total number of windows checked is $T = \sum_{i=0}^{5} \frac{(\mathrm{h}/1.5^i - 30)(\mathrm{w}/1.5^i - 20)}{4}$. For $\mathrm{w} \times \mathrm{h} = 1{,}280 \times 960$, $T = 5.1019 \times 10^5$. Consider an alternative scenario, where Esaliency sets the order of fixations. For each fixation, we let the detection algorithm checks all the windows that include the center fixation point in all six scales. For k fixations, the number of checked windows is $T' = k\frac{20 \times 30}{4} \times 6 = 900k$. Hence, $\frac{T'}{T} = 0.0018k$. That is, with Esaliency, the median number of windows tested before all pedestrians are detected is just $\frac{T'}{T} = 0.0378$, or less than 4 percent of the windows used in the sequential scan. (Note that the estimate is conservative: We could reduce the number of tested windows, for example, by taking only those that contain all or most of the segment.)

Note that the results for the StreetScenes database are not as good as those obtained for the images from the UWGT database. One obvious reason is that, in the UWGT database images, all "interesting objects" were marked and considered to be targets. Here, only pedestrians are considered targets, and other salient attention-attracting objects, such as cars and signs, are considered to be nontargets. Besides, most images in the UWGT database are of natural scenes, and are less crowded with nontarget salient objects. Actually, some of the pedestrians in the StreetScenes images were not marked, and therefore, although pedestrians are sometimes found earlier, this is not counted as a success (see, e.g., last image in Fig. 9).

Nevertheless, as shown above, using Esaliency undoubtedly makes the detection more efficient. We emphasize that the proposed bottom-up Esaliency algorithm does not use any knowledge about the pedestrians. The results could have been improved by adding such knowledge, by, say, locating sidewalks or zebra crossings first, or by optimizing the segmentation or the feature selection for the pedestrian context.

4.6 Esaliency versus Human Eye Fixations

The main goal of this paper was to develop an efficient and quantitatively reasoned attention process for computer vision. Yet it is tempting to relate the sequence of fixations associated with the proposed approach to that obtained by the human visual attention mechanism. Moreover, the similarity of the search pattern to that of the human search pattern, usually regarded as the gold standard, supports the proposed algorithm. We describe here a preliminary study of this relation.

We follow the approach proposed in [35], which evaluates an attention model by comparing the saliency map proposed by that model to an empirical *human saliency map*. The latter is constructed by recording human eye fixations over an (limited time displayed) image, and convolving each fixation point in the image with a Gaussian. Averaging the *human saliency maps* obtained from several (seven or more) subjects yields a *mean human saliency map* for each image. The maps are intensity normalized and downsampled (from 512×384 to 32×24) so that they can be easily compared to the saliency maps of the iLab algorithm [26], available in [1]. See Fig. 10. The correlation coefficient between two saliency maps serves as a quantitative scalar measure of similarity. See the upper rows of Table 4 for the correlation between Itti's model and the empirical *mean human saliency map*, as reported in [35].

To evaluate Esaliency, we ran it on the same images and created saliency maps in a similar way. Esaliency's "fixations" were specified as the centers of the attended candidate regions. See Fig. 10. See also the correlation coefficients between the resulting *Esaliency maps* and the human saliency maps in Table 4.

All correlations between the mean human saliency maps and those of the computational models are significant

Fig. 9. Examples of Esaliency's performance when locating pedestrians in the StreetScenes images. Best viewed on a color computer screen.

(p-value $\ll 0.05$). The correlations between the mean human maps and the Esaliency maps are higher than the correlations with iLab saliency in five out of six images, implying a somewhat better agreement.

5 Discussion

In this paper, we proposed a new, region-based bottom-up saliency measure. This measure is the (approximate) probability of an image region to be salient, estimated from preferences on the number of objects of interest in the scene, and from validated stochastic modeling of the likely target assignments. This quantitative approach differs from the traditional feature-based (or space-based) methods. Moreover, the resulting *Esaliency* estimates are based on global considerations, which are more justified.

We have validated our saliency measure using a variety of image databases. Esaliency's fixation path was validated by comparing it to human fixation paths, by comparing it to human selection of "interesting objects," and for object detection tasks. We compared Esaliency's performance to that of the dominant model in computerized visual attention [26], and showed similar or better results. We found that the proposed method is fast, reliable, and performs better in complex, cluttered scenes.

