Learning to Detect a Salient Object

Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, *Fellow*, *IEEE*, Xiaoou Tang, *Fellow*, *IEEE*, and Heung-Yeung Shum, *Fellow*, *IEEE*

Abstract—In this paper, we study the salient object detection problem for images. We formulate this problem as a binary labeling task where we separate the salient object from the background. We propose a set of novel features, including multiscale 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. Further, we extend the proposed approach to detect a salient object from sequential images by introducing the dynamic salient features. We collected a large image database containing tens of thousands of carefully labeled images by multiple users and a video segment database, and conducted a set of experiments over them to demonstrate the effectiveness of the proposed approach.

Index Terms—Salient object detection, conditional random field, visual attention, saliency map.

1 Introduction

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 [1], adaptive image display on small devices [2], image/video compression, advertising design [3], and image collection browsing [4]. Recent studies [5], [6], [7] demonstrated that visual attention helps object recognition, tracking, and detection as well. In this paper, we study one aspect of visual attention—salient object detection. Fig. 1 shows some examples of salient objects.

For instance, people are usually interested in the objects in images in Fig. 1, and 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, or objects with the most interest. In many applications, such as image display on small devices [2] and image collection browsing [4], people want to show the regions with the most interest, or the salient objects. In this paper, we try to locate these salient objects automatically with the supposition that a salient object exists in an image.

- *T. Liu is with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, and the Analytics and Optimization Department, IBM Research-China, Building 19A2F, Zhongguancun Software Park, 8 Dongbeiwang West Road, Haidian District, Beijing 100193, P.R. China. E-mail: liultie@cn.ibm.com.*
- *Z. Yuan and N. Zheng are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, 28 Xianning Xilu, Xi'an 710049, China. E-mail: yzejian@gmail.com, nnzheng@mail.xjtu.edu.cn.*
- *J. Sun is with the Visual Computing Group, Microsoft Research Asia, 5/F, Beijing Sigma Center, No. 49, Zhichun Road, Haidian District, Beijing 100190, P.R. China. E-mail: jiansun@microsoft.com.*
- *J. Wang is with the Media Computing Group, Microsoft Research Asia, 5/F, Beijing Sigma Center, No. 49, Zhichun Road, Haidian District, Beijing 100190, P.R. China. E-mail: jingdw@microsoft.com.*
- *X. Tang is with the Department of Information Engineering, Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: xtang@ie.cuhk.edu.hk.*
- *H.-Y. Shum is with the On-Line Service Division, R&D, Microsoft, One Microsoft Way, Redmond, WA 98052. E-mail: hshum@microsoft.com.*

Manuscript received 4 Dec. 2008; revised 23 Oct. 2009; accepted 29 Nov. 2009; published online 2 Mar. 2010.
Recommended for acceptance by A. Torralba.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-12-0834.
Digital Object Identifier no. 10.1109/TPAMI.2010.70.

1.1 Related Work

Most existing visual attention approaches are based on the bottom-up computational framework [8], [9], [10], [11], [12], [13], [14], [15], [16], where visual attention is supposed to be 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 [13], self-information [8], or graph-based random walk [9] using multiple features. After normalization and linear/nonlinear combination, a master map [17] or a saliency map [14] 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 nonlinear operations. Recently, a saliency model based on low, middle, and high-level image features was trained using the collected eye tracking data [18]. While these approaches have worked well in finding a few fixation locations in synthetic and natural images, they have not been able to accurately detect where the salient object should be.

For instance, the middle row in Fig. 1 shows three saliency maps computed using Itti's algorithm [13]. Note that the visual saliency concentrates on several small local regions with high-contrast structures, e.g., the background grid in Fig. 1a, the shadow in Fig. 1b, and the foreground boundary in Fig. 1c. Although the leaf in Fig. 1a commands much attention, the saliency for the leaf is low. Therefore, these saliency maps computed from low-level features don't have the notation of objects, and they are not good indications for where a salient object is located while perusing these images.

Figure-ground segregation is somehow related to salient object detection. However, the usually figure-ground

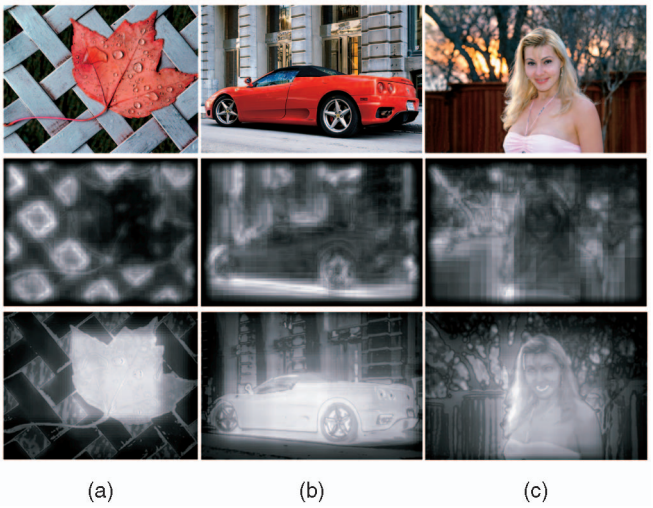


Fig. 1. Salient object detection. From top to bottom: input image with a salient object, saliency map computed by Itti's attention algorithm (http://www.saliencytoolbox.net), and saliency map computed by our salient object detection approach.

Fig. 2. Sample images in our image database for salient object detection.

segregation algorithm works with the supposition of the category of objects [19], [20], [21] or with interactions [22], [23]. If the object is assigned a given category, the specific features, for example, for cows, can be defined specially, and these features cannot be adopted for other categories. For interactive figure-ground segmentation, the appearance model is usually set up, where for our salient object detection, we do not have such an appearance model.

Visual attention is also studied for sequential images, where the spatiotemporal cues from image sequences are indicated to be helpful for visual attention detection. For instance, motion from objects or backgrounds helps to indicate the salient fixations [24], [25], [26]. Large motion [27] and motion contract [24] are supposed to induce prominent attention, respectively. Usually, the visual saliency from a single image is combined with the motion saliency for better visual attention detection, and different combination strategies are introduced in [27]. Video surprising [11] is also related, where it describes the KullbackLeibler divergence between the prior and posterior distribution of a feature map. These visual attention approaches suffer from the similar shortcoming to the visual attention approaches for single image. Automatic object discovery [28], [29], [30] deals with a similar salient object detection task for sequential images. The objects are extracted and tracked using motion-based layer segmentation in [28] and a generative model of objects by defining switch variables for combinatorial model selection is adopted in [29]. The unsupervised video object discovery [30] combines the topic model and the temporal model for videos.

1.2 Our Approach

In this paper, we investigate one aspect of visual attention, namely, salient object detection. We incorporate the high-level concept of the salient object into the process of saliency map computation. As can be observed in Fig. 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 or sequential images. First, we model the salient object detection problem by a condition random field (CRF), where a group of salient features are combined through 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 Fig. 1 shows the saliency maps computed by our approach. 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 salient features to define a generic salient object. We also define the salient features on the motion field similarly to capture the spatiotemporal cues. Then, we construct a large image database with 20,000+ well-labeled images for training and evaluation. To the best of our knowledge, it is the first time a large image database has been made available for quantitative evaluation.

The remainder of the paper is organized as follows: Section 2 introduces the formulation of the salient object detection problem, and the salient object features are presented in Section 3. Section 4 introduces the image database and the evaluation experiments. Section 5 discusses the connections between our approach and related approaches, and the conclusion follows in Section 6.

2 Formulation

Given an image I, we represent the salient object as a binary mask $A = \{a_x\}$. For each pixel x, $a_x \in \{1, 0\}$ is a binary label to indicate whether the pixel x belongs to the salient object. Similarly, the salient objects in sequential images, $\{I_1, \ldots, I_t, \ldots, I_N\}$, are represented by a sequence of binary masks $\{A_1, \ldots, A_t, \ldots, A_N\}$, with A_t corresponding to image I_t.

In this paper, we formulate the salient object detection problem as a binary labeling task by inspecting whether each pixel belongs to the salient object. We first present the conditional random field formulation to the single-image

case, and then extend it to the sequential image case by exploring the extra temporal information.

2.1 Formulation of Salient Object Detection in a Single Image

In the CRF framework [31], the probability of a labeling configuration $A = \{a_x\}$, given the observation image I, is modeled as a conditional distribution $P(A|I) = \frac{1}{Z}\exp(-E(A|I))$, where Z is the partition function. We define the energy $E(A|I)$ as a linear combination of a set of static salient features, including a number of K unary features $F_k(a_x, I)$ and a pairwise feature $S(a_x, a_{x'}, I)$:

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

where λ_k is the weight of the kth feature and x, x' are two adjacent pixels. Compared with Markov random field, one of the advantages of CRF is that the features $F_k(a_x, I)$ and $S(a_x, a_{x'}, I)$ can be arbitrary low-level or high-level features extracted from the whole image. CRF also provides an elegant framework to learn an optimal combination of multiple features.

