Saliency and Gist Features for Target Detection in Satellite Images

Zhicheng Li and Laurent Itti

***Abstract*—Reliably detecting objects in broad-area overhead or satellite images has become an increasingly pressing need, as the capabilities for image acquisition are growing rapidly. The problem is particularly difficult in the presence of large intraclass variability, e.g., finding "boats" or "buildings," where model-based approaches tend to fail because no good model or template can be defined for the highly variable targets. This paper explores an automatic approach to detect and classify targets in high-resolution broad-area satellite images, which relies on detecting statistical signatures of targets, in terms of a set of biologically-inspired low-level visual features. Broad-area images are cut into small image chips, analyzed in two complementary ways: "attention/saliency" analysis exploits local features and their interactions across space, while "gist" analysis focuses on global nonspatial features and their statistics. Both feature sets are used to classify each chip as containing target(s) or not, using a support vector machine. Four experiments were performed to find "boats" (Experiments 1 and 2), "buildings" (Experiment 3) and "airplanes" (Experiment 4). In experiment 1, 14 416 image chips were randomly divided into training (300 boat, 300 nonboat) and test sets (13 816), and classification was performed on the test set (ROC area: 0.977 ± 0.003). In experiment 2, classification was performed on another test set of 11 385 chips from another broad-area image, keeping the same training set as in experiment 1 (ROC area: 0.952 ± 0.006). In experiment 3, 600 training chips (300 for each type) were randomly selected from 108 885 chips, and classification was conducted (ROC area: 0.922 ± 0.005). In experiment 4, 20 training chips (10 for each type) were randomly selected to classify the remaining 2581 chips (ROC area: 0.976 ± 0.003). The proposed algorithm outperformed the state-of-the-art SIFT, HMAX, and hidden-scale salient structure methods, and previous gist-only features in all four experiments. This study shows that the proposed target search method can reliably and effectively detect highly variable target objects in large image datasets.**

***Index Terms*—Gist features, saliency features, satellite images, target detection.**

I. Introduction

OVERHEAD and satellite imagery have become ubiquitous, with applications ranging from intelligence gathering to consumer mapping and navigation assistance. With the overwhelming amount of satellite imagery available today, it has become impossible for human image analysts to examine all of the imagery, in search of interesting intelligence information. Thus, there is a pressing need for automatic algorithms to preprocess the data and to extract actionable intelligence from raw imagery, thereby facilitating and supporting human interpretation. This paper focuses on automatically detecting diverse types of targets with large intraclass variability in satellite images. This analysis is one of the currently highly time-consuming tasks that image analysts routinely perform manually. Providing new means to automate this task is expected to facilitate and render more efficient the interpretation of satellite image by human analysts.

The problem of target detection is a difficult challenge in computer vision [1]–[3]. For a given scene (image), the target detection task can be simply described as "where is the target?" Considering the feature types used for detection in static images, algorithms for target detection can be briefly summarized as belonging to three broad categories: A first, relatively straightforward approach is to use a provided (or trained) target template or model (hence, the feature is the image itself), to match against targets in the image of interest, at different locations, orientation and scales [4]–[6]. This type of method works well when the variability of targets is small (for example, detecting human faces [5], [6]). A second method for target detection is to use a model to extract a spatially sparse collection of invariant structural features (e.g., keypoint descriptors, bags of features) of the target even when viewpoint, pose, and lighting conditions vary [7]–[10]. In a third approach, using knowledge of target shape and characteristic geometry, several studies have proposed methods which learn and apply target geometric constraints on the keypoint feature locations [11], [12]. In practice, the detection algorithms usually overlap these categories, and some approaches are intermediate between the geometry-based and "bag of features" approaches retaining only some coarsely-coded location information or recording the locations of features relative to the target's center [3], [13]. In addition to these machine vision approaches, several biologically-inspired computational models have also started exploring target detection tasks in imagery, usually based on our knowledge of visual cortex, showing some promising experimental results [14]–[19]. Our approach extends these biologically-inspired frameworks.

Based on the special properties of satellite image, several algorithms have been proposed to detect the targets in such kind

Manuscript received August 05, 2010; revised November 19, 2010; accepted November 24, 2010. Date of publication December 13, 2010; date of current version June 17, 2011. This work was supported by Defense Advanced Research Projects Agency under Government Contract HR0011-10-C-0034, the National Geospatial Intelligence Agency under Grant 19-1082141, the National Science Foundation under CRCNS Grant BCS-0827764, the Army Research Office under Grant W911NF-08-1-0360, and the China Scholarship Council under Grant 2007103281. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kenneth K. M. Lam.

Z. Li was with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191 China. He is now with the Computer Science Department, University of Southern California, Los Angeles, CA 90089 USA (e-mail: lzcbuaa@gmail.com).

L. Itti is with the Computer Science Department, University of Southern California, Los Angeles, CA 90089 USA (e-mail: itti@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2010.2099128

of images. For example, for hyper-spectral satellite images, the features applied usually take advantage of the reflection characteristics of different materials [20]–[22] while for multispectral images, the features are usually extracted from fused spectra [23], [24]. However, the images discussed in this paper focus on the visible spectrum and, thus, the detection methods discussed in the previous paragraph are usually adopted. Despite all the recent advances in computer vision technologies, humans still perform orders of magnitude better than the best available vision systems in object and target detection, and for many target search applications humans remain the gold standard. As such, it is reasonable to examine the low-level mechanisms as well as the system-level computational architecture of human vision for inspiration. Early on, the human visual processing system already makes decisions to focus attention and processing resources onto those small regions within the field of view which look more interesting or visually "salient" [25]–[27]. When no specific search target, no search task, and no particular time or other constraint are specified to an observer, bottom-up (image-derived) information may play a predominant role in guiding attention toward potential generically interesting targets [28]. The mechanism of selecting a small set of candidate salient locations in a scene has recently been the subject of comprehensive research efforts and several computational models have been proposed [29]–[34]. One can make use of these models to predict possible target locations and target distributions. In this paper, saliency maps from several feature channels (intensity contrast, local edge orientation, etc.) are computed from a modified Itti-Koch saliency model [25], [31], [35]. Given a static or dynamic visual scene, this model creates a number of multiscale topographic feature maps which analyze the visual inputs along visual feature channels known to be represented in the primate brain [31] and thought to guide visual attention and search [36] (luminance contrast, color-opponent contrast, oriented edges, etc.). Center-surround mechanisms and long-range competition for salience operate separately within each feature channel, coarsely reproducing neuronal interactions within and beyond the classical receptive field of early sensory neurons [37], [38]. These interactions are critical in transforming raw feature responses (e.g., an edge map computed over the input scene) into salient feature responses, as they emphasize locations which are locally outliers to the global statistics of the scene. As a result, local feature responses (e.g., a color contrast response to a small red object in an image) are modulated globally depending on the entire scene's content (e.g., the response to the small red object might be inhibited if many other red objects are present in the scene, or might be amplified if all other objects in the scene are blue). After these interactions, the feature maps from all feature channels are combined into a single scalar topographic saliency map. Locations of high activity in the saliency map are more likely to attract attention and gaze [28], [29].

Thus far, saliency-based analysis of scenes has been predominantly applied to relatively small images, typically on the order of 1 megapixel (MP), with at least one study pushing to 24 MP [40]. Such smaller images are coarsely matching the amount of information which might arise from a primate retina (about 1 million distinct nerve fibers in each of the human optic nerves). With larger broad-area-search images, for example 400 MP–1000 MP satellite images, it becomes an interesting research question whether the mechanisms developed by the primate brain might scale up. Here, we address this question by developing a new algorithm, which analyzes large images in small chips, thus, mimicking the processing which human image analysts might operate when they deploy multiple eye fixations on an image, analyzing each fixated location in turn. A second important research question is whether saliency maps might be useful at all for object classification, as opposed to being limited to just attention guidance as described previously. Here we hypothesize that, within each chip, the chip's saliency map may provide a coarse indication of the structure of the visual contents of the chip. Hence, rather than attempting to shift an attention spotlight to different salient locations within the chip, the hypothesis underlying the proposed algorithm is that a coarse analysis of the statistics of a chip's saliency map may provide sufficient clues for classifying the chip as containing or not a target. For example, target chips might have more numerous and sharper saliency peaks than nontarget chips. Our experiments and results test whether this approach is viable for complex target classification tasks where the intraclass heterogeneity is significant (e.g., find "boats," ranging from small pleasure craft to larger commercial or military ships). For each saliency map, mean, variance, number of local maxima, and average distance between the locations of local maxima are adopted to summarize saliency maps. These values to some extent represent the saliency intensity and the salient objects' spatial distribution. In the full algorithm described in the following, all of these values from different feature channels' saliency maps are combined together to form the "saliency features" part of the proposed algorithm.