We believe that the approach proposed here may serve as a solid foundation for other search and detection tasks. Here,

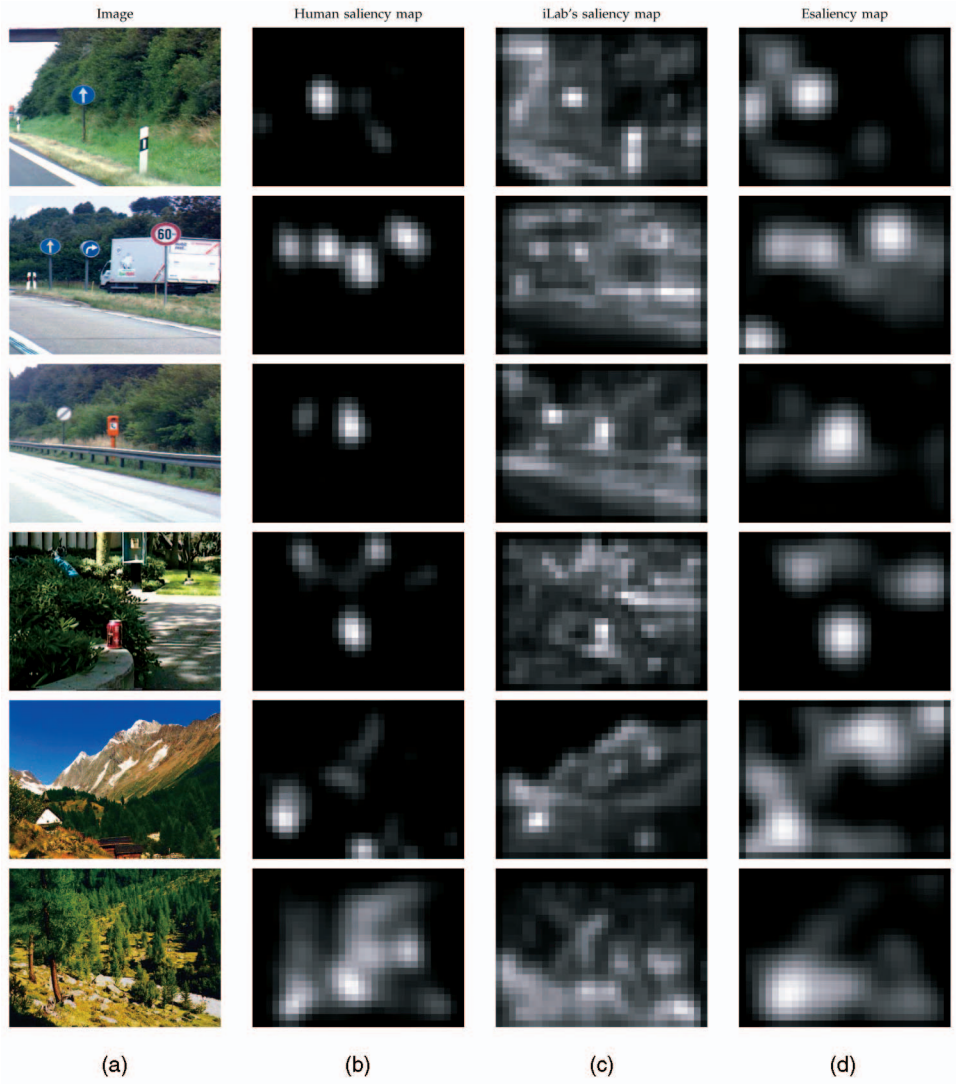

Fig. 10. Comparing computer saliency models and *human saliency maps*. (a) The input images. (b) The *mean human saliency maps* created from recording human eye fixations [35]. (c) iLab [26] saliency maps. (d) Esaliency saliency maps.

we focused on a pure bottom-up approach. For specific applications, top-down information is available and may be integrated naturally in the framework described here. This can be done by adapting the segmentation and the similarity measures to the specific context, by setting the expected μ_i values according to the candidate properties and the context (see, e.g., [50]), or even by integrating a categorizing mechanism into the search itself, making it more efficient by changing the priorities dynamically (see [5]).