Salient object feature. $F_k(a_x, I)$ indicates whether a pixel x belongs to the salient object. In the next section, we propose a set of local, regional, and global salient object features. The salient object feature $F_k(a_x, I)$ is formulated from a normalized feature map $f_k(x, I) \in [0, 1]$ for every pixel, and is written as follows:

$$F_k(a_x, I) = \begin{cases} f_k(x, I), & a_x = 0, \\ 1 - f_k(x, I), & a_x = 1. \end{cases} \quad (2)$$

Pairwise feature. $S(a_x, a_{x'}, I)$ exploits the spatial relationship between two adjacent pixels. Following the contrast-sensitive potential function in interactive image segmentation [22], we define $S(a_x, a_{x'}, I)$ as

$$S(a_x, a_{x'}, I) = |a_x - a_{x'}| \cdot \exp(-\beta d_{x,x'}), \quad (3)$$

where $d_{x,x'} = \|I_x - I_{x'}\|_2$ is the L2-norm of the color difference, β is a robust parameter that weights the color contrast and can be set as $\beta = (2\langle\|I_x - I_{x'}\|^2\rangle)^{-1}$ [32], with $\langle\cdot\rangle$ being the expectation operator. This feature function can be viewed as a penalty term when adjacent pixels are assigned with different labels. The more similar the colors of the two pixels are, the less likely it is that they are assigned different labels.

2.2 Formulation of Salient Object Detection in Sequential Images

We exploit the extra temporal cues to formulate salient object detection in sequential images. Besides the static salient features from a single image, the temporal features, called dynamic features, are further defined. Differently from previous work [24], [25], [26], we propose new dynamic features and learn a CRF model to combine the dynamic features and static features. Instead of building a complex 3D graph formulation, e.g., a large graph in interactive video cutout [33], [34], we integrate the cues from multiple images into a 2D graph for effective and efficient optimization.

Given the sequential images $\{I_t\}$, $t \in \{1, \ldots, N\}$, the probability of the sequential binary maps, $\{A_t\}$, $t \in \{1, \ldots, N\}$, can be modeled as a conditional distribution:

$$P(A_{1,\ldots,N}|I_{1,\ldots,N}) = \frac{1}{Z}\exp(-E(A_{1,\ldots,N}|I_{1,\ldots,N})), \quad (4)$$

where Z is the partition function. A reasonable supposition is that the salient object detection A_t can be inferred from the associated frame I_t and the previous frame I_{t-1}. Then, the energy function $E(A_{1,\ldots,N}|I_{1,\ldots,N})$ can be decomposed as

$$E(A_{1,\ldots,N}|I_{1,\ldots,N}) = \sum_{t=1}^{N} E(A_t|I_{1,\ldots,N}) = \sum_{t=1}^{N} E(A_t|I_{t-1}, I_t). \quad (5)$$

Here, $E(A_t|I_{t-1}, I_t)$ is composed of a static term and a dynamic term. The static term is the same as the single-image case. In the dynamic term, we compute a motion field M_t from a pair of successive images I_{t-1} and I_t, and build salient features from the motion field, and in addition, introduce an appearance coherent feature between the salient objects in the successive frames. Specifically, the energy $E(A_t|I_{t-1}, I_t)$ is formulated as a linear combination of static salient features $F_k(a_x, I_t)$, a pairwise feature $S(a_x, a_{x'}, I)$, and a set of dynamic salient features, including motion salient features $F_k(a_x, M_t)$ and appearance coherent features $F_k(a_x, I_{t-1}, I_t)$:

$$E(A_t|I_{t-1}, I_t) = \sum_x \left(\sum_{k=1}^{K} \lambda_k F_k(a_x, I_t) + \sum_{k=K+1}^{K+L} \lambda_k F_k(a_x, M_t) + \lambda_0 F(a_x, I_{t-1}, I_t) \right) + \sum_{x,x'} S(a_x, a_{x'}, I_t), \quad (6)$$

where $\{\lambda_k\}$ are the weights of the features, M_t is the motion field corresponding to image I_t, and x, x' are two adjacent pixels in image I_t. $F_k(a_x, I_t)$ are the static salient features, and $S(a_x, a_{x'}, I_t)$ describes the spatial relationship between two adjacent pixels. These two categories of features are defined as in (1). Differently from (1), more features from the temporal information are included. These are the motion salient features $F_k(a_x, M_t)$ from the motion field M_t and the appearance coherent feature $F(a_x, I_{t-1}, I_t)$ between the salient objects from two adjacent frames.

Motion salient feature. $F_k(a_x, M_t)$ is defined, similarly to (2), as the indicator of a normalized feature map $f_k(x, M_t) \in [0, 1]$, where M_t is the motion field of the image I_t and obtained based on the SIFT flow technique [35].

Appearance coherent feature. $F(a_x, I_{t-1}, I_t)$ models the appearance coherence of the salient objects from two adjacent frames, which is defined as an indicator of a normalized feature map $f(x, I_{t-1}, I_t) \in [0, 1]$, similarly to (2). This feature function $f(x, I_{t-1}, I_t)$ penalizes the pixels that are identified to be in the salient object, but with a large color difference between the surrounding regions from two adjacent frames. With this appearance coherent feature, the salient objects from two adjacent frames can be labeled more consistently.

2.3 Learning and Inference for The CRF Model

The objective functions of the salient object detection for single-image and sequential-image cases in (1) and (6) are essentially very similar to the perspective of the CRF formulation, i.e., a linear combination of a set of features. To get the linear combination of features, the goal of CRF

learning is to estimate the linear weights $\overrightarrow{\lambda} = \{\lambda_k\}$ under the Maximized Likelihood (ML) criteria. In the following, we present the parameter learning scheme for the single-image case. The parameter learning scheme for the sequential image case can be similarly obtained. Given N training image pairs $\{I^n, A^n\}_{n=1}^N$, the optimal parameters maximize the sum of the log-likelihood:

$$\overrightarrow{\lambda^*} = \arg\max_{\overrightarrow{\lambda}} \sum_n \log P(A^n|I^n; \overrightarrow{\lambda}). \tag{7}$$

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

$$\frac{d \log P(A^n|I^n; \overrightarrow{\lambda})}{d\lambda_k} = < F_k(A^n, I^n) >_{P(A^n|I^n; \overrightarrow{\lambda})} - < F_k(A^n, I^n) >_{P(A^n|G^n)}. \tag{8}$$

Then, the gradient descent direction is

$$\Delta\lambda_k \propto \sum_n \Big(\sum_{x, a_x^n} \big(F_k(a_x^n, I^n) p(a_x^n|I^n; \overrightarrow{\lambda}) - F_k(a_x^n, I^n) p(a_x^n|g_x^n)\big) \Big), \tag{9}$$

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 g_x^n and is defined as

$$p(a_x^n|g_x^n) = \begin{cases} 1 - g_x^n, & a_x = 0, \\ g_x^n, & a_x = 1. \end{cases} \tag{10}$$

Exact computation of marginal distribution $p(a_x^n|I^n; \overrightarrow{\lambda})$ is intractable. However, the pseudomarginal (belief) computed by belief propagation can be used as a good approximation [36], [19]. The tree-reweighted belief propagation [37] can be run under the current parameters in each step of gradient descent to compute an approximation of the marginal distribution $p(a_x^n|I^n; \overrightarrow{\lambda})$.

When the combination parameters of salient features are learned, we can infer the most probable labeling A to minimize the energy from (1) and (6). We still apply the tree-reweighted belief propagation to infer the label using the learned parameters, and we will discuss the details of implementations in Section 4.

3 Salient Object Feature

In this section, we instantiate the formulation of salient object detection by presenting the salient object features: static salient features for the single-image case and dynamic salient features specifically for the sequential images.

3.1 Static Salient Feature

We introduce local, regional, and global features that define a salient object. Since the 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.

Fig. 3. Multiscale contrast. From left to right: input image, contrast maps at multiple scales, and the feature map from linearly combining the contrasts at multiple scales.

3.1.1 Multiscale Contrast

Contrast is the most commonly used local feature for attention detection [13], [38], [39] because the contrast operator simulates the human visual receptive fields. Without knowing the size of the salient object, contrast is usually computed at multiple scales. In this paper, we simply define the multiscale contrast feature $f_c(x, I)$ as a linear combination of contrasts in the Gaussian image pyramid:

$$f_c(x, I) = \sum_{l=1}^{L} \sum_{x' \in N(x)} \|I^l(x) - I^l(x')\|^2, \tag{11}$$

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 Fig. 3. Multiscale contrast highlights the high-contrast boundaries by giving low scores to the homogenous regions inside the salient object.

