Learning to Detect A Salient Object

Tie Liu[1] Jian Sun[2] Nan-Ning Zheng[1] Xiaoou Tang[2] Heung-Yeung Shum[2]

[1]Xi'an Jiaotong University
Xi'an, P.R. China

[2]Microsoft Research Asia
Beijing, P.R. China

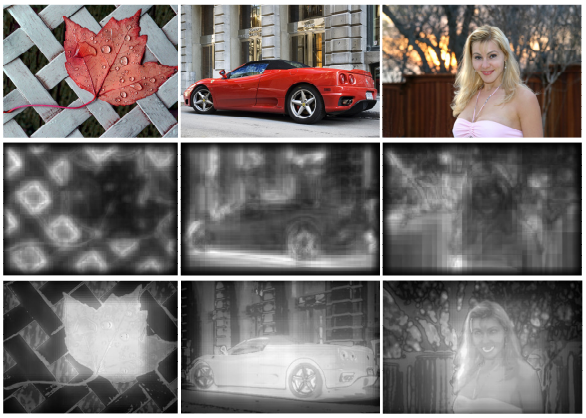
Abstract We study visual attention by detecting a salient object in an input image. We formulate salient object detection as an image segmentation problem, where we separate the salient object from the image background. We propose a set of novel features including multi-scale contrast, center-surround histogram, and color spatial distribution to describe a salient object locally, regionally, and globally. A Conditional Random Field is learned to effectively combine these features for salient object detection. We also constructed a large image database containing tens of thousands of carefully labeled images by multiple users. To our knowledge, it is the first large image database for quantitative evaluation of visual attention algorithms. We validate our approach on this image database, which is public available with this paper.

1. Introduction

"*Everyone knows what attention is...*"
—William James, 1890

The human brain and visual system pay more attention to some parts of an image. Visual attention has been studied by researchers in physiology, psychology, neural systems, and computer vision for a long time. There are many applications for visual attention, for example, automatic image cropping [23], adaptive image display on small devices [4], image/video compression, advertising design [7], and image collection browsing. Recent studies [18, 22, 26] demonstrated that visual attention helps object recognition, tracking, and detection as well.

Most existing visual attention approaches are based on the bottom-up computational framework [3, 6, 8, 9, 10, 11, 19, 25] because visual attention is in general unconsciously driven by low-level stimulus in the scene such as intensity, contrast, and motion. These approaches consist of the following three steps. The first step is *feature extraction*, in which multiple low-level visual features, such as intensity, color, orientation, texture and motion are extracted from the image at multiple scales. The second step is *saliency computation*. The saliency is computed by a center-surround operation [10], self-information [3], or graph-based random walk [6] using multiple features. After normalization and linear/non-linear combination, a master map [24] or a salient map [11] is computed to represent the saliency of each image pixel. Last, a few key locations on the saliency map are identified by winner-take-all, or inhibition-of-return, or other non-linear operations. While these approaches have worked well in finding a few fixation locations in both synthetic and natural images, they have not been able to accurately detect where visual attention should be.



(a) (b) (c)

Figure 1. Salient map. From top to bottom: input image, salient map computed by Itti's algorithm (http://www.saliencytoolbox.net), and salient map computed by our approach.

For instance, the middle row in Figure 1 shows three salient maps computed using Itti's algorithm [10]. Notice that the saliency concentrates on several small local regions with high contrast structures, e.g., the background grid in (a), the shadow in (b), and the foreground boundary in (c). Although the leaf in (a) commands much attention, the saliency for the leaf is low. Therefore, these salient maps computed from low-level features are not a good indication for where a user's attention is while perusing these images.

In this paper, we incorporate the high level concept of salient object into the process of visual attention computation. In Figure 1, the leaf, car, and woman attract the most visual attention in each respective image. We call them salient objects, or foreground objects that we are familiar with. As can

1-4244-1180-7/07/$25.00 ©2007 IEEE

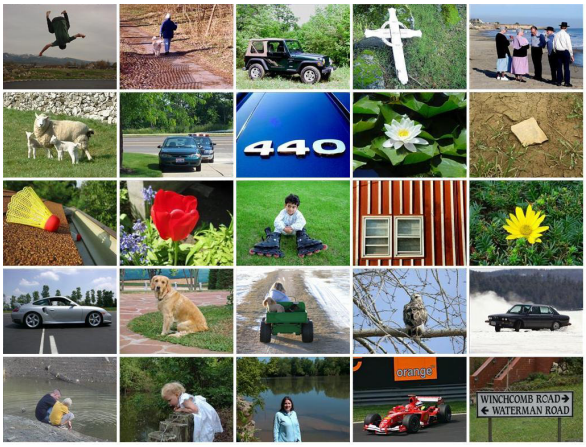


Figure 2. Sample images in our salient object image database.

be observed in Figure 2, people naturally pay more attention to salient objects in images such as a person, a face, a car, an animal, or a road sign. Therefore, we formulate *salient object detection* as a binary labeling problem that separates a salient object from the background. Like face detection, we learn to detect a familiar object; unlike face detection, we detect a familiar yet unknown object in an image.

