Report Learning to Detect A Salient Object

Advisor

Dr. Maryam Abedi

Student

Mohammad Shahpouri

October

2021

Contents

List of Figures

List of Equations

1 Image Database

Salient object representation. They represented 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.
Image source. They selected 20,840 images for labeling.
Labeling consistency. for each labeled image, a saliency probability map was computed $G = \{g_x | g_x \in [0, 1]\}$ 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.
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.

2 CRF for Salient Object Detection

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

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

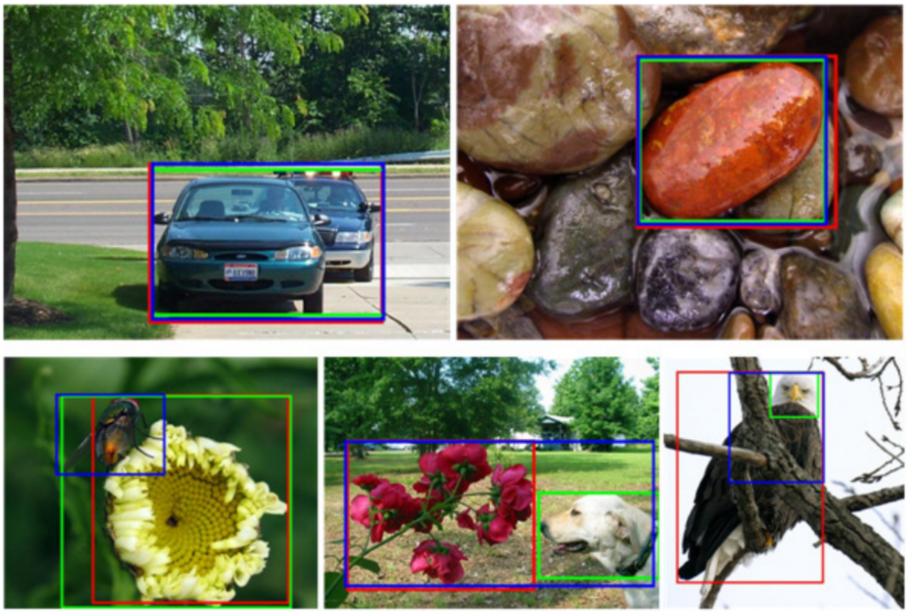


Figure 1: Labeled images from 3 users. Top: two consistent labeling examples. Bottom: three inconsistent labeling examples.

Where λ_k is the weight of the kth feature, and x, x' are two adjacent pixels. **Salient object feature.** $F_k(a_x, I)$ indicates whether or not a pixel x belongs to the salient object.

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

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]

$$S(a_x, a_{x'}, I) = |a_x - a_{x'}| \,.\, \exp\left(-\beta d_{x,x'}\right) \tag{5}$$

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 $\beta = (2\left\langle \|I_x - I_{x'}\|^2 \right\rangle)^{-1}$ [3], where $\langle . \rangle$ is the expectation operator. This feature function is a penalty term when adjacent pixels are assigned with different labels.

2.1 CRF Learning

To get an optimal linear combination of features λ should be estimated. The optimal parameters maximize the sum of the log-likelihood:

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

The derivative of the log-likelihood with respect to the parameter λ_k

$$\begin{aligned} &\frac{d \log P(A^n|I^n; \overrightarrow{\lambda})}{d\lambda_k} = \\ &< F_k(A^n, I^n) >_{P(A^n|I^n; \overrightarrow{\lambda})} - < F_k(A^n, I^n) >_{P(A^n|G^n)} \end{aligned} \tag{7}$$

Then, the gradient descent direction is:

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

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 x distribution and $p(a_x^n|g_x^n)$ is from the labeled ground-truth:

$$p(a_x^n|g_x^n) = \begin{cases} 1 - g_x^n & a_x = 0 \\ g_x^n & a_x = 1 \end{cases} \tag{9}$$

3 Salient Object Features

In this section, they introduced local, regional, and global features that define a salient object.