3.1.2 Center-Surround Histogram

As shown in Fig. 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 Fig. 4a. 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 a 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 an integral histogram introduced recently [40]. Fig. 4a shows that the salient object (the girl) is most distinct using the χ^2 histogram distance. We have also tried the intensity histograms and histograms of oriented gradient [41]. 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 $\{0.5, 0.75, 1.0, 1.5, 2.0\}$.

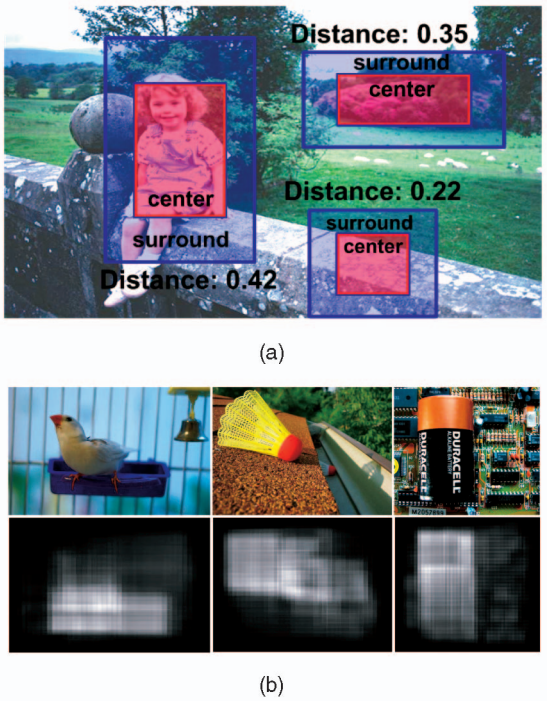


Fig. 4. Center-surround histogram. (a) Center-surround histogram distances with different locations and sizes. (b) Top row are input images and bottom row are center-surround histogram feature maps.

We find the most distinct rectangle, $R^*(x)$, centered at each pixel x by varying the size and aspect ratio:

$$R^*(x) = \arg\max_{R(x)} \chi^2(R(x), R_S(x)). \tag{12}$$

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

$$f_h(x, I) \propto \sum_{\{x' | x \in R^*(x')\}} w_{xx'} \chi^2(R^*(x'), R^*_S(x')), \tag{13}$$

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 $\sigma^2_{x'}$, which is set to one-third of the size of $R^*(x')$. Finally, the feature map $f_h(\cdot, I)$ is also normalized to the range $[0, 1]$.

Fig. 4b shows several center-surround feature maps. The salient objects are well located by the center-surround histogram feature. The last image in Fig. 4b is an especially 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. Fig. 5 shows the average distances on the image set $\mathcal{A}$, and this image set is introduced in Section 4. It is no surprise that the salient object has a large center-surround histogram distance.

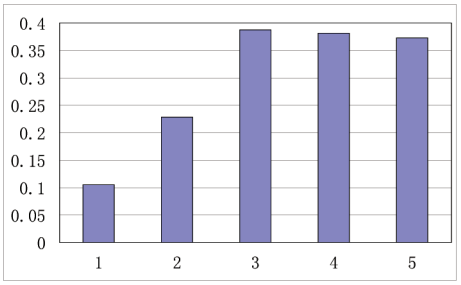


Fig. 5. The average center-surround histogram distance on the image set $\mathcal{A}$. 1. A randomly selected rectangle. 2. A rectangle centered at the image center with 55 percent ratio of area to image. 3-5. Rectangles labeled by three users.

3.1.3 Color Spatial Distribution

The center-surround histogram is a regional feature. Is there a global feature related to the salient object? We observe from Fig. 2 that the more widely a color is distributed in the image, the less possible it is that 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:

$$p(c|I_x) = \frac{w_c \mathcal{N}(I_x|\mu_c, \Sigma_c)}{\sum_c w_c \mathcal{N}(I_x|\mu_c, \Sigma_c)}. \tag{14}$$

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

$$V_h(c) = \frac{1}{|X|_c} \sum_x p(c|I_x) \cdot |x_h - M_h(c)|^2, \tag{15}$$

$$M_h(c) = \frac{1}{|X|_c} \sum_x p(c|I_x) \cdot x_h, \tag{16}$$

where x_h is the 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:

$$f_s(x, I) \propto \sum_c p(c|I_x) \cdot (1 - V(c)). \tag{17}$$

The feature map $f_s(\cdot, I)$ is also normalized to the range $[0, 1]$. Fig. 6b 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 also be small because the image is cropped from the whole scene. To reduce this artifact, a center-weighted, spatial-variance feature is defined as

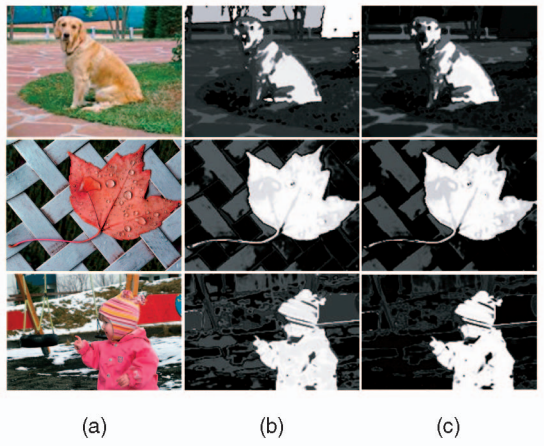


Fig. 6. Color spatial-distribution feature. (a) Input images. (b) Color spatial variance feature maps. (c) Center-weighted, color spatial variance feature maps.

$$f_s(x, I) \propto \sum_c p(c|I_x) \cdot (1 - V(c)) \cdot (1 - D(c)), \tag{18}$$

where $D(c) = \sum_x p(c|I_x)d_x$ is the weight which assigns less importance to colors nearby image boundaries and is also normalized to $[0, 1]$, similarly to $V(c)$. d_x is the distance from pixel x to the image center. As shown in Fig. 6c, 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 Fig. 7. Obviously, the smaller a color variance is, the higher the probability the color belongs to the salient object is.

3.2 Dynamic Salient Feature

3.2.1 Motion Salient Features

The motion field and the features derived from it are useful to induce visual attention. For example, large motion and motion contract are supposed to induce visual attention in [27], [24], and a constant velocity motion model is assumed for the salient object in [30]. Motion magnitude is a possible cue, but may not be sufficient. For example, in Fig. 8a, the region with larger motion magnitude includes the salient object. In contrast, the region with smaller motion magnitude includes the salient object in Fig. 8b. In this paper, we view the motion field as an image and define the local, regional, and global salient features from it.

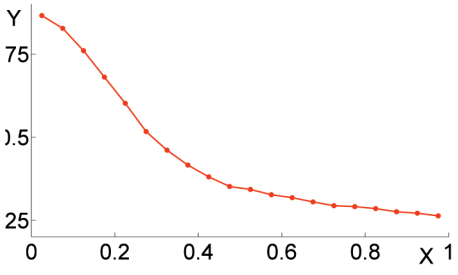


Fig. 7. Color spatial variance (x-coordinate) versus average saliency probability (y-coordinate) on the image set $\mathcal{A}$. The saliency probability is computed from the "ground truth" labeling.

Fig. 8. Motion map. (a) The salient object with a large motion. (b) The background with a large motion.

We compute the motion field M using the SIFT flow [35]. It can be observed that the motion fields have some special properties for the salient feature computation. For example, the motion fields from the salient object tend to be consistent because the regions from the salient object are inclined to have a similar motion, and the motion fields in the regions of object boundaries are usually disordered. To measure this consistency, motion variance $V(x, M)$ in a small rectangle surrounding x is computed, and a weight is assigned to each pixel as follows:

$$W(x, M) = \exp(-\epsilon_c \|V(x, M)\|^2), \tag{19}$$

where $V(x, M)$ is computed on a 2D motion vector from a window (5×5 in this paper) centered at x and $\epsilon_c = 0.2$. As in the first row of Fig. 9, the motion from the surrounding region of pixel x is more cluttered and the weight of pixel x is smaller.