We present a supervised approach to learn to detect a salient object in an image. First, we construct a large image database with 20,000+ well labeled images for training and evaluation. To our knowledge, it is the first time a large image database is available for quantitative evaluation. The user labeled information is used to supervise the salient object detection. It can be viewed as top-down information in the training phase. Second, to overcome the challenge that we do not know what a specific object or object category is, we propose a set of novel local, regional, and global features to define a generic salient object. These features are optimally combined through Condition Random Field (CRF) learning. Moreover, the segmentation is also incorporated into the CRF to detect a salient object with unknown size and shape. The last row in Figure 1 shows the salient maps computed by our approach.

2. Image Database

People may have different ideas about what a salient object in an image is. To address the problem of "what is the most likely salient object in a given image", we take a voting strategy by labeling a "ground truth" salient object in the image by multiple users. And in this paper, we focus on the case of a single salient object in an image.

Salient object representation. Formally, we represent the salient object as a binary mask $A = \{a_x\}$ in a given image I. For each pixel x, $a_x \in \{1, 0\}$ is a binary label to indicate whether or not the pixel x belongs to the salient object. For labeling and evaluation, we ask the user to draw a rectangle to specify a salient object. Our detection algorithm also outputs a rectangle.

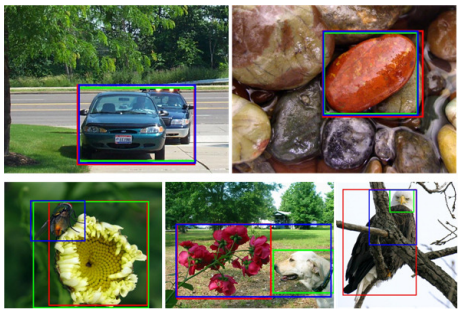


Figure 3. Labeled images from 3 users. Top: two consistent labeling examples. Bottom: three inconsistent labeling examples.

Image source. We have collected a very large image database with 130,099 high quality images from a variety of sources, mostly from image forums and image search engines. Then we manually selected 60,000+ images each of which contains a salient object or a distinctive foreground object. We further selected 20,840 images for labeling. In the selection process, we excluded any image containing a very large salient object so that the performance of detection can be more accurately evaluated.

Labeling consistency. For each image to be labeled, we ask the user to draw a rectangle which encloses the most salient object in the image according to his/her own understanding. The rectangles labeled by different users usually are not the same. To reduce the labeling inconsistency, we vote a "ground truth" labeling from the rectangles drawn by multiple users.

In the first stage, we asked three users to label all 20,840 images individually. On average, each user took 10-20 seconds to draw a rectangle on an image. The whole process took about three weeks. Then, for each labeled image, we compute a saliency probability map $G = \{g_x | g_x \in [0, 1]\}$ of the salient object using the three user labeled rectangles:

$$g_x = \frac{1}{M} \sum_{m=1}^{M} a_x^m, \tag{1}$$

where M is the number of users and $A^m = \{a_x^m\}$ is the binary mask labeled by the mth user. Figure 3 shows two highly consistent examples and three inconsistent examples. The inconsistent labeling is due to multiple disjoint foreground objects for the first two examples at the bottom row. The last example at the bottom row shows that an object has hierarchical parts that are of interest. We call this image set $\mathcal{A}$. In this paper, we focus on consistent labeling of a single salient object for each image.

To measure the labeling consistency, we compute statistics C_t for each image:

$$C_t = \frac{\sum_{x \in \{g_x > t\}} g_x}{\sum_x g_x}. \tag{2}$$

C_t is the percentage of pixels whose saliency probabilities are above a given threshold t. For example, $C_{0.5}$ is the per-

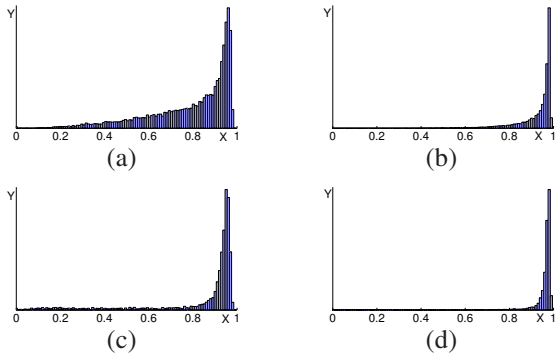


Figure 4. Labeling consistency. (a) (b) $C_{0.9}$ (agreed by all 3 users) and $C_{0.5}$ on image set $\mathcal{A}$. (c) (d) $C_{0.9}$ (agreed by at least 8 of 9 users) and $C_{0.5}$ on image set $\mathcal{B}$.

centage of the pixels agreed on by at least half of the users. $C_{0.9} \approx 1$ means that the image is consistently labeled by all the users. Figures 4(a) and 4(b) show the histograms of $C_{0.9}$ and $C_{0.5}$ on the image set $\mathcal{A}$. As we can see, the labeled results are quite consistent, e.g., 92% of the labeling results are consistent between at least two users (Figure 4 (b)) and 63% of the labeling results are highly consistent among all three users (Figure 4 (a)).