Finally, to the best of our knowledge, the mechanism described here is the first quantitative and practical model for object-based attention, i.e., an attention process that assigns priorities to structural units that are the result of a perceptual

TABLE 4
Correlation Coefficients between *Human Saliency Maps* [35] and Computational Model Saliency Maps

(a) iLab vs. human

	road1	road 2	road 3	coke	swissalps	forest
All subjects	**0.232**	**0.3624**	**0.482**	**0.4**	**0.523**	**0.436**
Best subject	0.321	0.348	0.462	0.45	0.608	0.477
Worst subject	0.079	0.194	0.082	0.154	0.134	-0.078

(b) Esaliency vs. human

	road1	road 2	road 3	coke	swissalps	forest
All subjects	**0.579**	**0.709**	**0.647**	**0.671**	**0.371**	**0.615**
Best subject	0.616	0.741	0.715	0.681	0.514	0.476
Worst subject	0.305	0.381	0.217	0.322	0.059	-0.029

(a) Correlation coefficients between human saliency maps and iLab saliency maps. (b) Correlation coefficients between human saliency maps and Esaliency's saliency maps.

organization preprocessing stage. We intend to research whether it may indeed explain some aspects of human perception. The preliminary saliency comparisons described in Section 4.6 are encouraging. Our mechanism's correspondence to eye fixations can be further improved by incorporating known properties of human visual search, such as preference for the image center (Section 4.4), preference for short interfixation distance (proximity) [28], and the inhibition of locations that are similar to already attended locations, as suggested in [19] and as modeled, e.g., in [6].

ACKNOWLEDGMENTS

The authors would like to thank Nabil Ouerhani for the human saliency maps [35], Robert Cowell [3] for sharing the XBAIES software package and for his advice, and Stanley Bileschi for sharing the StreetScenes database. This work was supported by the Israeli Science Foundation (ISF) and by the European Commission (EC) network of excellence, MUSCLE.