Compared with the salient features defined for a single image, all the local, regional, and global salient features are defined similarly on weighted 2D motion vectors, including the motion magnitude and the motion direction. In the following, we present the formulation and only highlight the difference from the image:

Multiscale contrast of weighted motion field. It is defined on weighted motion vectors as follows:

$$f_{M_c}(x, M) = \sum_{l=1}^{L} \sum_{x' \in N(x)} W_x^l W_{x'}^l \|M^l(x) - M^l(x')\|^2, \tag{20}$$

where M^l is the lth-level motion in the pyramid and W_x^l is the weight at pixel x. We also test the multiscale contrast on motion magnitude or motion direction. They do not outperform the feature on the 2D motion vector because neighborhood pixels may have the same motion magnitude

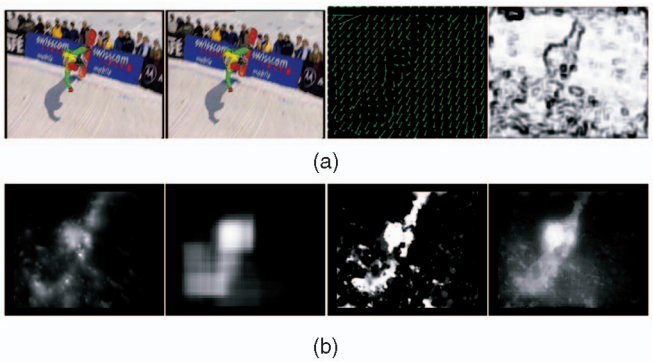


Fig. 9. Motion salient features. From left to right: (a) Two adjacent images, the motion field, and the motion weight map; (b) the local, regional, global, and combined motion salient features.

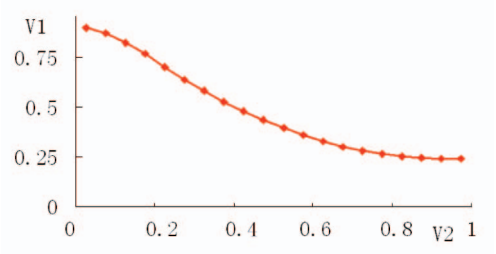


Fig. 10. Histogram distance of appearance features (x-coordinate) versus average saliency ratio (y-coordinate).

but different directions, and the salient feature from orientation does not perform well, especially when the motion magnitude is small.

Center-surround histogram of weighted motion field. It captures the statistic difference of motion field in a regional extension. We compute the histogram of motion vectors where horizontal and vertical motion are both normalized and used. The regional salient feature is defined as

$$f_{Mh}(x, M) \propto \sum_{\{x'|x \in R^*_M(x')\}} w_{xx'} W_{x'} \chi^2\big(R^*_M(x'), R_{M^*_S}(x')\big), \quad (21)$$

where R^*_M has the largest center-surround histogram distance on motion vectors, $w_{xx'}$ is the weight for the spatial distance, and $W_{x'}$ is the weight of pixel x'.

Spatial distribution of weighted motion field. It captures the global distribution of the motion field in an image. There are usually several different prominent motions in one frame, such as the motions from the background, object, or disturbs. Similarly to the spatial distribution of color, the wider a motion is distributed in the image, the less possible it is that a salient object corresponds to this motion. To get the spatial distribution, these motion vectors in which each vector is weighted by W_x are first clustered into several GMMs. The spatial variance $V_M(m)$ of each Gaussian component m is computed similar to the static feature, and the final spatial distribution feature is defined as

$$f_{Ms}(x, M) \propto \sum_m W_x p(m|M_x) \cdot (1 - V_M(m)). \quad (22)$$

Usually, there are fewer components for motion than for color because there are not many independent moving regions in one image. We use 3-5 motion components in this paper.

3.2.2 Appearance Coherent Feature

It is observed that salient objects from two consecutive frames probably have similar appearance features. The observation is verified on our labeled image pairs. We first compute the color histogram $R_t(x)$ in the labeled rectangle on the current image I_t, and the bth bin of $R_t(x)$ is computed as: $R_t(x)^b = \sum_{x' \in R_t} f(x', I_t)\delta(I_{x'} = b)$, where $f(x', I_t)$ is set to 1 if x' is in the labeled rectangle and 0 otherwise. Second, we randomly select one rectangle with the same size in the previous image I_{t-1} and compute the color histogram of $R_{t-1}(x')$ similarly. Third, we compute the saliency ratio V_1 and the χ^2 distance between the color histograms $R_t(x)$ and $R_{t-1}(x')$ as two variables:

Fig. 11. Appearance coherent feature. (a) and (c) are an image pair, (b) and (d) are the corresponding static salient features, and (e) is the appearance coherent feature.

$$V_1 = \frac{\sum_{x'} f(x', I)\delta(x' \in R_t)}{\sum_{x'} \delta(x' \in R_t)} \cdot \frac{\sum_{x'} f(x', I_{t-1})\delta(x' \in R_{t-1})}{\sum_{x'} \delta(x' \in R_{t-1})},$$

$$V_2 = \chi^2(R_t(x), R_{t-1}(x')), \quad (23)$$

where $f(x, I)$ and $f(x', I_{t-1})$ come from the labeled ground truths. We then create a statistic of the relationship between the two variables V_1 and V_2, as shown in Fig. 10.

To integrate the appearance coherence into the energy defined on a 2D graph, we try to penalize the pixels that are identified to be in the salient object by static salient features but with a big color histogram difference. First, we compute the weighted color histogram $R_t(x)$ from an $N \times N$ patch surrounding pixel x, and the bth bin of the color histogram $R_t(x)$ is computed as: $R_t(x)^b = \sum_{x' \in R_t} f(x', I)\delta(I_{x'} = b)$, where $f(x', I)$ is the static salient feature defined on a single image. Second, we search the patch $R_{t-1}(x^*)$ in image I_{t-1} to satisfy: $x^* = arg\ max_{x'}\ \chi^2(R_t(x), R_{t-1}(x'))$, where $x' \in N(x)$ and $N(x)$ are the set of the neighboring pixels of x, and $R_{t-1}(x')$ is computed similar to $R_t(x)$. Finally, the appearance coherent feature is computed as

$$f(x, I_t, I_{t-1}) \propto \frac{f(x, I_t) + f(x^*, I_{t-1})}{2} \exp(-\chi^2(R_t(x), R_{t-1}(x^*))), \quad (24)$$

where $f(x, I_t)$ and $f(x^*, I_{t-1})$ are the static salient features from I_t and I_{t-1}. Fig. 11 gives an example of the appearance coherent feature.

4 Evaluation

4.1 Data Set

4.1.1 Image Data Set

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. To test the performance, we further selected 20,840 images that contain a clear, unambiguous object of interest, which is helpful for building the ground truth. 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.

Fig. 2 gives some example images, and each image contains an unambiguous salient object. These salient objects differ in category, color, shape, size, etc. In other words, there is no more prior knowledge or constraint on these objects except that they are the most salient. This image database is different from the UIUC Cars data set, or the PASCAL VOC 2006 data set, where images containing a

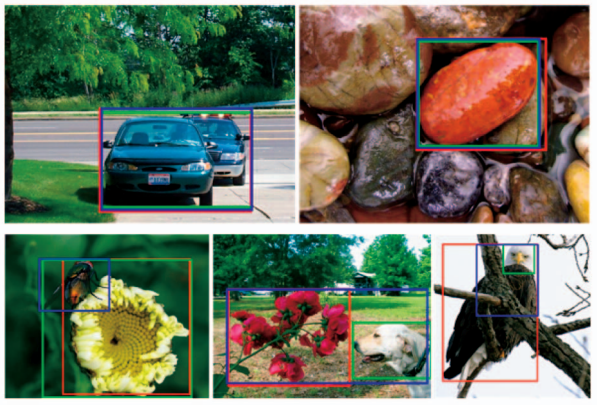


Fig. 12. Labeled images from three users. Top: two consistent labeling examples. Bottom: three inconsistent labeling examples.

specific category of objects are collected together. As clarified in the above section, we do not judge whether an object exists or discriminates from multiple objects. Specifically, we aim to locate the salient object, with the assumption that one salient object exists in the given image.

4.1.2 Sequential Image Data Set

We collected a video database with $2{,}000+$ video segments from a variety of sources, e.g., video sharing Web sites. Further, we selected 100 video segments that include salient object sequences, such as racing car, long jump, kids sequences, and so on. Example images from these video segments are shown in Fig. 14. Each video segment contains about 100-500 frames with the same salient objects. We also label the ground truth by hand for parameter learning and result evaluation, and $30{,}000+$ image pairs are collected for labeling. One trait of these image pairs is that the image quality is not as good as the image quality from the above image data set because all images are taken from video segments on Web sites. Another trait is that the salient objects are much smaller and the average of the ratios between the sizes of salient objects and images is about 0.1, which results in the salient object detection tasks in those video segments being very challenging.

4.2 Ground Truth Construction

For labeling the ground truth, we ask the user to draw a bounding rectangle to specify a salient object. Our detection algorithm also outputs a rectangle around the salient object. As addressed in [43], one advantage is that it is much easier to provide ground truth annotation for bounding boxes than, e.g., for pixelwise segmentations. At the same time, the rectangle representation of the salient object satisfies many applications, such as adaptive image display on small devices and image collage. We still represent the salient object piecewise as A_t in the problem formulation, and we will transform the final binary result to a bounding rectangle for further evaluation and applications where this strategy can avoid cutting off the spindly edge of the salient object.