In the second stage, we randomly selected 5000 highly consistent images (i.e., $C_{0.9} > 0.8$) from the image set $\mathcal{A}$. Then, we asked nine different users to label the salient object rectangle. Figures 4(c) and 4(d) show the histograms of $C_{0.9}$ and $C_{0.5}$ on these images. Compared with the image set $\mathcal{A}$, this set of images has less ambiguity of what the salient object is. We call these images as image set $\mathcal{B}$.

After the above two-stage labeling process, the salient object in our image database is defined based on the "majority agreement" of multiple users and represented as a saliency probability map.

Evaluation. With the saliency probability map G, for any detected salient object mask A, we define region-based and boundary-based measurements.

We use the precision, recall, and F-measure for region-based measurement. Precision/Recall is the ratio of correctly detected salient region to the detected/"ground truth" salient region: $\textit{Precision} = \sum_x g_x a_x / \sum_x a_x, \textit{Recall} = \sum_x g_x a_x / \sum_x g_x$. F-measure is the weighted harmonic mean of precision and recall, with a non-negative α: $F_\alpha = \frac{(1+\alpha)\times \textit{Precision}\times \textit{Recall}}{\alpha \times \textit{Precision} + \textit{Recall}}$. We set $\alpha = 0.5$ following [17]. The F-measure is an overall performance measurement.

For the boundary-based measurement, we use boundary displacement error (BDE) [5], which measures the average displacement error of corresponding boundaries of two rectangles. The displacement is averaged over the different users.

3. CRF for Salient Object Detection

We formulate the salient object detection problem as a binary labeling problem by separating the salient object from the background. In the Conditional Random Field (CRF) framework [13], the probability of the label $A = \{a_x\}$ given the observation image I is directly modeled as a conditional distribution $P(A|I) = \frac{1}{Z}\exp(-E(A|I))$, where Z is the partition function. To detect a salient object, we define the energy $E(A|I)$ as a linear combination of a number of K salient features $F_k(a_x, I)$ and a pairwise feature $S(a_x, a_{x'}, I)$:

$$E(A|I) = \sum_x \sum_{k=1}^{K} \lambda_k F_k(a_x, I) + \sum_{x,x'} S(a_x, a_{x'}, I), \quad (3)$$

where λ_k is the weight of the kth feature, and x, x' are two adjacent pixels. Compared with Markov Random Field (MRF), one of advantages of CRF is that the feature functions $F_k(a_x, I)$ and $S(a_x, a_{x'}, I)$ can use arbitrary low-level or high-level features extracted from the whole image. CRF also provides an elegant framework to combine multiple features with effective learning.

Salient object feature. $F_k(a_x, I)$ indicates whether or not a pixel x belongs to the salient object. In next section, we propose a set of local, regional, and global salient object features to detect the salient object. Each kind of salient object feature provides a normalized feature map $f_k(x, I) \in [0, 1]$ for every pixel. The salient object feature $F_k(a_x, I)$ is defined as follows:

$$F_k(a_x, I) = \begin{cases} f_k(x, I) & a_x = 0 \\ 1 - f_k(x, I) & a_x = 1 \end{cases}. \quad (4)$$

Pairwise feature. $S(a_x, a_{x'}, I)$ models the spatial relationship between two adjacent pixels. Following the contrast-sensitive potential function in interactive image segmentation [2], we define $S(a_x, a_{x'}, I)$ as:

$$S(a_x, a_{x'}, I) = |a_x - a_{x'}| \cdot \exp(-\beta d_{x,x'}), \quad (5)$$

where $d_{x,x'} = ||I_x - I_{x'}||$ is the L2 norm of the color difference. β is a robust parameter that weights the color contrast, and can be set as $\beta = (2\langle||I_x - I_{x'}||^2\rangle)^{-1}$ [1], where $\langle\cdot\rangle$ is the expectation operator. This feature function is a penalty term when adjacent pixels are assigned with different labels. The more similar the colors of the two pixels are, the less likely they are assigned different labels. With this pairwise feature for segmentation, the homogenous interior region inside the salient object can also be labeled as salient pixels.