3.1 Multi-scale contrast

They simply defined the multi-scale contrast feature $f_c(x, I)$ as a linear combination of contrasts in the Gaussian image pyramid:

$$f_c(x, I) = \sum_{l=1}^{L} \sum_{x' \in N(x)} \left\| I^l(x) - I^l(x') \right\| \tag{10}$$

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 $[0, 1]$.

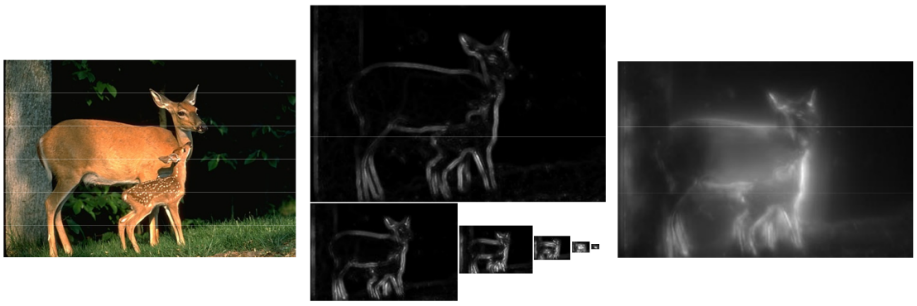


Figure 2: Multi-scale contrast. From left to right: input image, contrast maps at multiple scales, and the feature map from linearly combining the contrasts at multiple scales.

3.2 Center-surround histogram

They enclosed salient object by a rectangle R and constructed a surrounding contour R_S with the same area of R, as shown in Figure 3 (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, they use the χ^2 distance between histograms of RGB color: $\chi^2(R, R_S) = \frac{1}{2} \sum \frac{(R^i - R^i_S)^2}{R^i + R^i_S}$

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

$$R^*(x) = \arg\max_{R(x)} \chi^2(R(x), R_S(x)) \tag{11}$$

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 $f_h(x, I)$

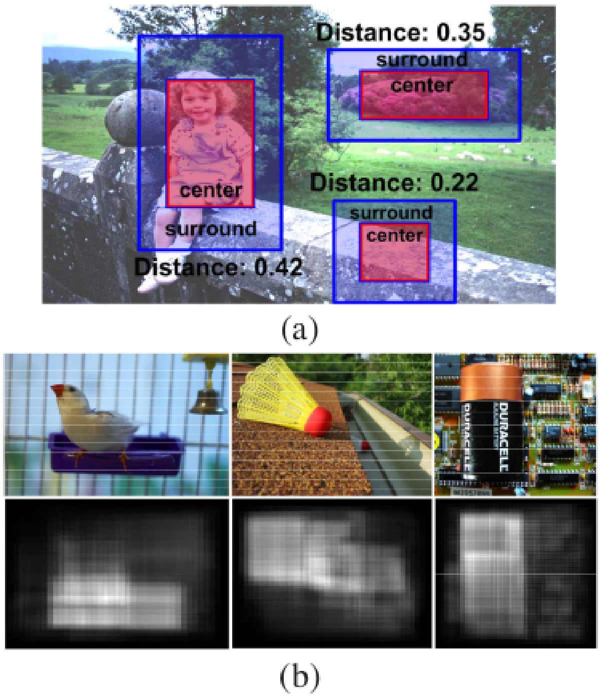


Figure 3: Center-surround histogram. (a) center-surround histogram distances with different locations and sizes. (b) top row are input images and bottom row are center-surround histogram feature maps.

is defined as a sum of spatially weighted distances:

$$f_h(x, I) \propto \sum_{\{x' | x \in R^*(x')\}} w_{xx'} \chi^2(R^*(x'), R_S^*(x')) \tag{12}$$

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 third of the size of $R^*(x')$.Finally, the feature map $f_h(\cdot, I)$ is also normalized to the range $[0, 1]$. Figure 3 (b) shows several center-surround feature maps.