4.2.1 Labeling Consistency in Image Data Set

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. In this paper, we focus on the case of a single salient object in an image. For each image to be labeled, we ask the user to draw a rectangle that 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.

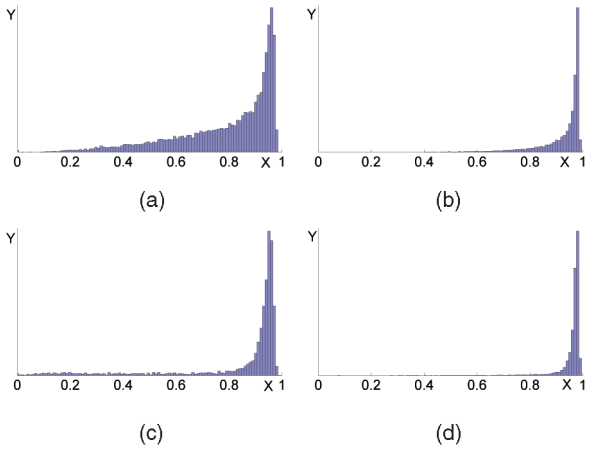


Fig. 13. Labeling consistency for image data set. (a) and (b) $C_{0.9}$ (agreed upon by all three users) and $C_{0.5}$ on image set $\mathcal{A}$. (c) and (d) $C_{0.9}$ (agreed upon by at least eight of nine users) and $C_{0.5}$ on image set $\mathcal{B}$.

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:

$$g_x = \frac{1}{M}\sum_{m=1}^{M} a_x^m, \tag{25}$$

where M is the number of users and $A^m = \{a_x^m\}$ is the binary mask labeled by the mth user. Fig. 12 shows two highly consistent examples and three inconsistent examples. The inconsistent labeling is due to multiple disjointed foreground objects for the first two examples at the bottom row. The last example in 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:

$$C_t = \frac{\sum_{x\in\{g_x>t\}} g_x}{\sum_x g_x}, \tag{26}$$

where C_t is the percentage of pixels whose saliency probabilities are above a given threshold t. For example, $C_{0.5}$ is the percentage 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. Figs. 13a and 13b show the histograms of $C_{0.9}$ and $C_{0.5}$ on the image set $\mathcal{A}$. As can be seen, the labeled results are quite consistent, e.g., 92 percent of the labeling results are consistent between at least two

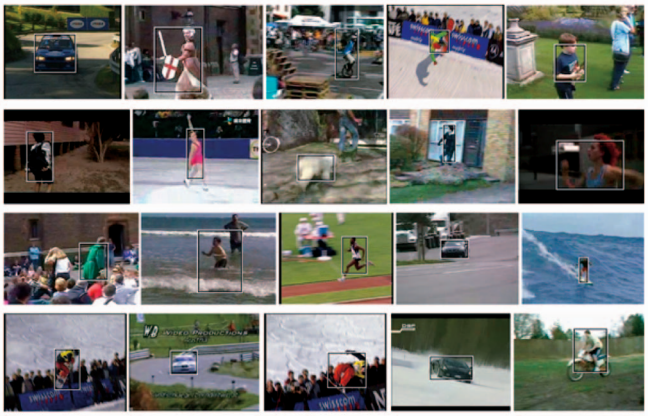


Fig. 14. Sample images from experimental video segments and our detection results on these images.

users (Fig. 13b) and 63 percent of the labeling results are highly consistent among all three users (Fig. 13a).

In the second stage, we randomly selected 5,000 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. Figs. 13c and 13d 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. The whole labeled image database is publicly available.[1]

4.2.2 Labeling Continuity in Sequential Image Data Set

For image pairs from video segments, people may have less disputation about what the salient object is because the motion helps to address the salient object. We ask only one user to label these sequential images, and it takes about two weeks to label all image pairs. In most cases, the movement of the salient object is smooth, which means that the labeled rectangles from two adjacent frames are also continuous. To describe the labeling continuity, we compute the statistic $C = \frac{Region\ area(R_{t-1} \cap R_t)}{Region\ area(R_{t-1} \cup R_t)}$, where R_{t-1} and R_t are two labeled rectangles for two adjacent frames. Fig. 15a shows the histogram about C on these image pairs. We also get statistics on the maximal boundary distance to describe the labeling continuity for sequential images, and the histogram is shown in Fig. 15b. We find that the maximal boundary distance is less than 10 pixels for 95 percent of image pairs, and this is also used as a reference when we define the appearance coherent features.

4.3 Evaluation Criteria

With the labeled 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

1. http://research.microsoft.com/jiansun/SalientObject/salient_object.htm.

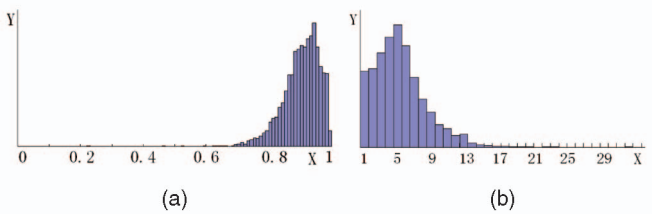


Fig. 15. Labeling continuity for the video data set. (a) The histogram of C. (b) The histogram of maximal boundary distance. Both are on two adjacent labeled rectangles R_{t-1} and R_t.

of a correctly detected salient region to the detected/"ground truth" salient region:

$$Precision = \sum_x g_x a_x \Big/ \sum_x a_x, Recall = \sum_x g_x a_x \Big/ \sum_x g_x. \tag{27}$$

The F-measure is the weighted harmonic mean of precision and recall, with a nonnegative α:

$$F_\alpha = \frac{(1+\alpha) \times Precision \times Recall}{\alpha \times Precision + Recall}. \tag{28}$$

We set $\alpha = 0.5$ following [44]. The F-measure is an overall performance measurement.

For the boundary-based measurement, we use boundary displacement error (BDE) [45], which measures the average displacement error of the corresponding boundaries of two rectangles. The displacement is averaged over the different users.

4.4 Implementation of CRF Learning and Inference

For the image data set, 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, which are excluded from the testing phase. For sequential image data set, we randomly select 20 video segments with $5{,}000+$ image pairs to construct a training set, and use others for testing. We do many different splits in terms of a training/test data set, and find that the different splits almost do not affect the evaluation results. The key factor is the amount of training data, and the parameter learning algorithm can converge well when the amount of training data is more than 2,000.

Because the ground truth of salient objects is labeled by rectangles, this strategy lacks the precise alignment between object boundaries and labeled rectangles. Instead of learning the parameter of the pairwise feature [46], we normalize the sum of λ_k by experience, e.g., $\sum_k \lambda_k = 1$. Furthermore, we observe that the pixels from the boundaries of labeled rectangles are less believable because the surrounding rectangle may label some pixels near the boundaries as the salient object by mistake. To reduce this effect, we use a Gaussian function to give the weight of pixels when we compute $\Delta\lambda_k$ in (9). This strategy helps to speed up the convergence of the learning algorithm.

We use the tree-reweighted belief propagation to infer the labeling because it is used for CRF learning. We find that there are small differences between the learned parameters if we use different algorithms to compute the marginal distribution $p(a_x^n | I^n; \vec{\lambda})$. To output a rectangle for the evaluation, we exhaustively search for a smallest

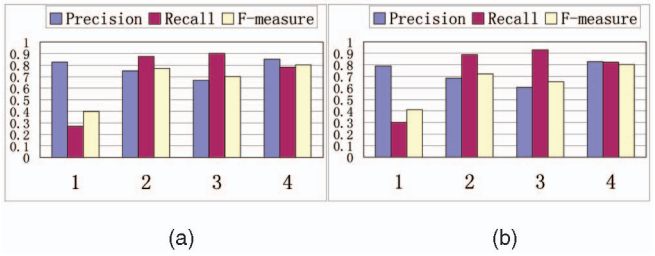


Fig. 16. Evaluation of salient object features. 1. Multiscale contrast. 2. Center-surround histogram. 3. Color spatial distribution. 4. Combination of all features. (a) Image set $\mathcal{A}$. (b) Image set $\mathcal{B}$.

rectangle containing at least 95 percent salient pixels in the binary label map produced by the CRF model.

4.5 Salient Object Detection from a Single Image

4.5.1 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. Fig. 16 shows the precision, recall, and F-measure of these CRFs on the image sets $\mathcal{A}$ and $\mathcal{B}$. As can be seen, the multiscale contrast feature has high precision but 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 Fig. 16. The best linear weights we learned are: $\overrightarrow{\lambda} = \{0.24, 0.54, 0.22\}$.