Parallel with attention guidance and mechanisms for saliency computation, studies of scene perception have shown that observers can recognize the "gist" of a real-world scene from a single, possibly very brief glance. For example, following presentation of a photograph for just a fraction of a second, a human observer may report that it is an indoor meeting room or an outdoor scene of a beach [41]–[45]. Such a report from the first glance onto an image is remarkable considering that it summarizes the quintessential characteristics of an image, a process previously thought to require deep visual and cognitive analysis. With very brief exposures (100 ms or below), reports are typically limited to a few general semantic attributes (e.g., indoors, outdoors, playground, mountain) and a coarse evaluation of the distributions of visual features (e.g., grayscale, colorful, large masses, many small objects) [46]–[48]. Gist may be computed in brain areas which have been shown to preferentially respond to "places," that is, visual scene types with a restricted spatial layout [49]. Like Siagian-Itti's gist formulation in computer vision [50], here we use the term "gist" to represent a low-dimensional (compared with the raw image pixel array) scene representation feature vector which is acquired over very short time. In our target detection scenario, this feature vector is computed for every image chip, and we explore how well it may represent the overall information of the chip so as to support classification (e.g., chips containing boats might have significantly different gist signatures than chips which do not). Saliency and gist features appear to be complementary opposites [50]: saliency fea-

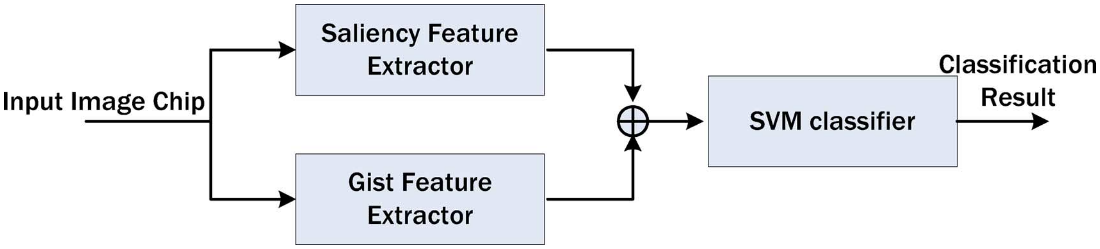


Fig. 1. Diagram of the image classification system applied to every image chip.

tures tend to capture and summarize the intensity and spatial distribution of those objects within a chip which stand out by being significantly different from their neighbors, while gist features capture and summarize the overall statistics and contextual information over the entire chip.

Given the proposed chip-based analysis approach, the task of answering "where is the target?" is equivalent to answering "does this image chip include the target?" for every chip in a large image. To achieve this decision making task, a Support Vector Machine (SVM) [51], [52] is adopted as the classifier, while the biologically inspired saliency-gist features are explored to form the feature vector in the feature space. The system overview diagram can be seen in Fig. 1.

II. Design and Implementation

Here we first describe the two computational models proposed to compute the saliency features and gist features separately.

A. Saliency Feature Computation

We compute saliency maps using several variants of the general Itti-Koch [31] architecture, and we then compute basic saliency map statistics for each variant. While in the original model only simple biological features (color, intensity, orientation) were employed, we here develop several new features which might be more effective in supporting the target/non-target classification task. The block diagram of the proposed model is shown in Fig. 2. In this model, an image is analyzed along multiple low-level feature channels to give rise to multiscale feature maps, which, as in the original Itti-Koch model, detect potentially interesting local spatial outliers. Ten feature channels are adopted in this paper: intensity, orientation (0°, 45°, 90° and 135°, combined into one "orientation" channel), local variance, entropy, spatial correlation, T-junctions, L-junctions, X-junctions, endpoints and surprise. Note that color information is not used since the images often are greyscale. Some of these feature channels (variance, entropy, spatial correlation) are computed by analyzing 16×16 image patches, giving rise to a map that is 16 times smaller than the original image horizontally and vertically (one map pixel per 16×16 image patch). The remaining feature channels are computed using image pyramids and center-surround differences, as in the original Itti-Koch algorithm: for each of these feature channels, center-surround scales are obtained from dyadic pyramids with nine scales, from scale 0 (the original image) to scale 8 (the image reduced by factor to $2^8 = 256$ in both the horizontal and vertical dimensions). Six center surround difference maps are then computed as point-to-point difference across pyramid scales, for combination of three center scales $(c = \{2,3,4\})$ and two center-surround scale differences $(\delta = \{3,4\})$. Each feature map is additionally endowed with internal dynamics that provide a strong spatial within-scale competition for activity, followed by within-feature, across-scale competition. In this way, initially possibly very noisy feature maps are reduced to sparse representations of only those locations which strongly stand out from their surroundings. Feature maps then contribute additively to the corresponding saliency maps (SMs) that represent the conspicuity of each location in their channel. Finally, a saliency map feature extractor is applied to summarize each saliency map into a 4D vector with mean, variance, number of local maxima and average distance between locations of local maxima. All those feature vectors from the ten model variants are combined into a 40D vector referred to as the "saliency features." More information about the model is described in details in the following.

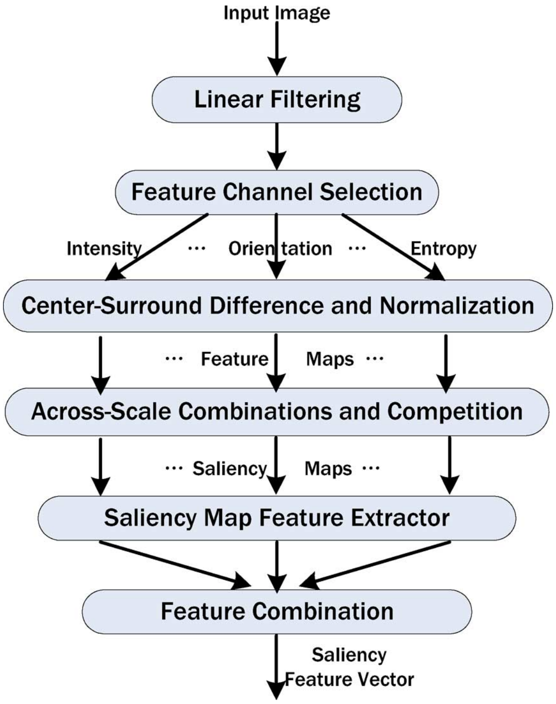


Fig. 2. Block diagram of the saliency features computation model applied to every image chip.

Intensity Channel: With the image chip as input, nine spatial scales are created using a dyadic Gaussian pyramid [25], which progressively low-pass filters and subsamples the input image, yielding horizontal and vertical image-resolution factors ranging from 1:1 (scale zeros) to 1:256 (scale nine).

Intensity represents the amount of light reflected by the corresponding point on the object in the direction of the camera view and multiplied by some constant factor that depends on the parameters of the imaging system. In our experiments, the range of the intensity value is from 0 to 65 535 (16-bit image) or from 0 to 255 (8-bit image) for all images I_s $(s = 0, 1, \ldots, 8)$ at every spatial scale. This channel is essentially as previously described [25].

Orientation Channel: Orientation features are generally very effective feature in identifying objects, as demonstrated for example by humans' ability to understand line drawings. Here we adopt Gabor filters $(\theta_k = 0^\circ, 45^\circ, 90^\circ, 135^\circ)$ to extract the orientation feature. For each image I in the image pyramid, the orientation feature maps can be obtained as follows [25]:

$$M_{O,k} = \text{Gabor}(I, \theta_k). \tag{1}$$

Local Variance Channel: Local variance channel is used to capture local pixel intensity variance over 16×16 image patches of the image chip of interest. This feature is of interest here as it has previously been shown to attract human attention [53], [54]. For each 16×16 image patch, the local variance feature map can be computed as follows:

$$\begin{aligned} M_V(i,j) \\ = \text{sqrt}\left(\frac{\sum_{sz} I^2(i,j) - S_{sz} * \text{ Mean}(I_{sz}(i,j))}{S_{sz} - 1}\right) \end{aligned} \tag{2}$$

here S_{sz} is the total pixel number of pixel (i,j)'s neighborhood with size of sz ($\text{sz} = 16 \times 16$ in our implementation).

Entropy Channel: Entropy as implemented here also provides a simple measure of information content in small 16×16 image patches. We follow the definition proposed by Privitera and Stark [54] who showed that such measure of entropy also correlates with human eye fixations. Note that many more sophisticated measures of entropy could be computed at the chip, image, or image sequence level, but this one has the advantage of being simple and motivated by previous human gaze tracking experiments. In image processing, entropy always indicates the probability distribution of the image intensity. The entropy value can be computed with the formula described in the following:

$$M_E(i,j) = -\sum_{I \in I_{sz}} p(I) * \log(p(I)) \tag{3}$$

where I_{sz} means the neighborhood of the pixel at (i,j) location, $p(I)$ stands for the probability of possible intensity I in its neighborhood.

Spatial Correlation Channel: For two random variables X and Y, their correlation can be formulated as

$$\begin{aligned} \rho_{X,Y} &= \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \\ &= \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}. \end{aligned} \tag{4}$$

Here, spatial correlation is computed at every location between a local 16×16 patch and other patches at a given radius from the local patch. It represents the similarity between the local patch and its neighbors. In the spatial correlation saliency map, low spatial correlation is a simple measure of high salience, i.e., low similarity.