3.1. CRF Learning

To get an optimal linear combination of features, the goal of CRF learning is to estimate the linear weights $\vec{\lambda} = \{\lambda_k\}_{k=1}^K$ under the Maximized Likelihood (ML) criteria. Given N training image pairs $\{I^n, A^n\}_{n=1}^N$, the optimal parameters maximize the sum of the log-likelihood:

$$\vec{\lambda}^* = \arg\max_{\vec{\lambda}} \sum_n \log P(A^n|I^n; \vec{\lambda}). \quad (6)$$

The derivative of the log-likelihood with respect to the parameter λ_k is the difference between two expectations:

$$\frac{d \log P(A^n|I^n; \overrightarrow{\lambda})}{d\lambda_k} =$$

$$< F_k(A^n, I^n) >_{P(A^n|I^n;\overrightarrow{\lambda})} - < F_k(A^n, I^n) >_{P(A^n|G^n)} . \quad (7)$$

Then, the gradient descent direction is:

$$\Delta\lambda_k \propto \sum_n (\sum_{x,a_x^n} (F_k(a_x^n, I^n) p(a_x^n|I^n; \overrightarrow{\lambda}) - F_k(a_x^n, I^n) p(a_x^n|g_x^n))), \quad (8)$$

where $p(a_x^n|I^n; \overrightarrow{\lambda}) = \int_{A^n \setminus a_x^n} P(A_x^n|I^n; \overrightarrow{\lambda})$ is the marginal distribution and $p(a_x^n|g_x^n)$ is from the labeled ground-truth:

$$p(a_x^n|g_x^n) = \begin{cases} 1 - g_x^n & a_x = 0 \\ g_x^n & a_x = 1 \end{cases} .$$

Exact computation of marginal distribution $p(a_x^n|I^n; \overrightarrow{\lambda})$ is intractable. However, the pseudo-marginal (belief) computed by belief propagation can be used as a good approximation [21, 14]. The tree-reweighted belief propagation[12] can be run under the current parameters in each step of gradient descent to compute an approximation of the marginal distribution $p(a_x^n|I^n; \overrightarrow{\lambda})$.

4. Salient Object Features

In this section, we introduce local, regional, and global features that define a salient object. Since scale selection is one of the fundamental issues in feature extraction, we resize all images so that the max(width,height) of the image is 400 pixels. In the following, all parameters are set with respect to this basic image size.

4.1. Multi-scale contrast

Contrast is the most commonly used local feature for attention detection [10, 15, 16] because the contrast operator simulates the human visual receptive fields. Without knowing the size of salient object, contrast is usually computed at multiple scales. In this paper, we simply define the multi-scale contrast feature $f_c(x, I)$ as a linear combination of contrasts in the Gaussian image pyramid:

$$f_c(x, I) = \sum_{l=1}^{L} \sum_{x' \in N(x)} ||I^l(x) - I^l(x')||^2 \quad (9)$$

where I^l is the lth-level image in the pyramid and the number of pyramid levels L is 6. $N(x)$ is a 9×9 window. The feature map $f_c(\cdot, I)$ is normalized to a fixed range $[0, 1]$. An example is shown in Figure 5. Multi-scale contrast highlights the high contrast boundaries by giving low scores to the homogenous regions inside the salient object.

Figure 5. Multi-scale contrast. From left to right: input image, contrast maps at multiple scales, and the feature map from linearly combining the contrasts at multiple scales.

4.2. Center-surround histogram

As shown in Figure 2, the salient object usually has a larger extent than local contrast and can be distinguished from its surrounding context. Therefore, we propose a regional salient feature.

Suppose the salient object is enclosed by a rectangle R. We construct a surrounding contour R_S with the same area of R, as shown in Figure 6 (a). To measure how distinct the salient object in the rectangle is with respect to its surroundings, we can measure the distance between R and R_S using various visual cues such as intensity, color, and texture/texton. In this paper, we use the χ^2 distance between histograms of RGB color: $\chi^2(R, R_S) = \frac{1}{2} \sum \frac{(R^i - R_S^i)^2}{R^i + R_S^i}$. We use histograms because they are robust global description of appearance. They are insensitive to small changes in size, shape, and viewpoint. Another reason is that the histogram of a rectangle with any location and size can be very quickly computed by means of integral histogram introduced recently [20]. Figure 6 (a) shows that the salient object (the girl) is most distinct using the χ^2 histogram distance. We have also tried the intensity histogram and oriented gradient histogram. We found that the former is redundant with the color histogram and the latter is not a good measurement because the texture distribution in a semantic object is usually not coherent.