Fig. 17 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.

Fig. 17. Examples of salient features. From left to right: input image, multiscale contrast, center-surround histogram, color spatial distribution, and binary salient mask by CRF.

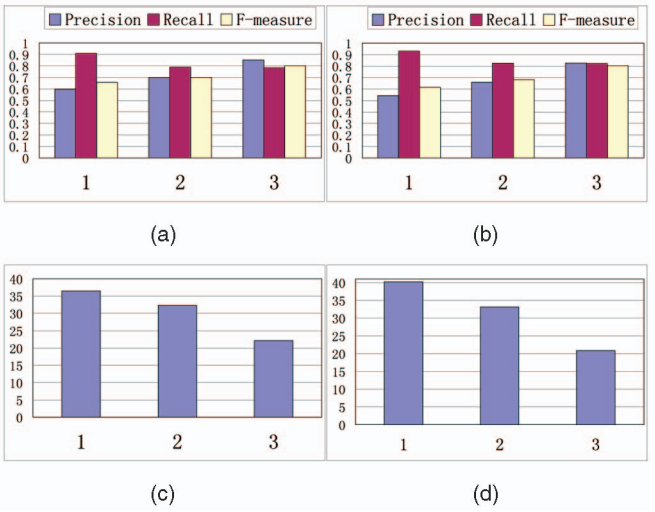


Fig. 18. Comparison of different algorithms. Region-based (precision, recall, and F-measure) and boundary-based (BDE) evaluations. 1. FG. 2. SM. 3. Our approach. (a) Precision/recall, image set $\mathcal{A}$. (b) Precision/recall, image set $\mathcal{B}$. (c) BDE, image set $\mathcal{A}$. (d) BDE, image set $\mathcal{B}$.

4.5.2 Comparison with Other Approaches

We compare our algorithm with two leading approaches. One is the contrast and fuzzy growing-based method [39], which we call "FG." This approach directly outputs a rectangle. Another approach is based on the salient model presented in [13] (we use a matlab implementation from http://www.saliencytoolbox.net), and we call it "SM." Because the output of Itti's salient model is a saliency map, we convert the saliency map to a rectangle containing 95 percent of the fixation points, which are determined by the winner-take-all algorithm [13]. We also resolve the rectangles directly through complete searching by maximizing $\sum_{x \in R}(1 - F(x)) + \sum_{x \notin R} F(x)$, where R is the resolve rectangle and $F(x) \in [0, 1]$ is the normalized saliency map. This method can be applied on our saliency map and Itti's saliency map, but the results do not outperform the corresponding results using the current method.

Fig. 18 shows the evaluation results of three algorithms on both image sets $\mathcal{A}$ and $\mathcal{B}$. On image set $\mathcal{A}$, our approach reduces by 42 and 34 percent the overall error rates on F-measure, and 39 and 31 percent BDEs, compared with FG and SM. Similarly, 49 and 38 percent overall error rates on F-measure and 48 and 37 percent BDEs are reduced on the image set $\mathcal{B}$.

Note that as shown in Figs. 16 and 18, 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 a very useful measure in attention detection. For example, a 100 percent recall rate can be achieved by simply selecting the whole image. So, an algorithm trying to achieve a high recall rate tends to select as large an salient region as possible, sacrificing the precision rate. The key objective of salient object detection should be to locate the position of a salient object as accurately as possible, i.e., with high precision. However, for images with a large salient object, high precision is also not too difficult to achieve. Again, for example, for an image with a salient object occupying 80 percent of the image area, just selecting the whole image as the salient area will give 80 percent

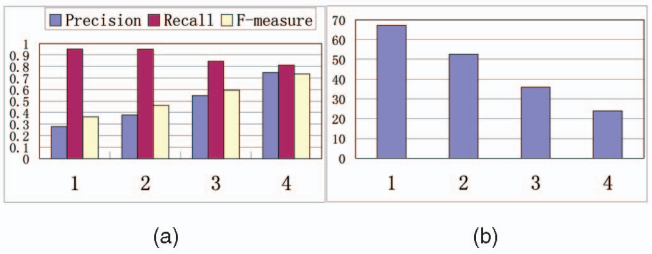


Fig. 19. Comparison on a small object (object/image ratio $\in [0, 0.25]$) data set from image set $\mathcal{A}$. 1. A rectangle centered at the image center and with 0.6 object/image ratio. 2. FG. 3. SM. 4. Our approach. (a) Precision/recall. (b) BDE.

precision with 100 percent recall rate. So, the real challenge for salient object detection is to achieve high precision on small salient objects. To construct such a challenging 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 data set are shown in Fig. 19, where we also show the performance of a rectangle fixed at the image center with 0.6 object/image ratio. Note 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 percent improvement) and BDE (55 and 33 percent reduction). Fig. 20 shows several examples with ground truth rectangles from one user for a qualitative comparison. We can see that the FG and SM approaches tend to produce a larger attention rectangle and our approach is much more precise.

Fig. 20. Comparison of different algorithms. From left to right: FG, SM, our approach, and ground truth.

Fig. 21. Our detection result on the images in Fig. 2.

Fig. 21 shows our detection results on the images in Fig. 2. Our results are also publicly available with the whole labeled database.

4.6 Salient Object Detection from Sequential Images

4.6.1 *Effectiveness of Salient Features*

To evaluate the effectiveness of the static and dynamic salient features for sequential images, we set up a "baseline method" model in the CRF framework with the following energy:

$$E(A_t|I_{t-1}, I_t) = \sum_x \sum_{k=1}^{K} \lambda_k F_k(a_x, \cdot) + \sum_{x,x'} S(a_x, a_{x'}, I_t), \quad (29)$$

where $F_k(a_x, \cdot)$ indicates different salient features and $S(a_x, a_{x'}, I_t)$ is the pairwise feature. We define $F_k(a_x, \cdot)$ from (29) with different features, and train four CRFs as follows:

- **C1:** The static salient features from the current image I_t are used: $F_k(a_x, \cdot) = F_k(a_x, I_t)$.
- **C2:** The motion salient features from the motion field M_t are used: $F_k(a_x, \cdot) = F_k(a_x, M_t)$.
- **C3:** The static and motion salient features are both used to detect a salient object, and $F_k(a_x, \cdot)$ is the combination of $F_k(a_x, I_t)$ and $F_l(a_x, M_t)$.
- **C4:** All of the static and dynamic salient features are used, and $F_k(a_x, \cdot)$ is the combination of $F_k(a_x, I_t)$,

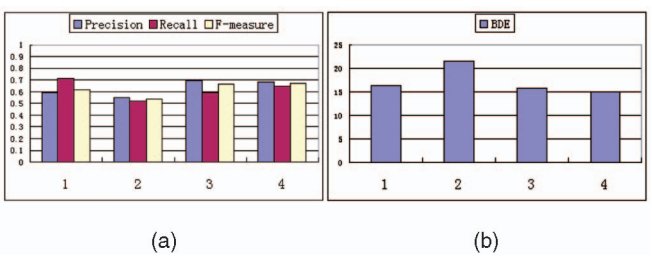


Fig. 22. Evaluation of different salient features for sequential images. (a) and (b) 1-4 corresponds to C1-C4, respectively, where different salient features are trained within the CRF framework. (a) Precision/recall. (b) BDE.

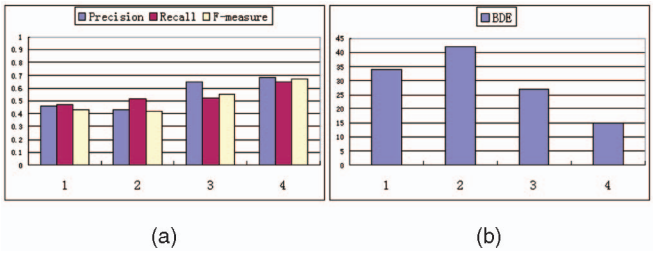


Fig. 23. Comparison with different approaches for sequential images. 1-3 correspond to D1-3, respectively, 4 is our approach C4. (a) precion/recall. (b) BDE.

$F_l(a_x, M_t)$, and $F_m(a_x, I_{t-1}, I_t)$. This is the proposed approach for salient object detection.

Fig. 22 shows the precision, recall, F-measure, and BDE of these CRFs on $30{,}000+$ image pairs. The salient objects in these image pairs are very small compared with the image database, and that is why the performance of C1 is not as good as the trained CRF in our image database. Approach C3, combining salient features on color and motion, improves 8-9 percent on F-measure compared with approach C1 with only the salient features on color. C3 also improves 23 percent on F-measure compared with C2 with only the salient features on motion. We can see that C1 outperforms C2 on the F-measure. With the appearance coherent features, the proposed approach C4 improves 9 percent on recall with a little sacrifice on precision and improves 1 percent on the overall criteria F-measure.