Junction Channels: In addition to the Orientation channel described previously, several "junction" channels are created to further characterize the edge contents of image chips. Taking the local edge responses in different directions over small image patches into consideration, four different kinds of junction channels are created, all included in the junction saliency map: L-junction, T-junction, X-junction and endpoint. The L-junction channel is sensitive to "corner" features: it responds at locations where two edges meet perpendicularly and end at the intersection point. The T-junction channel responds when two edges are perpendicular and only one of them ends at the intersection point. Likewise, the difference of X-junction from T-junction is that in X-junction both edges do not end at the intersection point. Finally, the Endpoint channel responds when an extended edge ends. All junction channels are computed using a common framework which considers the collection of edge responses from the four maps in the Orientation channel, at points neighboring the point of interest.

We consider the 8-neighborhood of a given point of interest (at a given scale between 0 and 8 in our pyramid framework), and the one of the four orientation responses at each of the eight neighbors which is along the line segment from the central point to the neighbor (e.g., at the neighbor above the central point, the vertical orientation response is considered; at the neighbor to the left of the central point, the horizontal orientation response is considered). The response characteristics of a given junction channel is then given by a disjunction (sum) of binary response patterns (binary filter masks) applied to the neighbors' responses. For example, the T-junction detector will respond to 1) for an upright T, responses to the left (and from the orientation channel for horizontal orientation), right (horizontal orientation), and below (vertical orientation) the point of interest, plus 2) for a T rotated 90° clockwise, responses above, below, and to the left, plus 3) for an upside-down T, responses above, left and right, plus 4) for a T rotated 90° counter-clockwise, responses above, below and to the right. The L-junction and X-junction channels are defined likewise, and the mask pattern for the endpoint channel is simpler, as it will simply require that an orientation response exists on one side of the point of interest but not on the other (for example, some vertical response above but none below).

Surprise Channel: We recently proposed an enhanced saliency model, which exploits a new Bayesian definition of surprise to predict human perceptual salience in space and time [55]–[57]. Very briefly, surprise quantifies the difference between prior and posterior beliefs of an observer as new data is observed. If observing new data causes the observer to significantly reevaluate his/her/its beliefs about the world, that observation will cause high surprise. Surprise complements Shannon's definition of information by emphasizing the effect of data observations onto the internal subjective beliefs of an observer, while Shannon information objectively characterizes the data itself (in terms of, e.g., how costly it would be to transmit from one point to another). Here, we use this new model as well, though we only consider the spatial domain since all images are static. Surprise is then computed for each 16×16 image patch by establishing prior beliefs from a large neighborhood of image patches, and computing the extent to which such beliefs are adjusted into posterior beliefs after information about the central patch of interest is observed. The surprise map computed under these conditions is similar to a regular saliency map, except that the Bayesian surprise computations are used for competition across space instead of the mechanism described in the following. The surprise map is, thus, an optimized weighted combination of intensity, orientation and junction features, to which a spatial surprise detector is applied.

Feature Maps Competition: In all maps except surprise (which has its own internal competition dynamics), a feature map competition mechanism tends to globally promote maps in which a small number of strong peaks of conspicuous locations is present, while globally suppressing maps which contain numerous comparable peak responses. To implement this, first normalize the feature map to a fixed range $[0 \ldots M]$, and then find the global maximum value M and the average value $\bar{m}$ of other local maximums, finally globally multiplying the map by $(M - \bar{m})^2$, as was previously described in detail [25].

Saliency Map Feature Extractor: For each of the ten variants of the model, the obtained saliency map is relatively high-dimensional data (for example, a 512×512 image chip's saliency map size is $32 \times 32 = 1024$D), and this becomes especially true when all ten channels' saliency maps are combined. To reduce the data dimensionality while keeping the most important information, we compute four summary statistic values to represent each saliency map: mean value m_k, standard deviation v_k over the saliency map's pixels, number n_k of local maxima (peaks) in the map, and the average Euclidean distance between the local maximum points d_k. The computation formulas are described as

$$m_k = \frac{1}{W \times H} \sum_{i,j} SM_k(i,j) \tag{5}$$

$$v_k = \sqrt{\frac{1}{Sz - 1} \sum_{i,j} (SM_k(i,j) - v_k)^2} \tag{6}$$

$$d_k = \mathrm{mean}(\sqrt{(i_p - i_q)^2 + (j_p - j_q)^2)}) \qquad p, q < n_k, p \neq q \tag{7}$$

where W and H are the saliency map size, and Sz is the saliency map's area, (i_p, j_p) and (i_q, j_q) are local maximum points in saliency map, subscript k indicates the saliency map type (intensity, orientation, ...). A rational explanation of this is that the saliency map describes the conspicuity of the image and only the most salient points or regions will show on the saliency map, therefore, we can use these four values to represent the most important information of the saliency map. We may lose the salient objects' position information, however, we hypothesize that it might not affect the performance of the detection task greatly: the four statistics should capture some information about the distribution of salient objects in the image chip, no matter where they are, and may serve as a useful position-invariant (and somewhat rotation- and scale-invariant) descriptor of the image chip. Our experiments shown in the following will directly test this hypothesis. According to the previously shown analysis, the dimension of the combined saliency feature vector is: $\mathrm{Dim}_{\mathrm{sal}} = N_{\mathrm{feature\ Channels}} \times 4 = 10 \times 4 = 40$.

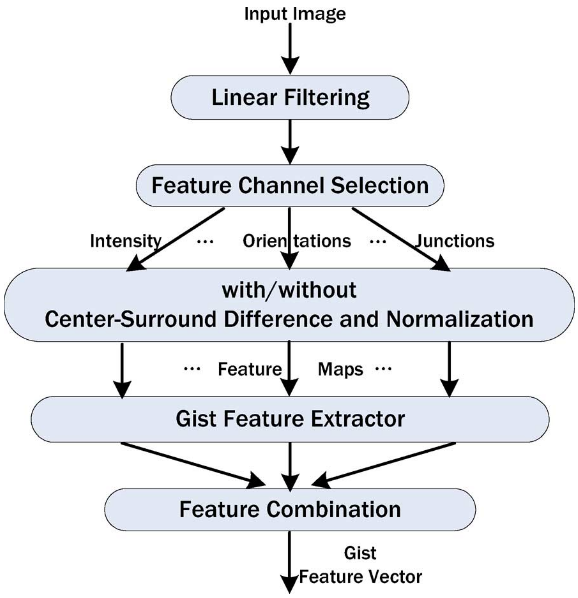


Fig. 3. Block diagram of gist features computation model applied to every image chip.

B. Gist Feature Computation

The gist feature computation model [50] is related to the saliency computation model, except that it embodies concepts of feature cooperation across space rather than competition. The gist computation model architecture used in the present paper is shown in Fig. 3 and the low-level features channels include intensity, four orientations (0°, 45°, 90°, and 135°), and four L-junctions (0°, 45°, 90°, and 135°), four T-junctions (0°, 45°, 90°, and 135°), four endpoints (0°, 45°, 90°, and 135°) and X-junction, therefore, 18 different feature channels are adopted.

Unlike the saliency feature extraction model, both center-surround and raw (before center-surround) pyramid levels are exploited. For the center-surround operation, six center surround difference maps are then computed within each pyramid as point-to-point difference across pyramid scales, for combination of three center scales ($c = \{2, 3, 4\}$) and two center-surround scale differences ($\delta = \{3, 4\}$). For the raw operation, the adopted raw pyramid scales range from 0 to 4.

Since gist features describe an image chip's overall information, we only use mean value to represent each of the gist feature maps

$$G_{k,s,c} = \frac{1}{W \times H} \sum_{i,j} GF_{k,s,c}(i,j) \tag{8}$$

where W and H are the gist feature map size, indices k, s, c denote feature map type, scale, center-surround type,