REFERENCES

[1] http://ilab.usc.edu/toolkit/downloads.shtml, 2009.
[2] http://www.cs.washington.edu/research/imagedatabase/groundtruth/, 2009.
[3] http://www.staff.city.ac.uk/rgc/software.html. 2000, 2009.
[4] T. Avraham and M. Lindenbaum, "Esaliency—A Stochastic Attention Model Incorporating Similarity Information and Knowledge-Based Preferences," *Proc. Int'l Workshop Representation and Use of Prior Knowledge in Vision, with European Conf. Computer Vision,* 2006.
[5] T. Avraham and M. Lindenbaum, "Dynamic Visual Search Using Inner Scene Similarity—Algorithms and Bounds," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 2, pp. 151-264, Feb. 2006.
[6] T. Avraham, Y. Yeshurun, and M. Lindenbaum, "Predicting Visual Search Performance by Quantifying Stimuli Similarities," *J. Vision,* vol. 8, no. 4, pp. 1-22, 2008.
[7] S. Bileschi, "StreetScenes: Towards Scene Understanding in Still Images," PhD thesis, Electrical Eng. and Computer Science Dept., Massachusetts Inst. of Technology, May 2006.
[8] O. Boiman and M. Irani, "Detecting Irregularities in Images and Video," *Proc. 10th Int'l Conf. Computer Vision,* 2005.
[9] N. Bruce and J.K. Tsotsos, "Saliency Based on Information Maximization," *Advances in Neural Information Processing Systems,* vol. 18, pp. 155-162, MIT Press, 2006.
[10] P.J. Burt, T.H. Hong, and A. Rosenfeld, "Segmentation and Estimation of Image Region Properties through Cooperative Hierarchical Computation," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 11, no. 12, pp. 802-809, Dec. 1981.
[11] M. Carrasco, D.L. Evert, I. Chang, and S.M. Katz, "The Eccentricity Effect: Target Eccentricity Affects Performance on Conjunction Searches," *Perception and Psychophysics,* vol. 57, no. 8, pp. 1241-1261, 1995.
[12] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image Segmentation Using Expectation-Maximization and Its Applications to Image Querying," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 8, pp. 1026-1038, Aug. 2002.
[13] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu, "Learning Bayesian Networks from Data: An Information-Theory Based Approach," *Artificial Intelligence,* vol. 137, pp. 43-90, 2002.
[14] C.K. Chow and C.N. Liu, "Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Trans. Information Theory,* vol. 14, no. 11, pp. 462-467, Nov. 1968.
[15] A. Cohen and R.B. Ivry, "Density Effects in Conjunction Search: Evidence for Coarse Location Mechanism of Feature Integration," *J. Experimental Psychology: Human Perception and Performance,* vol. 17, no. 4, pp. 891-901, 1991.
[16] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc. Series B,* vol. 39, no. 1, pp. 1-38, 1977.
[17] B.A. Draper and A. Lionelle, "Evaluation of Selective Attention under Similarity Transformations," *Computer Vision and Image Understanding,* vol. 100, nos. 1/2, pp. 152-171, 2005.
[18] J. Duncan, "Selective Attention and the Organization of Visual Information," *J. Experimental Psychology: General,* vol. 113, pp. 501-517, 1984.
[19] J. Duncan and G.W. Humphreys, "Visual Search and Stimulus Similarity," *Psychological Rev.,* vol. 96, pp. 433-458, 1989.
[20] *Learning in Graphical Models,* M.I. Jordan, ed. Kluwer Academic, 1998.
[21] C.W. Eriksen and J.D.S. James, "Visual Attention within and around the Field of Focal Attention: A Zoom Lens Model," *Perception and Psychophysics,* vol. 40, no. 4, pp. 225-240, 1986.
[22] K.H. Fecteau and D.P. Munoz, "Salience, Relevance, and Firing: A Priority Map for Target Selection," *Trends in Cognitive Sciences,* vol. 10, no. 8, pp. 382-390, 2006.
[23] S. Frintrop, A. Nüchter, and H. Surmann, "Visual Attention for Object Recognition in Spatial 3D Data," *Proc. Second Int'l Workshop Attention and Performance in Computational Vision,* 2005.
[24] L. Paletta, H. Bischof, G. Fritz, and C. Seifert, "Entropy Based Saliency Maps for Object Recognition," *Proc. Early Cognitive Vision Workshop,* 2004.
[25] L. Itti and C. Koch, "Feature Combination Strategies for Saliency-Based Visual Attention Systems," *J. Electronic Imaging,* vol. 10, no. 1, pp. 161-169, 2001.
[26] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.
[27] T. Kadir and M. Brady, "Saliency, Scale and Image Description," *Int'l J. Computer Vision,* vol. 45, no. 2, pp. 83-105, 2001.
[28] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology,* vol. 4, pp. 219-227, 1985.
[29] F. Liu and M. Gleicher, "Region Enhanced Scale-Invariant Saliency Detection," *Proc. IEEE Int'l Conf. Multimedia and Expo,* 2006.
[30] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics," *Proc. Eighth Int'l Conf. Computer Vision,* vol. 2, pp. 416-423, July 2001.
[31] K. Nakayama and G.H. Silverman, "Serial and Parallel Processing Visual Feature Conjunction," *Nature,* vol. 320, pp. 264-265, 1986.
[32] V. Navalpakkam and L. Itti, "A Goal Oriented Attention Guidance Model," *Lecture Notes in Computer Science,* pp. 453-461, Springer, 2002.
[33] U. Neisser, *Cognitive Psychology.* Appleton-Century-Crofts, 1967.
[34] D. Nilsson, "An Efficient Algorithm for Finding the M Most Probable Configurations in Probabilistic Expert Systems," *Statistics and Computing,* vol. 8, no. 2, pp. 159-173, 1998.
[35] N. Ouerhani, R. von Wartburg, H. Hügli, and R.M. Müri, "Empirical Validation of Saliency-Based Model of Visual Attention," *Electronic Letters on Computer Vision and Image Analysis,* vol. 3, no. 1, pp. 13-24, 2004.
[36] S. Palmer and I. Rock, "Rethinking Perceptual Organization: The Role of Uniform Connectedness," *Psychonomic Bull. and Rev.,* vol. 1, no. 1, pp. 29-55, 1994.
[37] D. Parkhurst and E. Niebur, "What Could over 1000 Internet Users Tell Us about Visual Attention?" *J. Vision,* vol. 3, no. 9, 2003.
[38] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1998.
[39] M.I. Posner, C.R.R. Snyder, and B.J. Davidson, "Attention and the Detection of Signals," *J. Experimental Psychology: General,* vol. 109, no. 2, pp. 160-174, June 1980.
[40] R.C. Prim, "Shortest Connection Networks and Some Generalizations," *Bell System Technical J.,* vol. 36, pp. 1389-1401, 1957.
[41] J. Rissanen, "Modeling by Shortest Data Description," *Automatica,* vol. 14, pp. 465-471, 1978.
[42] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is Bottom-Up Attention Useful for Object Recognition?" *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 37-44, 2004.
[43] C. Schmid and J. Frederic, "Scale-Invariant Shape Features for Recognition of Object Categories," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 90-96, 2004.
[44] B.J. Scholl, "Objects and Attention: The State of the Art," *Cognition,* vol. 80, pp. 1-46, 2001.

