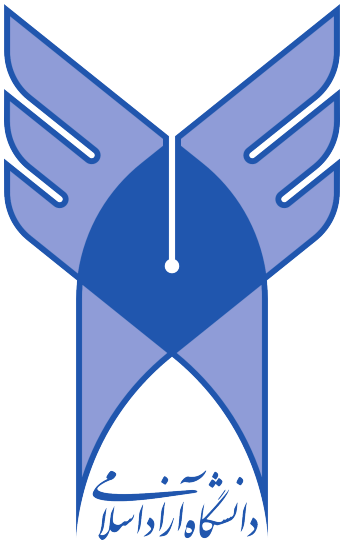

Saliency Aggregation: A Data-driven Approach report

Paper's Authors

Long Mai, Yuzhen Niu and Feng Liu

Advisor

Dr. Maryam Abedi

Student

Mohammad Shahpouri

October
2022

Contents

List of Figures

List of Equations

1 Saliency aggregation

Their method runs m saliency estimation algorithms, $\{M_i \| 1 \leq i \leq m\}$, on an input image I, and yields m saliency map, $\{S_i \| 1 \leq i \leq m\}$. $S_i(p)$ denotes saliency value at pixel p. The goal is to take these m saliency maps as input and produce a final saliency map S.

1.1 Standard saliency aggregation

the aggregated saliency value $S(p)$ at pixel p of image I is modeled as the probability:

$$S(p) = P(y_p = 1 | S_1(p), S_2(p), \ldots, S_m(p) \propto \frac{1}{Z} \sum_{i=1}^{m} \zeta(S_i(p)) \tag{1}$$

where $S_i(p)$ represents the saliency value of pixel p in the saliency map S_i , y_p is a binary random variable can be 1 if p is a salient pixel and 0 otherwise, and Z is a constant. According to [1] three different function for ζ in Equation 1 are implemented:

$$\zeta_1(x) = x, \ \zeta_2(x) = \exp(x), \ \text{and} \ \zeta_3(x) = \frac{-1}{\log(x)} \tag{2}$$

Standard saliency aggregation method are tested on two public saliency benchmarks, FT [2] and SS [3].

1.2 Data-driven saliency aggregation

1.2.1 Pixel-wise aggregation

final saliency value $S(p)$ is computed using the logistic model [4]:

$$P(y_p = 1 | x(p); \lambda) = \sigma(\sum_{i=1}^{m} \lambda_i S_i(p) + \lambda_{m+1}) \tag{3}$$

where λ is the set of model parameters which weigh the contribution of each individual saliency map. Here $\sigma(\cdot)$ denotes the sigmoid function:

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \tag{4}$$

1.2.2 Aggregation using conditional random field

Conditional Random Field (CRF) is applied to aggregate saliency analysis results from multiple methods. Each pixel is modeled as a node. Each node is associated with a saliency feature vector $\mathbf{x}(p) = (S_1(p), S_2(p), \ldots, S_m(p))$ and a binary random label y_p. The conditional distribution of labels $Y = \{y_p | p \in I\}$ on the features $X = \{x_p | p \in I\}$ as follows:

$$P(Y|X;\theta) = \frac{1}{Z} \exp\left(\sum_{p\in I} f_d(\mathbf{x}_p, y_p) + \sum_{p\in I}\sum_{q\in N_p} f_s(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)\right) \tag{5}$$

where p is a pixel in image I, $x_p$is its feature, and y_p is its saliency label. θ is the CRF model parameters. $f_d(\mathbf{x}_p, y_p)$ is the feature function that defines the relationship between the feature and label. $f_s(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)$ is another feature function that defines the feature-dependent relationship between the labels of neighboring pixels p and q. N_p is the set of pixels that are directly connected to p. It is considered as 8-connection neighborhood here. Z is a constant.

Feature function $f_d(\mathbf{x}_p, y_p)$, based on only the input saliency maps S_i, is defined as follows:

$$f_d(\mathbf{x}_p, y_p) = \sum_{i=1}^{m} \lambda_i S_i(p) y_p + \lambda_{m+1} y_p \tag{6}$$

where λ_i is a subset of the CRF model parameters and $S_i(p)$ is the saliency value at pixel p in the saliency map S_i.

The feature function $f_s(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)$ has two components, f_e and f_c, to model the data-dependent relationship between the labels of neighboring pixels.

$$f_s(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) = f_e(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) + f_c(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) \tag{7}$$

Left hand side of Equation 7, $f_e(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)$, determines if two pixels have different saliency values according to an individual saliency method, they take different saliency labels in the aggregation result. Particularly, if a pixel takes a high saliency value than its neighbor in an individual saliency map, it is also likely to take a more salient label after aggregation.

$$f_e(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) = \sum_{i=1}^{m} \alpha_i (1(y_p = 1, y_q = 0) - 1(y_p = 0, y_q = 1))(S_i(p) - S_i(q)) \tag{8}$$

where α_i are CRF model parameters and $1(\cdot)$ is an indicator function.
$f_c(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q)$ inpired by [5] idea to give same label to neighboring pixels with similar colors

$$f_c(\mathbf{x}_p, \mathbf{x}_q, y_p, y_q) = -1(y_p \neq y_q) \exp(-\eta \|I(p) - I(q)\|) \tag{9}$$

where $\|I(p) - I(q)\|$ is the color difference between pixel p and q in the RGB color space. η is set as $(2 < \|I(p) - I(q)\|^2 >)^{-1}$, where $< \cdot >$ denotes the expectation operator.

The CRF aggregation model parameters $\Theta = \{\lambda, \alpha\}$ are optimized to maximize the likelihood on the training data.