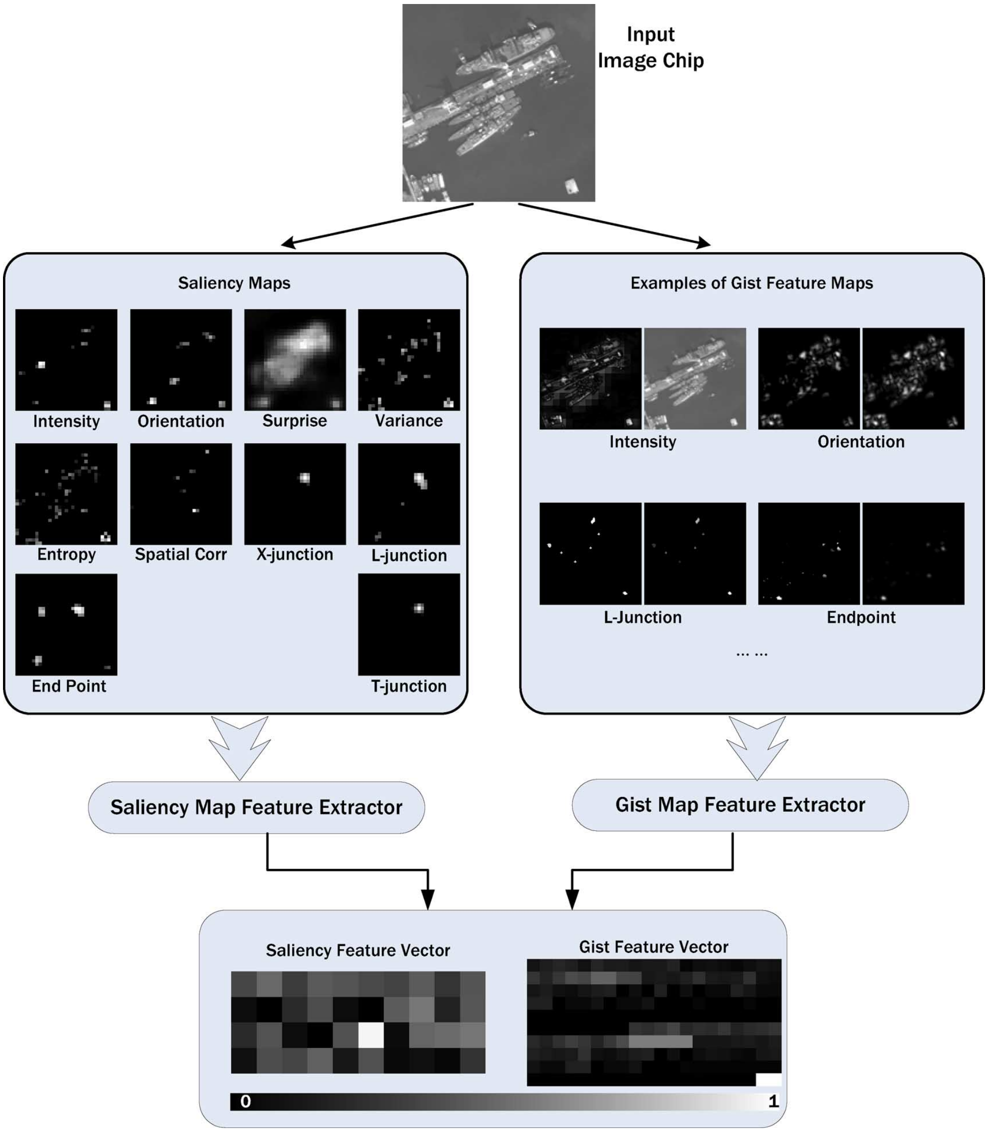


Fig. 4. Example of complete saliency-gist feature extraction for an image chip. Note that the saliency maps shown already have been subjected to spatial competition; hence, for example, out of the initially many responses in the T-junction channel at various locations and for various spatial scales, one ends up winning the competition strongly and dominating the other ones in the particular example image chip shown. The gist feature map examples shown in this figure are presented in pairs, for the center-surround and raw no-center-surround computations. In each pair, the left map is the center-surround result while the right map is the no-center-surround result. There are four pairs shown in this figure: intensity, 45° orientation, 135° L-junction and endpoint. The scales of center surround are 2 (center) and 5 (surround) while the scale of no center-surround is 2.

respectively. Therefore, the gist feature vector dimensions are $\mathrm{Dim}_{\mathrm{gist}} = N_{\text{feature Channels}} \times (N_{\text{Center Scales}} \times N_{\text{Surround Scales}} + N_{No\ CS\ \text{Scales}}) = 18 \times (3 \times 2 + 5) = 198$.

We simply combine the saliency features and gist features together to form the final saliency-gist feature vector, which is a $40 + 198 = 238$ dimensions vector. One example of the complete process for one input image is illustrated in Fig. 4. Before using these feature vectors to detect targets, it is necessary to normalize the feature values alone feature types. The normalized feature then can be sent to the classifier to implement detection task. Considering the high nonlinearity of the feature vectors' distribution, RBF (radial basic function) based SVM were adopted to complete the classification task. In this paper, SVM provided by [52] were adopted for its easy to use. Furthermore, for the normalized input, the parameters of SVM can be optimized automatically and no tuning is needed.

III. Experiments and Results

We test the proposed model with four experiments of challenging broad area search in satellite images. Mainly, the search tasks are challenging because of high intraclass variability in the target category: boats in experiments 1 and 2 (from small vessels to large ships), buildings in experiment 3, and airplanes in experiment 4. To compare our algorithm to the state of the art, we decided to employ the HMAX [14], [18], SIFT [7], and the hidden scale salient structure object detection algorithm [16] as references. We opted for HMAX and SIFT because of their popularity in target detection and in generalization over object cate-

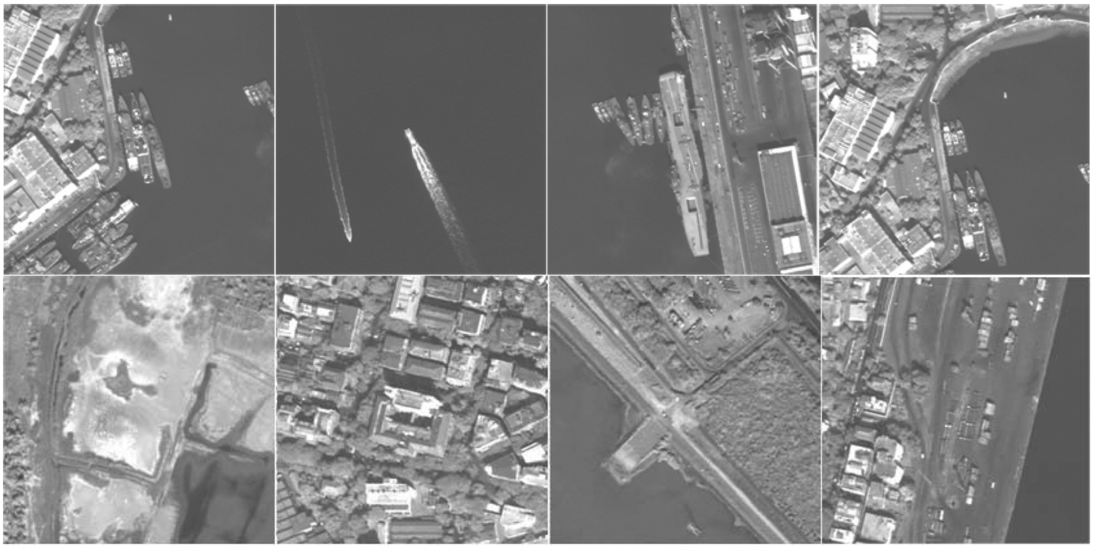


Fig. 5. Examples of target image chip and no target image chip for experiment 1, detecting image chips which contain one or more boat(s) of any size and type. The top-row image chips include one or more target boat, while the bottom-row images do not include any target.

gories from limited training data. The hidden scale salient structure method is similar to our research and performs very well in target detection for satellite images. All these references' source code is available and, thus, easy to implement for our experiments. To complement our analysis, we also compare our algorithm to Siagian-Itti's gist features proposed in [50] to show how much is gained from our very simple 4D summaries of saliency maps and from the new gist features used here.

A. *Experiment 1*

The first dataset (dataset 1) used to test the proposed model includes 14 416 image chips (500 × 500) which were cut out of one large broad-area satellite image (size 21 500 × 27 500) with a slide window step size of 200 pixels (hence, two successive chips overlap by 300 pixels). All target centers in the broad-area image were manually labeled as ground truth (if several boats were connected together, then we treated them as one target); the boats' sizes ranged from tens of pixels to hundreds of pixels. Among these image chips, 705 included targets (various boats). Examples of target image chips and nontarget image chips can be seen in Fig. 5. To compare the effectiveness of the proposed saliency-gist approach to the state of the art, we compare it with the gist feature proposed in [50] (here we call it standard gist feature), the HMAX feature [14], [18], the SIFT feature [7] and the hidden scale salient structure feature [16].

In the classification step, N positive image chips (which include one or more targets) and N negative image chips (which do not include any target) are randomly selected from the dataset and used as the training samples, while all remaining image chips are treated as test data. The commonly used measurement to evaluate the precision of classification are percentage of true positive (TP) and true negative (TN) which are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\% \tag{9}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \times 100\% \tag{10}$$

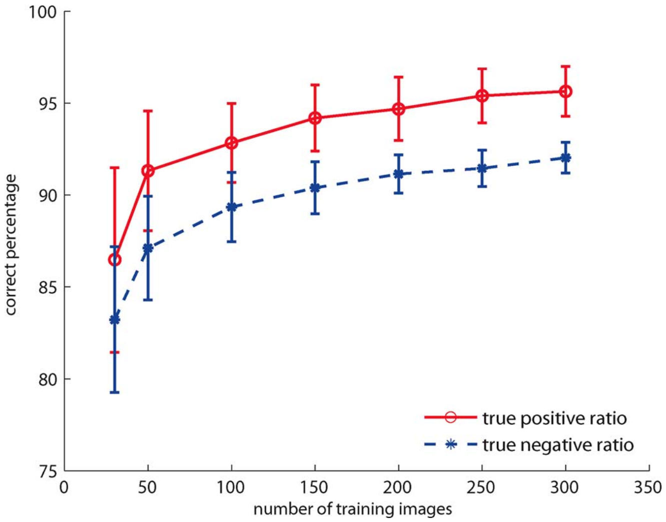


Fig. 6. Classification results for experiment 1 (detecting boats), for different numbers of training images from the pool of 705 total available chips containing one or more targets (error bars are computed from 100 runs for each number of training examples, selecting the examples randomly for each run).

where TPR and TNR stands for true positive ratio and true negative ratio. The classification results with different numbers of training samples are shown in Fig. 6. It is easy to see that when we increase the number of training samples, the classification rate improves. It is worth noting how, even with a small number of training samples, the results do not catastrophically degrade but rather remain quite high (above 80% hits and correct rejections).