To handle varying aspect ratios of the object, we use five templates with different aspect ratios $\{0.5, 0.75, 1.0, 1.5, 2.0\}$. We find the most distinct rectangle $R^*(x)$ centered at each pixel x by varying the size and aspect ratio:

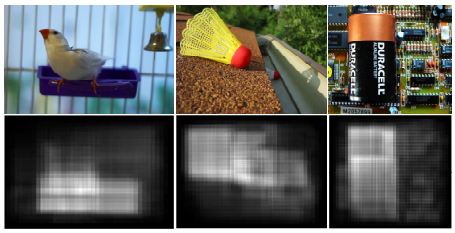
$$R^*(x) = \arg\max_{R(x)} \chi^2(R(x), R_S(x)). \quad (10)$$

The size range of the rectangle $R(x)$ is set to $[0.1, 0.7] \times \min(w, h)$, where w, h are image width and height. Then, the center-surround histogram feature $f_h(x, I)$ is defined as a sum of spatially weighted distances:

$$f_h(x, I) \propto \sum_{\{x'|x \in R^*(x')\}} w_{xx'} \chi^2(R^*(x'), R_S^*(x')), \quad (11)$$

where $R^*(x')$ is the rectangle centered at x' and containing the pixel x. The weight $w_{xx'} = \exp(-0.5\sigma_{x'}^{-2}||x - x'||^2)$ is a Gaussian falloff weight with variance $\sigma_{x'}^2$, which is set to one

(a)



(b)

Figure 6. Center-surround histogram. (a) center-surround histogram distances with different locations and sizes. (b) top row are input images and bottom row are center-surround histogram feature maps.

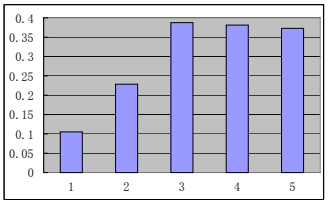


Figure 7. The average center-surround histogram distance on the image set $\mathcal{A}$. 1. a randomly selected rectangle. 2. a rectangle centered at the image center with 55% ratio of area to image. 3-5. rectangles labeled by three users.

third of the size of $R^*(x')$. Finally, the feature map $f_h(\cdot, I)$ is also normalized to the range $[0, 1]$.

Figure 6 (b) shows several center-surround feature maps. The salient objects are well located by the center-surround histogram feature. Especially, the last image in Figure 6 (b) is a difficult case for color or contrast based approaches but the center-surround histogram feature can capture the "object-level" salient region.

To further verify the effectiveness of this feature, we compare the center-surround histogram distance of a randomly selected rectangle, a rectangle centered at the image center, and three user-labeled rectangles in the image. Figure 7 shows the average distances on the image set $\mathcal{A}$. It is no surprise that salient object has a large center-surround histogram distance.

4.3. Color spatial-distribution

The center-surround histogram is a regional feature. Is there a global feature related to the salient object? We observe from Figure 2 that the wider a color is distributed in the image, the less possible a salient object contains this color. The global spatial distribution of a specific color can be used to describe the saliency of an object.

To describe the spatial-distribution of a specific color, the simplest approach is to compute the spatial variance of the color. First, all colors in the image are represented by Gaussian Mixture Models (GMMs) $\{w_c, \mu_c, \Sigma_c\}_{c=1}^{C}$, where $\{w_c, \mu_c, \Sigma_c\}$ is the weight, the mean color and the covariance matrix of the cth component. Each pixel is assigned to a color component with the probability:

$$p(c|I_x) = \frac{w_c \mathcal{N}(I_x|\mu_c, \Sigma_c)}{\sum_c w_c \mathcal{N}(I_x|\mu_c, \Sigma_c)}. \quad (12)$$

Then, the horizontal variance $V_h(c)$ of the spatial position for each color component c is:

$$V_h(c) = \frac{1}{|X|_c} \sum_x p(c|I_x) \cdot |x_h - M_h(c)|^2, \quad (13)$$

$$M_h(c) = \frac{1}{|X|_c} \sum_x p(c|I_x) \cdot x_h, \quad (14)$$

where x_h is x-coordinate of the pixel x, and $|X|_c = \sum_x p(c|I_x)$. The vertical variance $V_v(c)$ is similarly defined. The spatial variance of a component c is $V(c) = V_h(c) + V_v(c)$. We normalized $\{V(c)\}_c$ to the range $[0, 1]$ ($V(c) \leftarrow (V(c) - \min_c V(c))/(\max_c V(c) - \min_c V(c))$). Finally, the color spatial-distribution feature $f_s(x, I)$ is defined as a weighted sum:

$$f_s(x, I) \propto \sum_c p(c|I_x) \cdot (1 - V(c)). \quad (15)$$

The feature map $f_s(\cdot, I)$ is also normalized to the range $[0, 1]$. Figure 8 (b) shows color spatial-distribution feature maps of several example images. The salient objects are well covered by this global feature. Note that the spatial variance of the color at the image corners or boundaries may be also small because the image is cropped from the whole scene. To reduce this artifact, a center-weighted, spatial-variance feature is defined as:

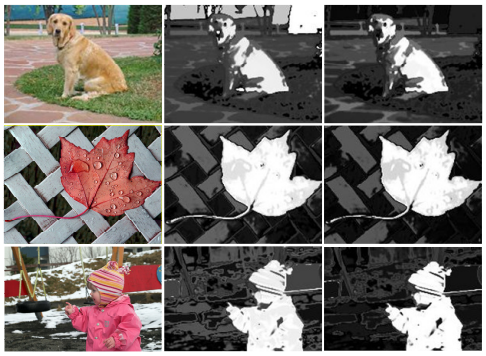
$$f_s(x, I) \propto \sum_c p(c|I_x) \cdot (1 - V(c)) \cdot (1 - D(c)), \quad (16)$$

where $D(c) = \sum_x p(c|I_x) d_x$ is a weight which assigns less importance to colors nearby image boundaries and is also normalized to $[0, 1]$, similar to $V(c)$. d_x is the distance from pixel x to the image center. As shown in Figure 8 (c), center-weighted, color spatial variance shows a better prediction of the saliency of each color.

To verify the effectiveness of this global feature, we plot the color spatial-variance versus average saliency probability curve on the image set $\mathcal{A}$, as shown in Figure 9. Obviously, the smaller a color variance is, the higher probability the color belongs to the salient object.

5. Evaluation

We randomly select 2,000 images from image set $\mathcal{A}$ and 1,000 images from image set $\mathcal{B}$ to construct a training set,



(a) (b) (c)

Figure 8. Color spatial-distribution feature. (a) input images. (b) color spatial variance feature maps. (c) center-weighted, color spatial variance feature maps.

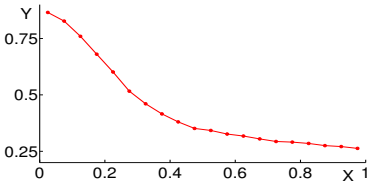


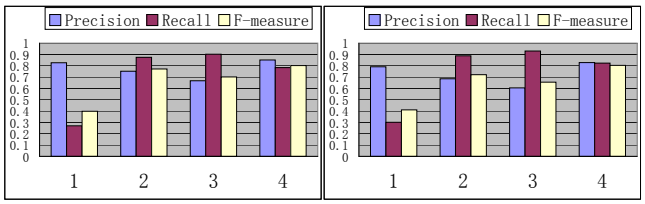
Figure 9. Color spatial variance (x-coordinate) v.s. average saliency probability (y-coordinate) on the image set $\mathcal{A}$. The saliency probability is computed from the "ground truth" labeling.

which are excluded from the testing phase. To output a rectangle for the evaluation, we exhaustively search for a smallest rectangle containing at least 95% salient pixels in the binary label map produced by the CRF.

Effectiveness of features and CRF learning. To evaluate the effectiveness of each salient object feature, we trained four CRFs: three CRFs with individual features and one CRF with all three features. Figure 10 shows the precision, recall, and F-measure of these CRFs on the image sets $\mathcal{A}$ and $\mathcal{B}$. As can be seen, the multi-scale contrast feature has a high precision but a very low recall. The reason is that the inner homogenous region of a salient object has low contrast. The center-surround histogram has the best overall performance (on F-measure) among all individual features. This regional feature is able to detect the whole salient object, although the background region may contain some errors. The color spatial-distribution has slightly lower precision but has the highest recall. Later, we will discuss that for attention detection, recall rate is not as important as precision. It demonstrates the strength and weakness of the global feature. After CRF learning, the CRF with all three features produces the best result, as shown in the last bars in Figure 10. The best linear weights we learnt are: $\overrightarrow{\lambda} = \{0.24, 0.54, 0.22\}$.

Figure 11 shows the feature maps and labeling results of several examples. Each feature has its own strengths and limitations. By combining all features with the pairwise feature, the CRF successfully locates the most salient object.

Comparison with other approaches. We compare our



(a) image set $\mathcal{A}$ (b) image set $\mathcal{B}$

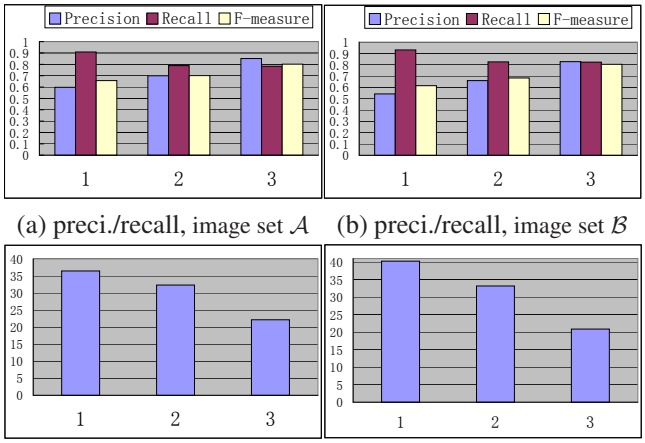
Figure 10. Evaluation of salient object features. 1. multi-scale contrast. 2. center-surround histogram. 3. color spatial distribution. 4. combination of all features.