1.2.3 Image-Dependent saliency aggregation

To improve the global saliency aggregation method they upgrade the aggregation model from $P(Y|X;\theta)$ into $P(Y|X;\theta(I))$ for each image I. Here, $\theta(I)$ indicates that the model parameters are customized to image I. Given an input image, the method first finds its k nearest neighbors in the training set and then trains a saliency aggregation model using these k images. The GIST descriptor is used to find similar images. L_2 distance is employed to compute the distance between GIST descriptors, as suggested in [6].

2 Experimental results

The method is experimented on two datasets FT [2] and SS [3].

2.1 Quantitative results

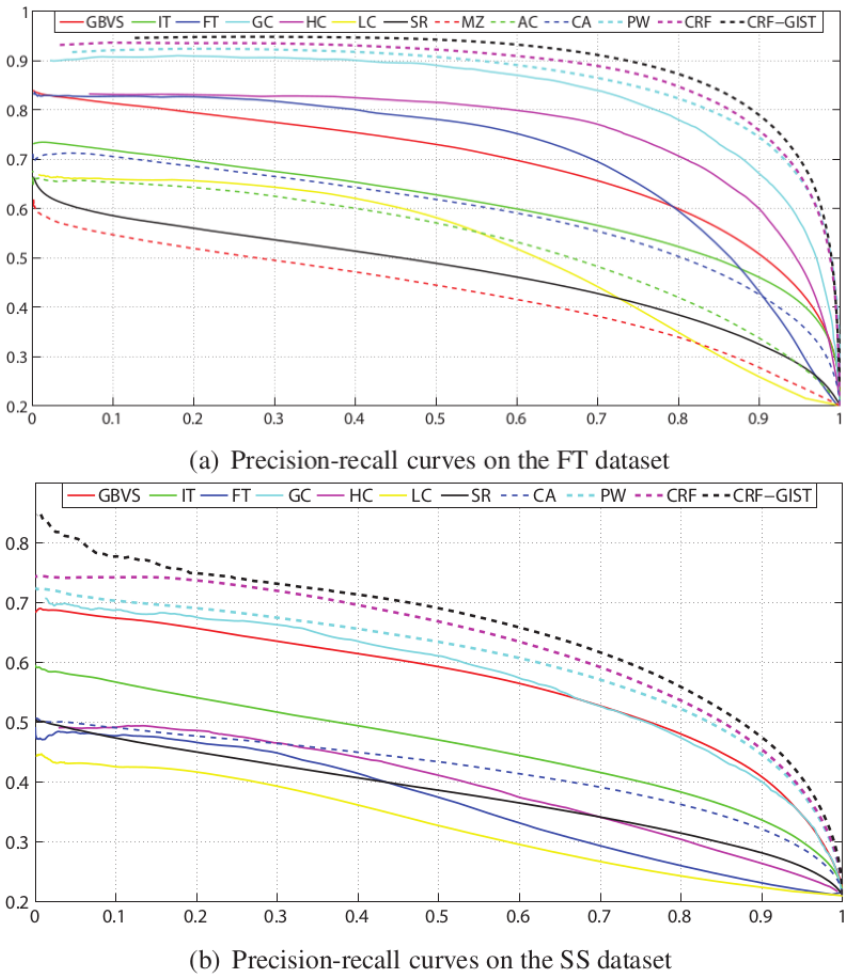

Figure 1. Precision-recall curves of the saliency aggregation approaches, including PW, CRF, and CRF-GIST.

2.2 Qualitative results

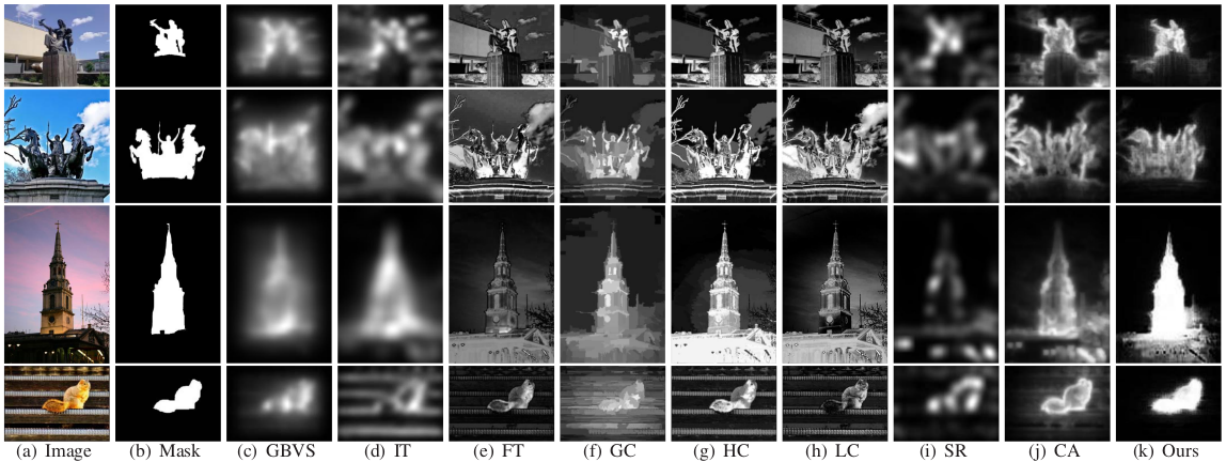

Figure 2. Saliency aggregation examples. (a) shows the input images, the ground-truth is (b), individual saliency maps (c-j), and the aggregation results using image-dependent CRF aggregation method.

References

[1] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Computer Vision – ECCV 2012* (A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), (Berlin, Heidelberg), pp. 414–429, Springer Berlin Heidelberg, 2012. 1

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, 2009. 1, 4

[3] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 454–461, 2012. 1, 4

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-VerlagBerlin, Heidelberg, 2006. 1

[5] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. 3

[6] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, May 2001. 3