For a classification system, pursuing higher TPR and lower false positive ratio (FPR) usually contradict each other: a higher TPR often causes higher FPR. With different decision criteria, the classification results may vary. For example, in a warning system, pursuing higher TPR is preferred to pursuing lower FPR. Since the receiver operating characteristic (ROC) curve has the ability to show the comparison of TPR and FPR as the

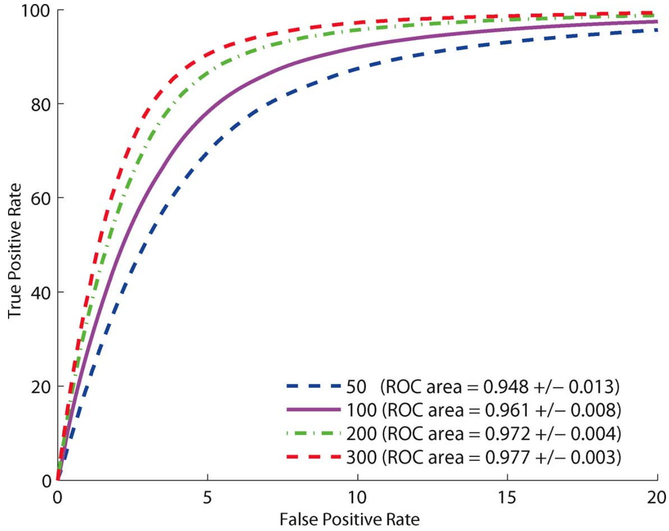


Fig. 7. ROC curve for the proposed system (zoomed-in on the horizontal axis) for different numbers of training samples, for experiment 1 (detecting boats). The corresponding ROC area values and standard deviations are labeled in the legend. (Standard deviations are computed from 100 runs for each number of training examples, selecting the examples randomly for each run).

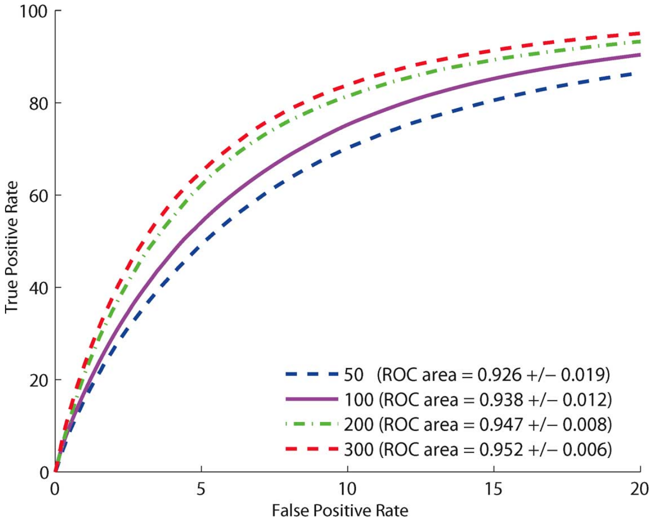


Fig. 9. ROC curve (zoomed-in on the horizontal axis) for different numbers of training samples, for experiment 2 (detecting boats, with training set from experiment 1). The corresponding ROC area values and standard deviations are labeled in the legend. (Standard deviations are computed from 100 runs for each number of training examples, selecting the examples randomly for each run).

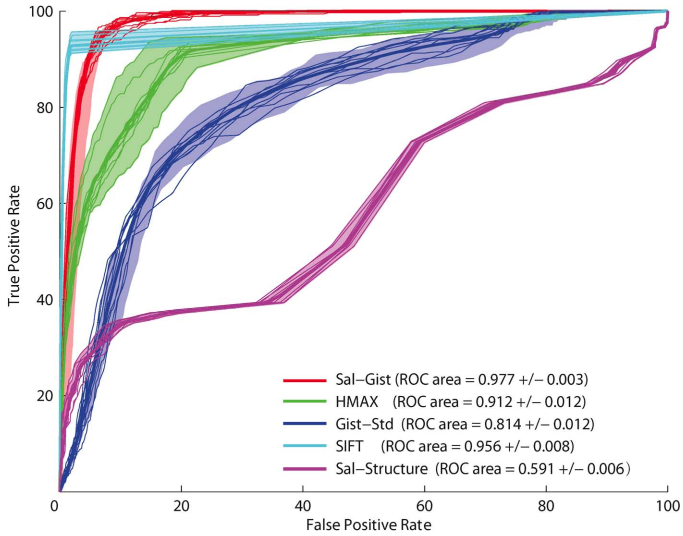


Fig. 8. ROC curve comparison among different feature types in experiment 1, detecting boats. 300 training samples were used for both the positive and negative target categories. The mean ROC area values (corresponding to the thick curves) and standard deviations are labeled in the legend. (Standard deviations are computed from 100 runs for saliency-gist feature, standard gist feature, SIFT feature and hidden scale salient structure feature, and from a smaller number of ten runs for HMAX feature because of the high run-time of HMAX). The shadow envelopes and ten thin curves for each model show the ROC curves which reach the maximum and minimum ROC area in the multiple runs of the experiment (using different randomly-chosen training samples from the training set). ROC performance for the proposed Sal-Gist algorithm is significantly better than for all other methods.

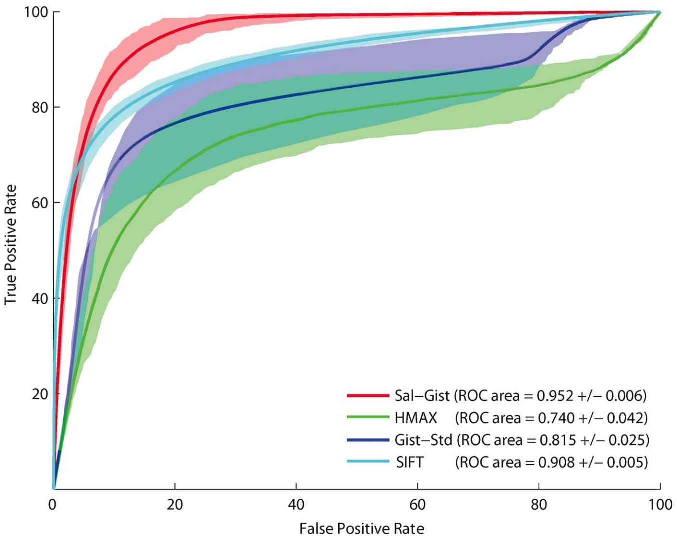


Fig. 10. ROC curve comparison among different feature types in experiment 2, detecting boats. 300 training samples were used for both the positive and negative target categories (from experiment 1's dataset). The corresponding mean ROC area values and standard deviations are labeled in the legend (100 runs for saliency-gist feature, standard gist feature and SIFT feature, ten runs for HMAX feature because of the high complexity). The shadow contours stand for the ROC curves which reach the maximum and minimum ROC area in multiple experiment runs.

classification decision criterion changes, it is widely adopted to compare performance of two different classification systems. A higher TPR while low FPR stands for a better classification system, and usually this can be described by the area under the ROC curve. An ROC area equals to 1 means a system that can perfectly classify the categories without any error, an ROC area equals to 0.5 stands for a random classification system, and, the bigger ROC area, the better classification performance. To compute ROC curves with our algorithm, we systematically vary distance to the decision boundary as the criterion parameter. Fig. 7 shows ROC curves for the proposed saliency-gist algorithm, as a function of the number of training examples. We can see that performance degrades gracefully as the number of training examples is decreased. The corresponding ROC curves and the ROC areas of classification with saliency-gist feature,

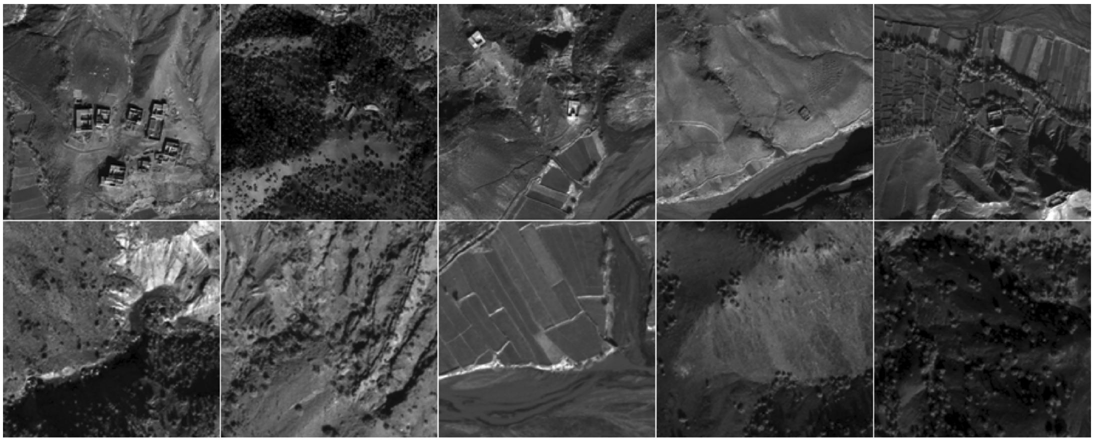


Fig. 11. Examples of target image and no target image for experiment 3, finding buildings of any type, size, and style. Top-row images include one or more target(s) while the bottom row images do not include any target.