Figure 11. Examples of salient features. From left to right: input image, multi-scale contrast, center-surround histogram, color spatial distribution, and binary salient mask by CRF.

algorithm with two leading approaches. One is the contrast and fuzzy growing based method [16], which we call "FG". This approach directly outputs a rectangle. Another approach is the salient model presented in [10], and we call it "SM" (We use a matlab implementation from http://www.saliencytoolbox.net). Because the output of this approach is a salient map, we convert the salient map to a rectangle containing 95% of the fixation points, which are determined by the winner-take-all algorithm [10].

Figure 12 shows the evaluation results of three algorithms on both image set $\mathcal{A}$ and $\mathcal{B}$. On image set $\mathcal{A}$, our approach reduced 42% and 34% overall error rates on F-measure, and 39% and 31% boundary displacement errors (BDEs), compared with FG and SM. Similarly, 49% and 38% overall error rates on F-measure, and 48% and 37% BDEs are reduced on the image set $\mathcal{B}$.

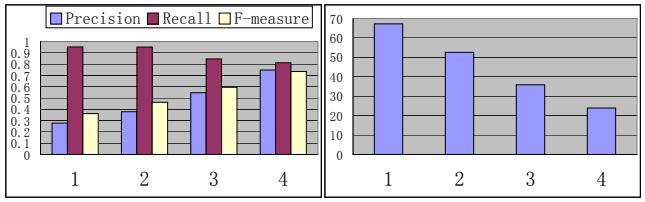
Notice that as show in Figure 10 and 12, the individual features (center-surround histogram and color spatial-distribution), FG, and SM all have higher recall rates than our final approach. In fact, recall rate is not much of a useful measure in attention detection. For example, a 100% recall rate can be achieved by simply select the whole image. So algorithm trying to achieve a high recall rate tends to select as large an attention region as possible sacrificing the precision rate. The key objective of attention detection should be to locate position of a salient object as accurately as possible, i.e. with high precision. However, for images with a large salient object, a high precision is also not too difficult to achieve. Again, for example, for an image with a salient ob-



(a) preci./recall, image set $\mathcal{A}$ (b) preci./recall, image set $\mathcal{B}$

(c) BDE, image set $\mathcal{A}$ (d) BDE, image set $\mathcal{B}$

Figure 12. Comparison of different algorithms. (a-b) and (c-d) are region-based (precision, recall, and F-measure) and boundary-based (BDE - boundary displacement error) evaluations. 1. FG. 2. SM. 3. our approach.



(a) preci./recall (b) BDE

Figure 13. Comparison on a small object (object/image ratio $\in [0, 0.25]$) dataset from image set $\mathcal{A}$. 1. a rectangle centered at the image center and with 0.6 object/image ratio. 2. FG. 3. SM. 4. our approach.

ject occupying 80% of the image area, just select the whole image as attention area will give 80% precision with 100% recall rate. So the real challenge for attention detection is to achieve high precision on small salient objects. To construct such a challenge data set, we select a small object subset with object/image ratio in the range $[0, 0.25]$ from the image set $\mathcal{A}$. The results on this small object dataset are shown in Figure 13, where we also show the performance of a rectangle fixed at the image center with 0.6 object/image ratio. Notice that both this center rectangle and FG achieve high recall rate but with very low precision and large BDE. Our method is significantly better than FG and SM in both precision (97% and 37% improvement) and BDE (55% and 33% reduction). Figure 14 shows several examples with ground truth rectangles from one user for a qualitative comparison. We can see that FG and SM approaches tend to produce a larger attention rectangle and our approach is much more precise.

Figure 15 shows our detection results on the images in Figure 2. The whole labeled database and our results are public available from: http://research.microsoft.com/~jiansun/.

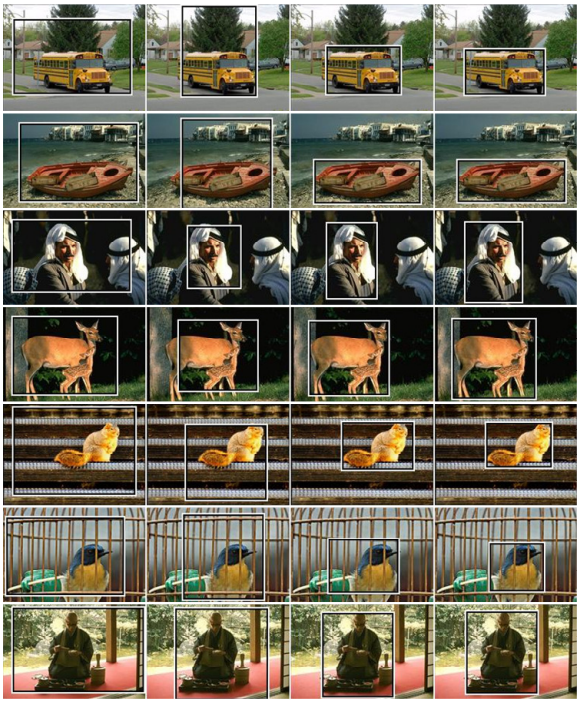


Figure 14. Comparison of different algorithms. From left to right: FG, SM, our approach, and ground-truth.