3.3 Color spatial-distribution

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.

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:

$$p(c|I_x) = \frac{w_c \mathcal{N}(I_x|\mu_c, \Sigma_c)}{\sum_c w_c \mathcal{N}(I_x|\mu_c, \Sigma_c)} \tag{13}$$

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

$$V_h c = \frac{1}{|X|_c} \sum_x p(c|I_x) \cdot |x_h - M_h(c)|^2 \tag{14}$$

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

Where x_h is x-coordinate of the pixel x, and $|X|_c = \sum_c p(c|I_x)$. The vertical variance $V_v(c)$ is similarly defined. The spatial variance of a component c is $V(c) = V_h(c) + V_v(c)$. $\{V(c)\}_c$ is normalized 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:

$$f_s(x, I) \propto \sum_c p(c|I_x) \cdot (1 - V(c)) \tag{16}$$

The feature map $f_s(\cdot, I)$ is also normalized to the range $[0, 1]$. 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:

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

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.

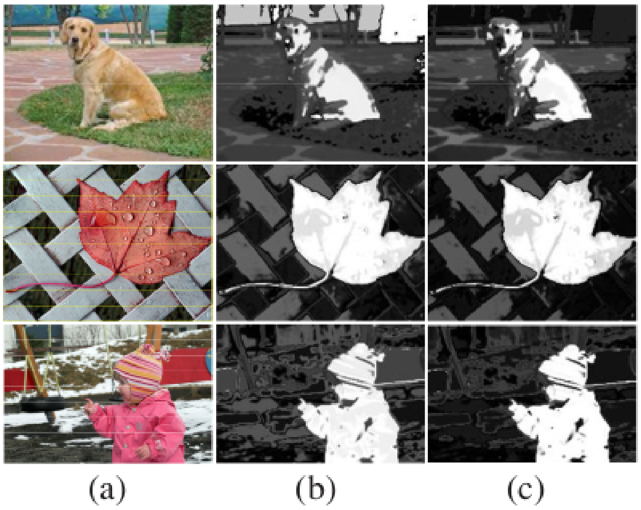
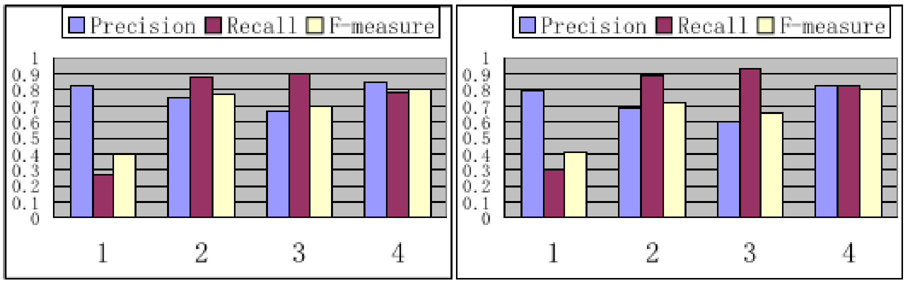


Figure 4: Color spatial-distribution feature. (a) input images. (b) color spatial variance feature maps. (c) center-weighted, color spatial variance feature maps.

4 Evaluation



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

Figure 5: Evaluation of salient object features. 1. multi-scale contrast. 2. center-surround histogram. 3. color spatial distribution. 4. combination of all features.

References

[1] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, (San Francisco, CA, USA), p. 282–289, Morgan Kaufmann Publishers Inc., 2001.

[2] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary amp; region segmentation of objects in n-d images," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1, pp. 105–112 vol.1, 2001.

[3] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive gmmrf model," in *Computer Vision - ECCV 2004* (T. Pajdla and J. Matas, eds.), (Berlin, Heidelberg), pp. 428–441, Springer Berlin Heidelberg, 2004.