HMAX feature, SIFT feature, hidden scale salient structure feature and standard gist features are shown in Fig. 8. (marked as sal-gist, HMAX, SIFT, sal-structure, and gist-std in the figure, respectively). It is clear from the figure that the saliency-gist feature outperforms the other features greatly (t-tests on the 100 ROC values obtained with each of the 100 randomly selected training sets, $p < 10^{-10}$ or better), hence, demonstrating appeal of the proposed approach. Also, from the figure we can see that the hidden scale salient structure method almost failed in this experiment. This is mainly because the targets (boats) are not salient compared with many inland buildings when using the salient structure algorithm in [16] and, thus, the algorithm misclassified many buildings as boat targets.

B. Experiment 2

This experiment tests how training on one broad-area image taken at one given time and location may generalize to testing on another broad-area image taken at another time and location. The second dataset (dataset 2) includes 11 385 image chips (500×500) which were cut out of another large broad-area satellite image (size 23 300×20 100, taken from the same country but on a different date and at a different place than the broad-area image of experiment 1), with the same slide window size as in dataset 1. We labeled the targets manually as ground truth like in experiment 1 and there are 1 049 image chips which include one or more target(s). In this experiment, training samples for the classifier are randomly selected from dataset 1, while all the image chips in dataset 2 are used as test set.

Fig. 9 shows that ROC performance improves with the number of training samples, as in experiment 1. With 300 training samples, ROC area was 0.952 here, as compared to 0.977 in experiment 1 (Fig. 8), suggesting good generalization capability to new, never seen images. Like in experiment 1, the comparison of detection results with the saliency-gist feature, standard gist features, HMAX features, and SIFT features (the hidden scale salient structure feature is not adopted to do the comparison in this experiment due to its poor performance in experiment 1) shows that the saliency-gist feature performs much better than other three features (Fig. 10, t-tests, $p < 10^{-12}$ or better).

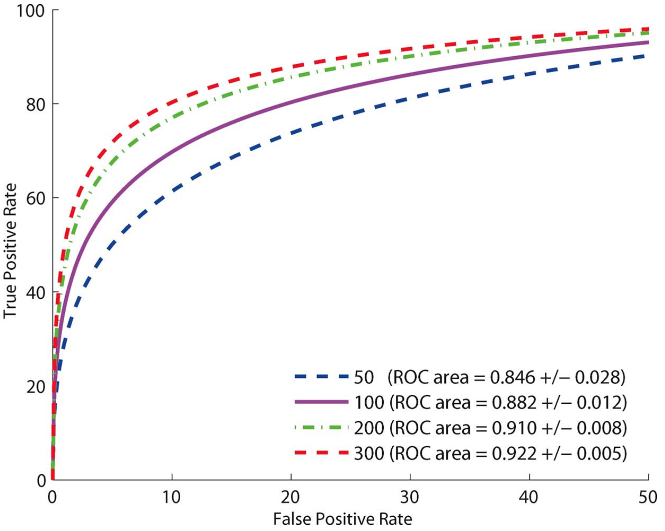


Fig. 12. ROC cures (zoomed in on the horizontal axis) for different numbers of training samples in experiment 3 (detecting buildings). The corresponding ROC area values and standard deviations are labeled in the legend. (Standard deviations are computed from 100 runs for each number of training examples, selecting the examples randomly for each run).

C. Experiment 3

In this experiment, targets are simply defined as "buildings" in satellite images. This experiment, thus, tests the ability of our same algorithm to classify very different types of targets; the intraclass variability here is also arguably even larger than in experiments 1 and 2 (see Fig. 11). The dataset (dataset 3) used here includes 108 885 image chips (this experiment used a smaller chip size of 256×256 because the targets were also smaller than in experiments 1 and 2) with 6 323 of them being positive examples. Fig. 11 shows examples of buildings and negative examples. Like in experiments 1 and 2, the image chips were cut from a broad-area satellite image (size 16 512×27 520, taken from a different country and a different year than the images of experiments 1 and 2). The slide window size here was 64 pixels. Ground-truth information for this dataset (locations of buildings) was provided to us by an outside corporation. The ROC curves for different numbers of training samples are plotted in Fig. 12. As we can see, performance again improves

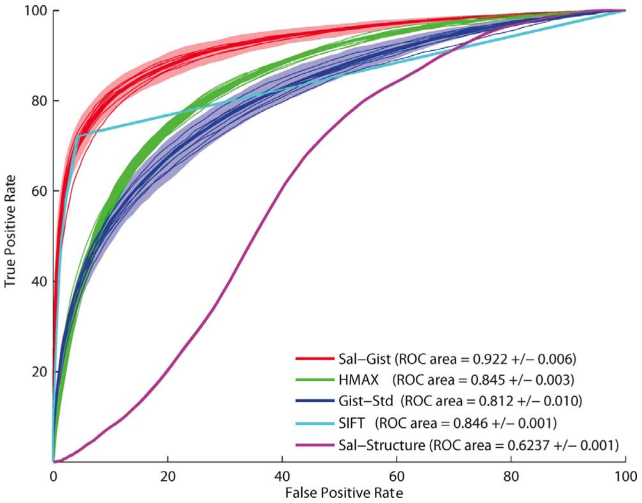


Fig. 13. ROC curve comparison among different feature types in experiment 3, detecting buildings. 300 positive and 300 negative training examples were used. The experiment parameters are the same as in experiments 1 and 2. The shadow contours of SIFT and hidden scale salient structure feature are quite small and can not seen in this figure.

with the size of the training set. Again, we compare the detection results with the standard gist features, the HMAX features, the SIFT features and the hidden scale salient structure features. The corresponding ROC curves of classification with these feature types are shown in Fig. 13. It is clear from the figure that the saliency-gist features again outperform the other two features greatly (t-tests, $p < 10^{-17}$ or better).

To illustrate the detection result in a more straightforward and global way, we adopt a probability map representation (PM) to show the results. A probability map is a matrix which depicts the probability value for each image chip to contain a target. The rescaled broad-area satellite image and some example target buildings are shown in Fig. 14(a), and the corresponding probability map is shown in Fig. 14(b), the red points in the images stand for the labeled targets' center location. This simple representation reinforces the ROC results and suggests a high performance of the algorithm, as shown by the overlap between red ground truth locations and brighter locations in the PM (higher probability of target according to our algorithm). During search for buildings, exploring the image in decreasing order of target probability per our algorithm would isolate more targets faster than a naïve scan from left to right and top to bottom.

D. Experiment 4

An aerial image of an airport is adopted in this experiment to detect the "airplanes" (see Fig. 15). The dataset (dataset 4) used here include 2 601 image chips, of which 1 382 of them include a target. For each chip, the size is 64 × 64 due to the small target size. Compared to the previous experiments, the target is relatively easier to detect because intra class variances (both in shape and area) are small. Here we compared the detection performance among saliency-gist feature, hidden scale salient structure feature and SIFT. Ten positive and ten negative examples were randomly selected as training data while the rest were taken as test data. The detection results from different methods are plotted in Fig. 16 (100 runs for each). From the figure we can see that all three methods perform very well while the proposed method performs even better than the others (no shadow contours plotted here because the difference of results from different method is small while the variance of result from SIFT is relatively big which may cause the whole figure not clear).