4.6.2 Comparison with Other Approaches

A general CRF model is defined in (29), where different static and dynamic features can be included to learn a detector. We compare our approach with the following approaches:

- **D1:** We use Itti's salient model to compute the static saliency map following [47], [24], [27], and the multiple-scale motion contrast from [24] to compute the dynamic saliency map. We also test the motion saliency from [47], where the difference between the pixel's motion and the global motion from the whole image is computed as the motion saliency, and experiments indicate that it does not outperform the multiple-scale motion contrast.
- **D2:** Differently from D1, we use the saliency map from the temporal surprise in [11] as the dynamic salient feature. However, the surprise computation using all features is extremely costly on a large number of image sequences. A reasonable simplification is to use only four combined saliency maps to compute the temporal surprise. We use the publicly available Bayesian Surprise Matlab toolkit (http://sourceforge.net/projects/surprise-mltk) for the implementation.
- **D3:** Other related work includes the tracking algorithm with a hand initialization in the first frame, and we report the results of the typical mean-shift tracking algorithm [48].

Fig. 23 shows the comparison of our approach with D1-3. Our approach improves 52 percent on F-measure and reduces 54 percent on BDEs compared with D1. We also evaluate the results with only the motion saliency in (29) and find that the multiscale motion contrast has a very low recall. This is the main reason that motion saliency is not well leveraged in D1. Our approach also improves 59 percent on F-measure and reduces 64 percent on BDEs compared with D2. The difference between D1 and D2 is the motion saliency, and we find that video surprise with the goal of eye movements cannot help to locate the small salient object well, and further, the video surprise does not strengthen the static saliency much because it is computed based on the static saliency. We find that the collected image sequences are also very challenging for the mean-shift tracking algorithm, because of the following traits: large motion of object and camera, object rotation and appearance change, illumination change, and so on. Our approach improves 20 percent on F-measure and reduces 45 percent on BDEs compared with D3. The results imply that the salient features can help visual object tracking for those challenging videos.

5 Discussion

In this section, we discuss the connection and clarify the difference between our approach for salient object detection and other related work.

5.1 Salient Object versus Visual Saliency

A visual saliency map is computed from multiscale image features in Itti's model [13], [12], which is one of the most representative works on computational modeling of visual attention. Itti's model and those similar to it are based on the biologically plausible computational models of attention, with a particular emphasis on bottom-up control of attentional deployment. They state as a goal the determination of fixation and eye movements over an image. We summarize the difference between salient object detection and visual saliency computation with Itti's model in the following.

First, a salient object is essentially one important aspect of visual attention, and the goal is to locate the salient object in an image or sequential images to help in displaying images on a small device or browsing image collection. It is different from Itti's model, which has as a goal the determination of fixation and eye movements over an image. The recent study [42] also analyzes their connection and indicates that Itti's visual saliency model is closely related to interesting object detection. Second, we propose a different solution to the problem. We adopt a binary mask to indicate a salient object using the salient feature maps which are combined with learned parameters. These feature maps are different from the visual saliency maps or the conspicuous maps in Itti's model that are based on biological theory.

5.2 Salient Object Detection versus Figure-Ground Segregation

The figure-ground segregation task is similar to salient object detection as both aim to find the objects, but they are essentially different. The main difference is that our approach detects a salient object automatically, without any prior knowledge about its category, its shape, or size and that the conventional figure-ground segregation algorithms require the supposition of the category of objects [19], [20], [21] or user interactions [22], [23]. On the other hand, the visual features adopted for the detection differ greatly. For salient object detection, we propose generic salient features without discrimination of object categories. For figure-ground segregation of an object with a given category, the specific features, for example, for cows, may be defined

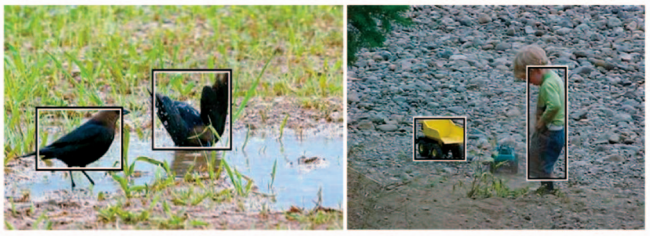


Fig. 24. Multiple salient object detection. (a) Two birds are detected at the same time. (b) The toy car is detected first, and using the updated feature maps, the boy is detected second.

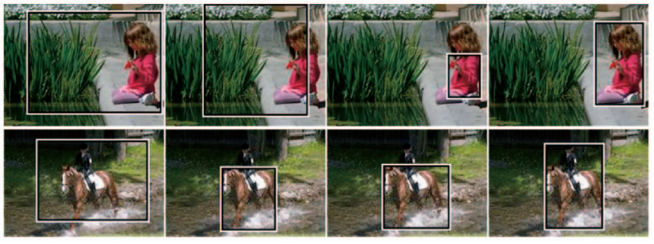


Fig. 25. Failure cases. From left to right: FG, SM, our approach, and ground truth.

specifically and these features cannot be adopted for other categories. Due to the above differences, the figure-ground segregation algorithm is not comparable to our approach.

6 Conclusion

In this paper, we have presented a supervised approach for salient object detection which is formulated as a binary labeling problem using a set of local, regional, and global salient object features. A CRF model was learned and evaluated on a large image database containing 20,000+ well-labeled images by multiple users. We also extend this supervised approach to detect a salient object sequence from sequential images, where dynamic salient features are included to help detect the salient object.

There are several possible remaining issues for further investigation. We plan to experiment with nonrectangular shapes for salient objects and a nonlinear combination of features. In particular, we are extending our single salient object detection framework to detect any number of salient objects, including no salient object at all. Fig. 24 shows two initial results. In Fig. 24a, our current CRF approach can directly output two disjointed connected components so that we can easily detect them simultaneously. In Fig. 24b, we use the inhibition-of-return strategy [13] to detect the salient objects one-by-one. Finally, Fig. 25 shows two failure cases which demonstrate one of the challenges in the salient object detection—hierarchical salient object detection.

Acknowledgments

Tie Liu and Zejian Yuan were supported by a grant from the National Natural Science Foundation of China (No. 90820017). Zejian Yuan and Nanning Zheng were supported by grants from the National Basic Research Program of China (No. 2007CB311005) and the National High-Tech Research and Development Plan of China (No. 2006AA01Z192). The authors appreciate the helpful comments from reviewers.

References

[1] A. Santella, M. Agrawala, D. Decarlo, D. Salesin, and M. Cohen, "Gaze-Based Interaction for Semi-Automatic Photo Cropping," *Proc. Conf. Human Factors in Computing Systems*, pp. 771-780, 2006.

[2] 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, 2002.

[3] L. Itti, "Models of Bottom-Up and Top-Down Visual Attention," PhD dissertation, California Inst. of Technology Pasadena, 2000.

[4] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "Autocollage," *Proc. ACM SIGGRAPH*, pp. 847-852, 2006.

[5] V. Navalpakkam and L. Itti, "An Integrated Model of Top-Down and Bottom-Up Attention for Optimizing Detection Speed," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, pp. 2049-2056, 2006.

[6] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is Bottom-Up Attention Useful for Object Recognition?" *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, pp. 37-44, 2004.

[7] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional Selection for Object Recognition—A Gentle Way," *Proc. Second Int'l Workshop Biologically Motivated Computer Vision*, 2002.

[8] N. Bruce and J. Tsotsos, "Saliency Based on Information Maximization," *Advances in Neural Information Processing Systems*, pp. 155-162, MIT Press, 2005.

[9] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *Advances in Neural Information Processing Systems*, pp. 545-552, MIT Press, 2006.

[10] L. Itti and P. Baldi, "Bayesian Surprise Attracts Human Attention," *Advances in Neural Information Processing Systems*, pp. 547-554, MIT Press, 2005.

[11] L. Itti and P. Baldi, "A Principled Approach to Detecting Surprising Events in Video," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, pp. 631-637, 2005.

[12] L. Itti and C. Koch, "Computational Modelling of Visual Attention," *Neuroscience*, vol. 2, no. 3, pp. 194-203, 2001.

[13] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[14] C. Koch and S. Ullman, "Shifts in Selection in Visual Attention: Toward the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219-227, 1985.

[15] O.L. Meur, O.L. Callet, D. Barba, and D. Thoreau, "A Coherent Computational Approach to Model Bottom-Up Visual Attention," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802-817, May 2006.

[16] 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*, vol. 78, nos. 1/2, pp. 507-545, 1995.

[17] A. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97-136, 1980.

[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to Predict Where Humans Look," *Proc. IEEE Int'l Conf. Computer Vision*, 2009.