[45] A. Shashua and S. Ullman, "Structural Saliency: The Detection of Globally Salient Structures Using a Locally Connected Network," *Proc. Int'l Conf. Computer Vision,* pp. 321-327, 1988.
[46] C. Siagian and L. Itti, "Biologically-Inspired Face Detection: Non-Brute-Force-Search Approach," *Proc. First IEEE Int'l Workshop Face Processing in Video,* June 2004.
[47] P. Spirtes and C. Glymour, "An Algorithm for Fast Recovery of Sparse Causal Graphs," *Social Science Computer Rev.,* vol. 9, no. 1, pp. 62-72, 1991.
[48] A. Srivastava, A.B. Lee, E.P. Simoncelli, and S.C. Zhu, "On Advances in Statistical Modeling of Natural Images," *J. Math. Imaging and Vision,* vol. 18, pp. 17-33, 2003.
[49] Y. Sun and R. Fisher, "Object-Based Attention for Computer Vision," *Artificial Intelligence,* vol. 146, pp. 77-123, 2003.
[50] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual Guidance of Eye Movements and Attention in Real-World Scenes: The Role of Global Features on Object Search," *Psychological Rev.,* vol. 113, no. 4, pp. 766-786, 2006.
[51] A. Treisman, "Features and Objects: The 14th Barlett Memorial Lecture," *Quarterly J. Experimental Psychology,* vol. 40A, pp. 201-237, 1998.
[52] A. Treisman and G. Gelade, "A Feature Integration Theory of Attention," *Cognitive Psychology,* vol. 12, pp. 97-136, 1980.
[53] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, and F.J. Nuflo, "Modeling Visual Attention via Selective Tuning," *Artificial Intelligence,* vol. 78, nos. 1/2, pp. 507-545, 1995.
[54] A. Vailaya, M.A.T. Figueiredo, A.K. Jain, and H.J. Zhang, "Image Classification for Content-Based Indexing," *IEEE Trans. Image Processing,* vol. 10, no. 1, pp. 117-130, Jan. 2001.
[55] P. Viola and M.J. Jones, "Robust Real-Time Face Detection," *Int'l J. Computer Vision,* vol. 57, no. 2, pp. 137-154, May 2004.
[56] J. Vogel and B. Schiele, "A Semantic Typicality Measure for Natural Scene Categorization," *Proc. German Assoc. for Pattern Recognition Symp.,* pp. 195-203, 2004.
[57] K.N. Walker, T.F. Cootes, and C.J. Taylor, "Correspondence Using Distinct Points Based on Image Invariants," *Proc. British Machine Vision Conf.,* pp. 540-549, 1997.
[58] S. Watson and A. Kramer, "Object-Based Visual Selective Attention and Perceptual Organization," *Perception and Psychophysics,* vol. 61, pp. 31-49, 1999.
[59] M. Wertheimer, "Untersuchungen Zur Lehre Von Der Gestalt," *Psychologishe Forschung,* vol. 4, pp. 301-350, 1923.
[60] G. Westheimer, "Visual Acuity," *Adler's Physiology of the Eye, Clinical Application,* R.A. Moses and W.M. Hart, eds., chapter 17, The C.V. Mosby Company, 1987.
[61] J.M. Wolfe, "Guided Search 2.0: A Revised Model of Visual Search," *Psychonomic Bull. and Rev.,* vol. 1, no. 2, pp. 202-238, 1994.
[62] J.M. Wolfe, "Visual Search," *Attention,* H. Pashler, ed., Psychology Press, 1998.
[63] A.L. Yarbus, *Eye Movements and Vision.* Plenum Press, 1967.

Tamar Avraham received the BSc (summa cum laude) and the PhD degrees from the Computer Science Department at the Technion-Israel Institute of Technology in 1996 and 2008, respectively. She worked for several years in industry (Fibronics, Ltd., and CMT Medical Technologies, Ltd.) as a software engineer, a project manager, and a product manager. Currently, she is a research fellow at the Technion Research and Development Foundation. Her main research interest is in computer vision, with a focus on visual attention, scene analysis, statistics of natural images, and computer graphics applications that incorporate image and video analysis.

Michael Lindenbaum received the BSc, MSc, and DSc degrees from the Department of Electrical Engineering at the Technion-Israel Institute of Technology, in 1978, 1987, and 1990, respectively. From 1978 to 1985, he served in the IDF. He did his postdoctoral research at the NTT Basic Research Labs in Tokyo, Japan, and since 1991, he has been with the Department of Computer Science at the Technion. He was also a consultant to HP Labs, Israel, and spent a sabbatical in NECI, in 2001. He worked in digital geometry, computational robotics, learning, and various aspects of computer vision and image processing. Currently, his main research interest is computer vision, and especially statistical analysis of object recognition and grouping processes. He is a member of the IEEE.