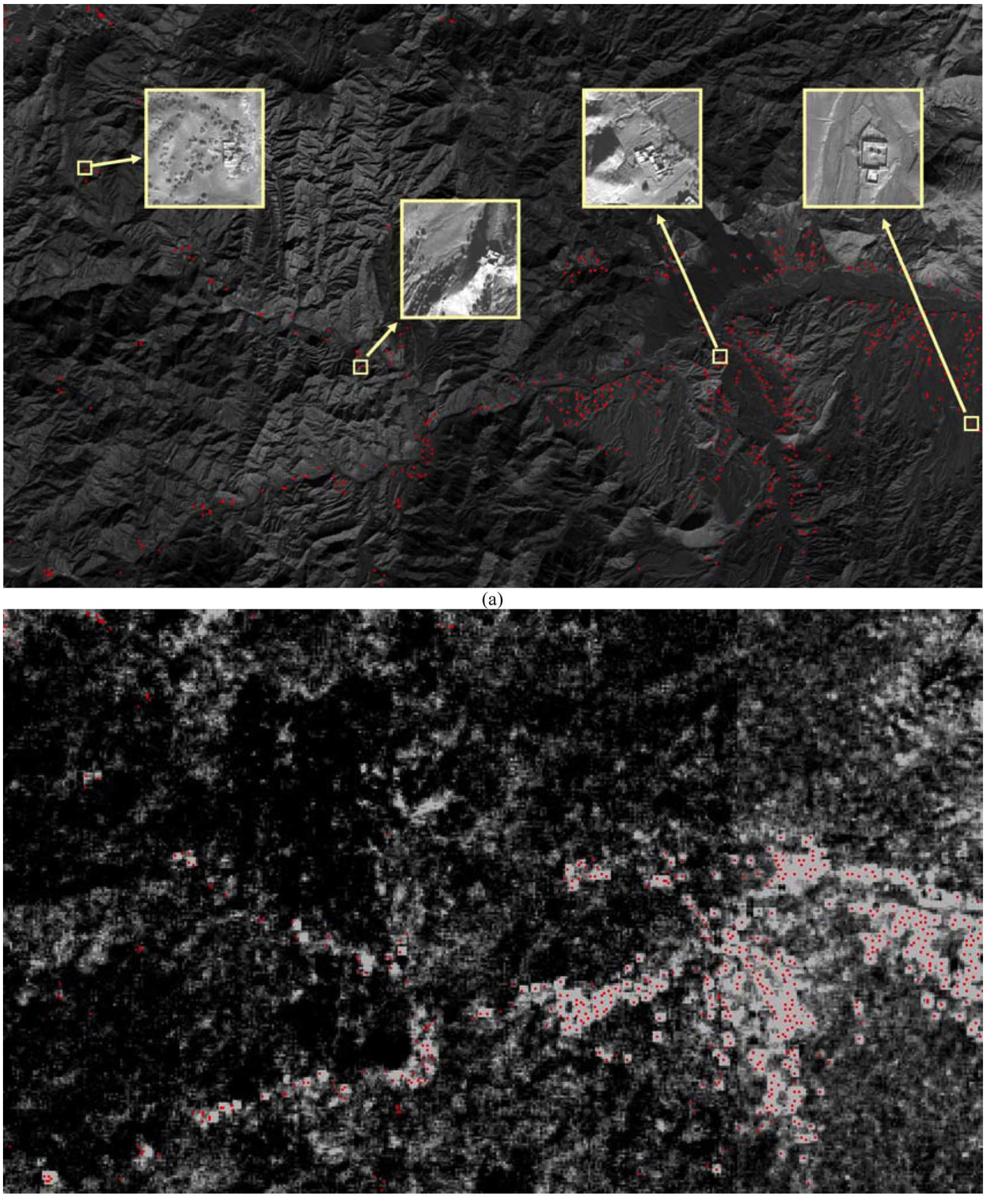
E. Saliency Versus Gist

As saliency-gist features yield great classification results, it is interesting to see the separate contributions of the saliency features and gist features. The ROC area of using saliency features only, gist features only (in our new implementation, which includes more feature channels than the older Gist-Std model), and combined saliency-gist features in all four experiments are shown in Table I. It can be seen from the table that the combined saliency-gist features outperforms both saliency features and gist features in all experiments. Hence, these results show that the saliency features and gist features are not fully redundant, even though they are computed using similar low-level feature detectors. In addition, the table shows that saliency features perform better in experiments 1, 2 and 4, while gist features perform better in experiment 3. Thus, in different cases, the classification results depends more on different types of information (saliency information and gist information), which again reinforces the benefits of using both types of features.

IV. Discussion

Our results show that the proposed algorithm performs better than the state-of-the-art (HMAX algorithm, SIFT algorithm, hidden scale salient structure algorithm and previously proposed gist algorithm alone) in difficult target search scenarios. This was achieved in situations where targets can vary greatly in their size, shape, and number of targets per chip. Overall, the proposed algorithm is conceptually very simple and at the same time very general, since the feature extraction stages were not designed or tuned for the specific types of images and targets tested here. Taking all results together suggests that the proposed system may be further applicable to a wide range of images and target types. Indeed, nothing in the proposed algorithm has been specifically developed or tuned for the boat or building or airplane targets tested here, or for the type of images processed in our experiments.

The success of the proposed approach may be due to our use of two complementary sets of biologically-inspired features: gist features largely discard spatial information, while saliency features summarize it. In the human brain, it is clear that object recognition relies on being able to compute invariants, but at the same time pose parameters are not lost: although one recognizes an upside-down face as being a face, one is also aware that it is upside-down. Our approach here seems to benefit from this dual view of the image data. Recently, some other biologically-inspired feature extraction methods [19] have started to use the "gestalt" information (continuity, symmetry, closure, repetition, etc.) to conduct object detection and have shown promising results. It is likely that combining these feature types will get even better detection performance. There are many other feature types which could be also added to our approach, including for example locally-binary pattern (LBP) features which have been particularly successful in texture segmentation [58].



(a)

(b)

Fig. 14. Illustration of "building" detection in experiment 3. (a) Rescaled broad-area satellite image (16 512 × 27 520 pixels) and some target examples. (b) Probability map of (a) computed by our algorithm. The red points are the true target center locations. In the probability map, lighter areas indicate higher probability of targets, while darker areas denote lower probability of targets according to the algorithm.

TABLE I
Comparison of ROC Areas of Different Types of Features in Four Experiments

	Saliency Feature	Gist Feature	Saliency-Gist Feature
Experiment 1	0.969 ± 0.003	0.943 ± 0.008	0.977 ± 0.003
Experiment 2	0.945 ± 0.007	0.903 ± 0.009	0.952 ± 0.005
Experiment 3	0.789 ± 0.007	0.905 ± 0.005	0.922 ± 0.005
Experiment 4	0.927 ± 0.031	0.942 ± 0.028	0.976 ± 0.003

The proposed algorithm does not take any complex procedure to combine the features extracted, although many research studies have proposed feature combination algorithms to improve classification performance [59], [60]. Here we only show that the combination of gist feature and salient feature are complementary and can achieve good performance in target detection. It is interesting that saliency and gist features both contribute significantly to performance, and are not fully redundant (Table I). This suggests a new use of saliency algorithms, for classification of images based on their saliency maps, as opposed to using the saliency maps to generate shifts of attention. It is interesting to think whether humans and other animals may use this as well. It is possible that human saliency maps in posterior parietal cortex, the pulvinar nucleus, the frontal eye fields,

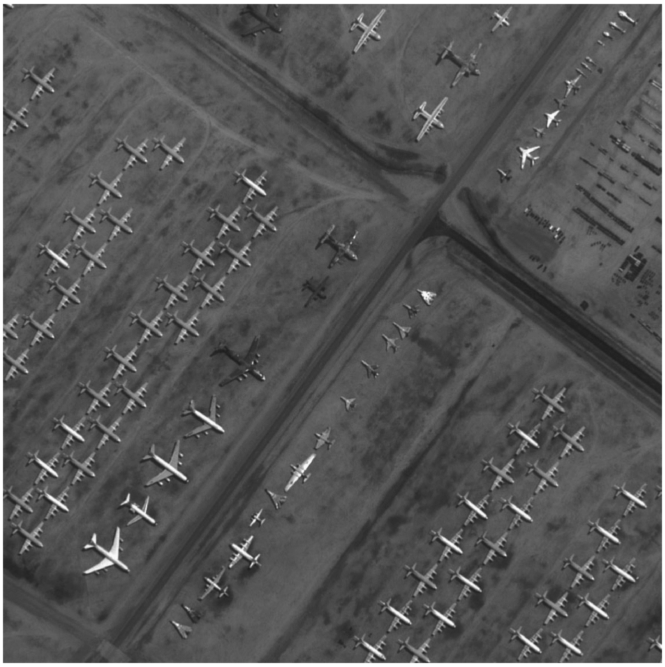


Fig. 15. Image used to detect the airplanes in experiment 4.

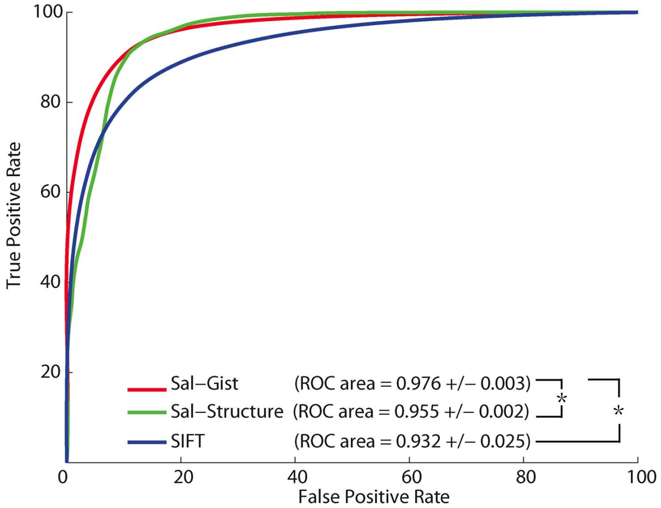


Fig. 16. ROC curve comparison among different feature types in experiment 4, detecting airplanes. Ten positive and ten negative training examples were used. * indicate statistically different ROC performance (t-test, $p < 10^{-10}$ or better).

or the superior colliculus [31] may also be analyzed in a holistic fashion and may contribute to the very rapid understanding of the rough layout of the scene. That is, the coarse structure of saliency maps may combine with the broad semantic information provided by the gist features to yield a coarse and rapid understanding of both a scene's gist and layout [61].

Our approach reinforces the idea, as shown by recent successes in the domains of statistical machine translation of text into foreign languages or of speech analysis [62], that relatively shallow statistical analysis of large datasets can yield surprisingly good classification and recognition results. Indeed, our algorithm does not try to understand the geometric structure or other specific high-level or cognitive feature of targets (e.g., buildings should have walls, tend to be rectangular, etc) and is not attempting recognition by components (breaking down target objects into elementary parts and their spatial arrangements [63].