[19] A. Levin and Y. Weiss, "Learning to Combine Bottom-Up and Top-Down Segmentation," *Proc. European Conf. Computer Vision*, pp. 581-594, 2006.

[20] E. Borenstein, E. Sharon, and S. Ullman, "Combining Top-Down and Bottom-Up Segmentation," *Proc. Computer Vision and Pattern Recognition Workshop*, 2004.

[21] B. Leibe, K. Mikolajczyk, and B. Schiele, "Segmentation Based Multi-Cue Integration for Object Detection," *Proc. British Machine Vision Conf.*, 2006.

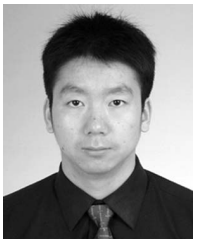
[22] Y. Boykov and M.P. Jolly, "Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 105-112, 2001.

[23] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts," *Proc. ACM SIGGRAPH*, pp. 309-314, 2004.

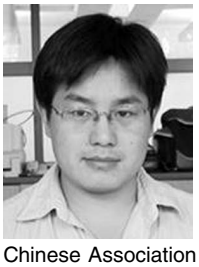
[24] R. Carmi and L. Itti, "Visual Causes versus Correlates of Attentional Selection in Dynamic Scenes," *Vision Research*, vol. 46, no. 26, pp. 4333-4345, 2006.

[25] Y. Ma and H. Zhang, "A Model of Motion Attention for Video Skimming," *Proc. Int'l Conf. Image Processing*, pp. 129-132, 2002.

[26] Y. Zhai and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues," *Proc. ACM Multimedia,* pp. 815-824, 2006.
[27] A. Bur, P. Wurtz, R.M. Miiri, and H. Hugli, "Dynamic Visual Attention: Competitive versus Motion Priority Scheme," *Proc. Int'l Conf. Computer Vision Systems,* 2007.
[28] S. Drouin, P. Hbert, and M. Parizeau, "Incremental Discovery of Object Parts in Video Sequences," *Proc. IEEE Int'l Conf. Computer Vision,* vol. 2, pp. 1754-1761, 2005.
[29] N. Jojic, J. Winn, and L. Zitnick, "Escaping Local Minima through Hierarchical Model Selection: Automatic Object Discovery, Segmentation, and Tracking in Video," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 117-124, 2006.
[30] D. Liu and T. Chen, "A Topic-Motion Model for Unsupervised Video Object Discovery," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* 2007.
[31] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proc. Int'l Conf. Machine Learning,* pp. 282-289, 2001.
[32] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive Image Segmentation Using an Adaptive GMMRF Model," *Proc. European Conf. Computer Vision,* pp. 428-441, 2004.
[33] Y. Li, J. Sun, and H.-Y. Shum, "Video Object Cut and Paste," *Proc. ACM SIGGRAPH,* pp. 595-600, 2007.
[34] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video Snapcut: Robust Video Object Cutout Using Localized Classifiers," *Proc. ACM SIGGRAPH,* 2009.
[35] C. Liu, J. Yuen, A.B. Torralba, J. Sivic, and W.T. Freeman, "Sift flow: Dense Correspondence across Different Scenes," *Proc. European Conf. Computer Vision,* no. 3, pp. 28-42, 2008.
[36] X. Ren, C. Fowlkes, and J. Malik, "Cue Integration for Figure/Ground Labeling," *Advances in Neural Information Processing Systems,* pp. 1121-1128, MIT Press, 2005.
[37] V. Kolmogorov, "Convergent Tree-Reweighted Message Passing for Energy Minimization," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 10, pp. 1568-1583, Oct. 2006.
[38] F. Liu and M. Gleicher, "Region Enhanced Scale-Invariant Saliency Detection," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 1477-1480, 2006.
[39] Y.-F. Ma and H.-J. Zhang, "Contrast-Based Image Attention Analysis by Using Fuzzy Growing," *Proc. Int'l Conf. Multimedia,* pp. 374-381, 2003.
[40] F. Porkili, "Integral Histogram: A Fast Way to Extract Histograms in Cartesian Spaces," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 829-836, 2005.
[41] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 886-893, 2005.
[42] L. Elazary and L. Itti, "Interesting Objects Are Visually Salient," *J. Vision,* vol. 8, pp. 1-15, 2008.
[43] C.H. Lampert, M.B. Blaschko, and T. Hofmann, "Beyond Sliding Windows: Object Localization by Efficient Subwindow Search," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 1-8, 2008.
[44] D.R. Martin, C.C. Fowlkes, and J. Malik, "Learning to Detect Natural Image Boundaries Using Local Brightness, Color, and Texture Cues," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 5, pp. 530-549, May 2004.
[45] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi, "Yet Another Survey on Image Segmentation: Region and Boundary Information Integration," *Proc. European Conf. Computer Vision,* pp. 408-422, 2002.
[46] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation," *Proc. European Conf. Computer Vision,* pp. 1-15, 2006.
[47] F. Liu and M. Gleicher, "Video Retargeting: Automating Pan and Scan," *Proc. ACM Multimedia,* pp. 241-250, 2006.
[48] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 5, pp. 603-619, May 2002.



Tie Liu received the BS, MS, and PhD degrees from Xian Jiaotong University, in 2001, 2004, and 2007, respectively. He is currently a staff researcher at the Analytics and Optimization Department at IBM Research—China. His areas of interest include machine learning, pattern recognition, multimedia computing, and computer vision. He is also interested in data analysis and mining.



Zejian Yuan received the MS degree in electronic engineering from Xi'an University of Technology in 1999, and the PhD degree in pattern recognition and intelligent system from Xi'an Jiaotong University, China, in 2003. He was a visiting scholar in the Advanced Robotics Lab of Chinese University of Hong Kong during 2008-2009. He is currently an associate professor in the Department of Automatic Engineering, Xi'an Jiaotong University, and a member of the Chinese Association of Robotics. His research interests include image processing, pattern recognition, as well as machine learning methods in computer vision.

Jian Sun received the BS, MS, and PhD degrees from Xian Jiaotong University in 1997, 2000, and 2003, respectively. He then joined Microsoft Research Asia in July 2003. His current two major research interests are interactive compute vision (user interface + vision) and Internet compute vision (large image collection + vision). He is also interested in stereo-matching and computational photography.

Jingdong Wang received the BSc and MSc degrees in automation from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the PhD degree in computer science from the Hong Kong University of Science and Technology in 2007. He is currently an associate researcher in the Media Computing Group, Microsoft Research Asia. His areas of interest include machine learning, pattern recognition, multimedia computing, and computer vision. In particular, he has worked on kernel methods, semi-supervised learning, data clustering, image segmentation, and image and video presentation, management, and search.

Nanning Zheng received the graduate degree in electrical engineering and the MS degree in information and control engineering from Xi'an Jiaotong University, China, in 1975 and 1981, respectively, and the PhD degree in electrical engineering from Keio University, Yokohama, Japan, in 1985. In 1975, he joined Xi'an Jiaotong University, where he is currently a professor and the director of the Institute of Artificial Intelligence and Robotics. His research interests include computer vision, pattern recognition and image processing, and hardware implementation of intelligent systems. He became a member of the Chinese Academy of Engineering in 1999, and he is the Chinese representative on the Governing Board of the International Association for Pattern Recognition. He also serves as an executive deputy editor of the *Chinese Science Bulletin*. He is a fellow of the IEEE.

Xiaoou Tang received the BS degree from the University of Science and Technology of China, Hefei, in 1990, the MS degree from the University of Rochester, New York, in 1991, and the PhD degree from the Massachusetts Institute of Technology, Cambridge, in 1996. He is a professor in the Department of Information Engineering, Chinese University of Hong Kong. He worked as the group manager of the Visual Computing Group at Microsoft Research Asia from 2005 to 2008. He received the Best Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009. He is a program chair of the IEEE International Conference on Computer Vision (ICCV) 2009 and an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* and *International Journal of Computer Vision (IJCV)*. He is a fellow of the IEEE. His research interests include computer vision, pattern recognition, and video processing.

Heung-Yeung Shum received the PhD degree in robotics from the School of Computer Science, Carnegie Mellon University in 1996. He worked as a researcher for three years in the Vision Technology Group at Microsoft Research Redmond. In 1999, he moved to Microsoft Research Asia, where his tenure began as a research manager and he subsequently moved up to assistant managing director, managing director of Microsoft Research Asia, distinguished engineer, and corporate vice president. His research interests include computer vision, computer graphics, human-computer interaction, pattern recognition, statistical learning, and robotics. He was the general cochair of Ninth International Conference on Computer Vision (ICCV) 2003 and a program chair of the International Conference of Computer Vision (ICCV) 2007. He is a fellow of the IEEE and ACM.