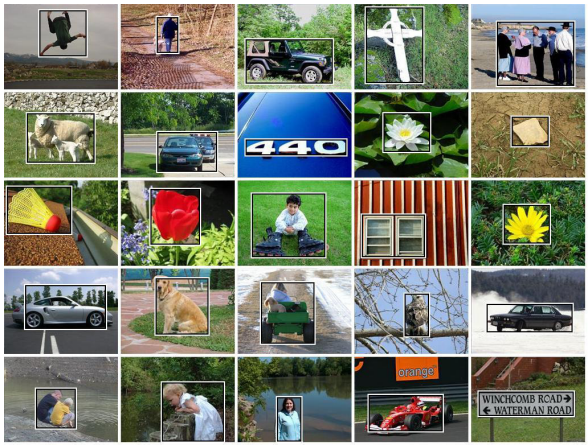


Figure 15. Our detection result on the images in Figure 2.

(a) (b)

Figure 16. Multiple salient object detection. (a) Two birds are detected at the same time. (b) The toy car is detected firstly, and using the updated feature maps, the boy is detected secondly.

6. Discussion and Conclusion

In this paper, we have presented a supervised approach for salient object detection, which is formulated as an image seg-

Figure 17. Failure cases. From left to right: FG, SM, our approach, and ground-truth.

mentation problem using a set of local, regional, and global salient object features. A CRF was learnt and evaluated on a large image database containing 20,000+ well-labeled images by multiple users.

Salient object detection has wider applications. For example, a more semantic, object-based image similarity can be defined with salient object detection for content-based image retrieval. Manually collecting and labeling training images in object recognition is very costly. With salient object detection, it is possible to automatically collect and label a substantial number of images.

There are several important remaining issues for further investigation of salient object detection. In future work, we plan to experiment with non-rectangular shapes for salient objects, and non-linear combination of features. More sophisticated visual features will further improve the performance. In particular, we are extending our single salient object detection framework to detect multiple salient objects or no salient object at all. Figure 16 shows two initial results. In Figure 16 (a), our current CRF approach can directly output two disjoint connected components so that we can easily detect them simultaneously. In Figure 16 (b), we use the inhibition-of-return strategy [10] to detect the salient objects one-by-one. Finally, Figure 17 shows two failure cases, which demonstrate one of the challenges in the salient object detection — hierarchical salient object detection.

Acknowledgments This work is performed when Tie Liu visited Microsoft Research Asia. Tie Liu and Nan-Ning Zheng were supported by a grant from the National Science Foundation of China (No.60021302).

References

[1] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. In *ECCV*, pages 428–441, 2004.

[2] Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *ICCV*, pages 105–112, 2001.

[3] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, pages 155–162, 2005.

[4] L. Chen, X. Xie, X. Fan, W. Ma, H. Shang, and H. Zhou. A visual attention mode for adapting images on small displays. Technical report, Microsoft Research, Redmond, WA, 2002.

[5] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi. Yet another survey on image segmentation: Region and boundary information integration. In *ECCV*, pages 408–422, 2002.

[6] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, 2006.

[7] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention*. PhD thesis, California Institute of Technology Pasadena, 2000.

[8] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *NIPS*, pages 547–554, 2005.

[9] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *CVPR*, pages 631–637, 2005.

[10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on PAMI*, 20(11):1254–1259, 1998.

[11] C. Koch and S. Ullman. Shifts in selection in visual attention: Toward the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.

[12] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. on PAMI*, 28(10):1568–1583, 2006.

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[14] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. In *ECCV*, pages 581–594, 2006.

[15] F. Liu and M. Gleicher. Region enhanced scale-invariant saliency detection. In *Proceedings of IEEE ICME*, 2006.

[16] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of ICMM*, pages 374–381, 2003.

[17] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. on PAMI*, 26(5):530–549, 2004.

[18] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *CVPR*, pages 2049–2056, 2006.

[19] O.L.Meur, O.L.Callet, D.Barba, and D.Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Trans. on PAMI*, 28(5):802–817, 2006.

[20] F. Porkili. Integral histogram: A fast way to extract histograms in cartesian spaces. In *CVPR*, pages 829– 836, 2005.

[21] X. Ren, C. Fowlkes, and J. Malik. Cue integration for figure/ground labeling. In *NIPS*, pages 1121–1128, 2005.

[22] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *CVPR*, pages 37–44, 2004.

[23] A. Santella, M. Agrawala, D. Decarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI*, pages 771–780, 2006.

[24] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[25] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo. Modelling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.

[26] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition - a gentle way. In *Biol. Motivated Comp. Vision*, 2002.