The proposed algorithm is mostly intended as a front-end, to be used to perform coarse preliminary analysis of large complex scenes. The data returned certainly is still far from representing a complete understanding of the scene's contents. However, our algorithm's output can be used in at least two practical ways: first, to compute statistics at the region level, like, e.g., finding areas in the world with high concentrations of boats, or determining which regions in a country have more buildings and, hence, may be more densely populated. Such basic statistics may be of great use on their own, for example when planning rescue efforts following a natural disaster, or may assist a human image analyst in performing deeper and more cognitively-driven surveys of imagery. Second, our algorithm can be used to rank image chips by interest (using the probability maps of Fig. 14), so as to focus limited resources onto the most promising image locations. Resources may be limited because of limited human personnel, human viewing time (e.g., when using rapid serial visual presentation of image chips [64], or computation time (e.g., using a more sophisticated and time-consuming object recognition back-end to validate high-probability chips). It is likely that our system could perform even better if one was to apply some of the recognition-by-components principles or other recognition back-end to the high-probability target chips returned by our algorithm.

Thus far, our algorithm has only been applied to greyscale visible imagery. With the increasing popularity of color and multispectral imagery, it remains to be tested in future work whether our simple approach will scale up to a larger number of spectral bands. All C++ source code for our algorithms is available on the authors' web site (http://iLab.usc.edu).

Acknowledgment

The authors would like to thank DigiGlobe for providing the aerial images. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States Government or any agency thereof.

References

[1] Y. Amit, *2D Object Detection and Recognition, Models, Algorithms and Networks*. Cambridge, MA: MIT Press, 2002.

[2] V. Vilaplana, F. Marques, and P. Salembier, "Binary partition trees for object detection," *IEEE Trans. Image Process.*, vol. 17, no. 11, pp. 2201–2216, Nov. 2008.

[3] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1475–1490, Nov. 2004.

[4] R. Brunelli, *Template Matching Techniques in Computer Vision: Theory and Practice*. Hoboken, NJ: Wiley, 2009.

[5] I. Craw, H. Ellis, and J. Lishman, "Automatic extraction of face features," *Pattern Recognit. Lett.*, vol. 5, pp. 183–187, 1987.

[6] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognit.*, vol. 25, no. 1, pp. 65–77, 1992.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[8] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[9] K. Grauman and T. Darrell, "Pyramid match kernels: Discriminative classification with sets of image features," in *Proc. ICCV*, 2005, vol. 2, pp. 1458–1465.

[10] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 416–431, Mar. 2006.

[11] G. Bouchard and B. Triggs, "Hierarchical part-based visual object categorization," in *Proc. CVPR*, 2005, vol. 1, pp. 710–715.

[12] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale invariant learning," in *Proc. CVPR*, 2003, vol. 2, pp. 264–271.

[13] A. C. Berg, T. L. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondence," in *Proc. CVPR*, 2005, vol. 1, pp. 26–33.

[14] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, pp. 1019–1025, 1999.

[15] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Proc. CVPR*, 2005, vol. 2, pp. 994–1000.

[16] B. Chalmond, B. Francesconi, and S. Herbin, "Using hidden scale for salient object detection," *IEEE Trans. Image Process.*, vol. 15, no. 9, pp. 2644–1655, Sep. 2006.

[17] J. Mutch and D. Lowe, "Object class recognition and localization using sparse features with limited receptive fields," *Int. J. Comput. Vis.*, vol. 80, no. 1, pp. 45–57, 2008.

[18] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 411–426, Mar. 2007.

[19] S. Bileschi and L. Wolf, "Image representations beyond histograms of gradients: The role of Gestalt descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007.

[20] D. Manolakis, D. Marden, and G. A. Shaw, "Hyperspectral image processing for automatic target detection applications," *Lincoln Lab. J.*, vol. 14, no. 1, 2003.

[21] C. Chang, H. Ren, and S. Chiang, "Real-time processing algorithms for target detection and classification in hyperstpectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 4, pp. 760–768, Apr. 2001.

[22] H. Li and J. H. Michels, "Parametric adaptive signal detection for hyperspectral imaging," *IEEE Trans. Signal Process.*, vol. 54, no. 7, pp. 2704–2715, Jul. 2006.

[23] J. Lanir and M. Maltz, "Analyzing target detection performance with multispectral fused images," in *Proc. SPIE*, 2006.

[24] S. Buganim and S. R. Rotman, "Matched filters for multispectral point target detection," in *Proc. SPIE*, 2006.

[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[26] L. Elazary and L. Itti, "Interesting objects are visually salient," *J. Vis.*, vol. 8, no. 3:3, pp. 1–15, 2008.

[27] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[28] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Vis. Cogn.*, vol. 12, pp. 1093–1123, 2005.

[29] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, pp. 97–137, 1980.

[30] J. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonom. Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.

[31] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[32] N. Bruce and J. Tsotsos, "Saliency, attention and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, 2009.

[33] D. Gao and N. Vasconcelos, "Integrated learning of saliency, complex features, and object detectors from cluttered scenes," in *Proc. IEEE CVPR*, 2005.

[34] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in highly dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.

[35] [Online]. Available: http://ilab.usc.edu/toolkit

[36] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?," *Nature Rev. Neurosci.*, vol. 5, pp. 495–501, 2004.

[37] M. W. Cannon and S. C. Fullenkamp, "Spatial interactions in apparent contrast: Inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations," *Vis. Res.*, vol. 31, pp. 1985–1998, 1991.

[38] A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis, "Visual cortical mechanisms detecting focal orientation discontinuities," *Nature*, vol. 378, pp. 492–496, 1995.

[39] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," *Vis. Res.*, vol. 46, no. 26, pp. 4333–4345, 2006.

[40] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, no. 10–12, pp. 1489–1506, 2000.

[41] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[42] M. C. Potter, "Meaning in visual search," *Science*, vol. 187, no. 4180, pp. 965–966, 1975.

[43] I. Biederman, "Do background depth gradients facilitate object identification?," *Perception*, vol. 10, pp. 573–578, 1982.

[44] B. Tversky and K. Hemenway, "Categories of the environmental scenes," *Cogn. Psychol.*, vol. 15, pp. 121–149, 1983.

[45] A. Oliva and P. Schyns, "Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli," *Cogn. Psychol.*, vol. 34, pp. 72–107, 1997.

[46] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cogn.*, vol. 7, pp. 17–42, 2000.

[47] T. Sanocki and W. Epstein, "Priming spatial layout of scenes," *Psychol. Sci.*, vol. 8, pp. 374–378, 1997.

[48] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520–522, 1995.

[49] R. Epstein, D. Stanley, A. Harris, and N. Kanwisher, "The parahippocampal place area: Perception, encoding, or memory retrieval?," *Neuron*, vol. 23, pp. 115–125, 2000.

[50] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[51] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Proc. IEEE CVPR*, 1997, vol. 1, pp. 130–136.

[52] [Online]. Available: http://www.kernel-machines.org

[53] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," *Network: Comput. Neural Syst.*, vol. 10, pp. 341–350, 1999.

[54] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison witheye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, Sep. 2000.

[55] L. Itti and P. F. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE CVPR*, 2005, vol. 1, pp. 631–637.

[56] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 547–554, 2006.

[57] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, 2009.

[58] T. Ojala and M. Pietikäinen, "Unsupervised texture segmentation using feature distributions," *Pattern Recognit.*, vol. 32, pp. 477–486, 1999.

[59] L. Wolf, S. Bileschi, and E. Meyers, "Perception strategies in hierarchical vision systems," in *Proc. IEEE CVPR*, 2006.

[60] I. Oh, J. Lee, and C. Suen, "Analysis of class separation and combination of class-dependent features for handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 1089–1094, Oct. 1999.

[61] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cogn.*, vol. 7, pp. 17–42, 2000.

[62] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Ling.*, vol. 29, no. 1, pp. 19–51, 2003.

[63] I. Biederman, "Recognition-by-components: A theory of human image understanding," *Psychol. Rev.*, vol. 94, no. 2, pp. 115–147, 1987.

[64] W. Einhaeuser, T. N. Mundhenk, P. F. Baldi, C. Koch, and L. Itti, "A bottom-up model of spatial attention predicts human error patterns in rapid scene recognition," *J. Vis.*, vol. 7, no. 10, pp. 1–13, Jul. 2007.

Zhicheng Li received the B.S. degree in electronics and information from Northwestern Polytechnical University, Xi'an, China, in 2005, and is currently pursuing the Ph.D. degree in school of automation science and electrical engineering, Beihang University, Beijing, China.

He is a Visiting Scholar in the Computer Science Department, University of Southern California, Los Angeles, since 2007. His main research interests include visual attention modeling, video compression, and target detection.

Laurent Itti received the M.S. degree in image processing from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1994, and the Ph.D. degree in computation and neural systems from the California Institute of Technology (Caltech), Pasadena, in 2000.

He is currently an Associate Professor of Computer Science, Psychology, and Neuroscience at the University of Southern California, Los Angeles. His research interests are in biologically-inspired computational vision, in particular in the domains of visual attention, gist, saliency, and surprise, with applications to video compression, target detection, and robotics.