Learning optimal seeds for diffusion-based salient object detection: A Report

Paper's Writers

Song Lu, Vijay Mahadevan, Nuno Vasconcelos

Advisor

Dr. Maryam Abedi

Student

Mohammad Shahpouri

September
2022

Contents

List of Tables

List of Equations

1 Optimal seeds for object saliency

Combination of pre-attentive saliency maps and mid-level vision cues for object perception are proposed as an algorithm for learning saliency seeds.

1.1 Feature based saliency diffusion

An image $\mathbf{x}$ is first segmented into N superpixels $\{x_i\}$, $i = (1 \dots N)$, using the algorithm of [1], and represented as a graph $G = (V, E)$ where each node corresponds to a superpixel. Following [2], affinity matrix is adopted

$$w_{ij} = \begin{cases} \nu(\mathbf{x}_i, \mathbf{x}_j) & \text{if } j \in \mathcal{N}_i \text{ or } \exists k \in \mathcal{N}_i | j \in \mathcal{N}_k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathcal{N}_i$ is the set of neighbors of superpixel x_i. The function $\nu(\mathbf{x}_i, \mathbf{x}_j)$ is a measure of visual similarity of two superpixels. They adopt a classifier to determine if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same object [3, 4]. A boosted decision tree are utilized as the classifier, which operates on a feature space that accounts for color difference, texture difference, and geometric properties such as distance.

1.2 Seed representation

Seed vector $\mathbf{s}$ consider as a linear combination

$$\mathbf{s} = \mathbf{F}(\mathbf{x})\mathbf{w} \tag{2}$$

where $\mathbf{F}$ is an $N \times K$ matrix, whose columns are the responses of K features to image $\mathbf{x}$. The weight vector $\mathbf{w}$ determines the contribution of the different features to the seed vector. With these seeds the saliency map can be written as

$$\begin{aligned} \mathbf{y}(\mathbf{x}) &= \mathbf{A}(\mathbf{x})\mathbf{F}(\mathbf{x})\mathbf{w} \\ &= \sum_i w_i \mathbf{A}(\mathbf{x})\mathbf{f}_i(\mathbf{x}) \end{aligned} \tag{3}$$

where $\mathbf{f}_i(\mathbf{x})$ is the i^{th} column of $\mathbf{F}(\mathbf{x})$ and contains the saliency information derived from feature i and $\mathbf{A}(\mathbf{x})$ is a diffusion matrix, as defined in Table 1.

Table 1. Diffusion matrices for graph-based similarity propagation.

Method	Propagation matrix $\mathbf{A}$
Quadratic energy models	$(\mathbf{K}+\lambda(\mathbf{D}-\mathbf{W}))^{-1}\mathbf{K}$
Random walks	$(\mathbf{I}-\alpha\mathbf{W}\mathbf{D}^{-1})^{-1}$
Manifold ranking	$(\mathbf{I}-\alpha\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2})^{-1}$

1.3 Learning

Learning is accomplished by optimizing $\mathbf{w}$

$$\begin{aligned}\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\alpha\|w\|^2 \\ + \sum_k \sum_{\{ij|\delta_i^k=1,\delta_j^k=0\}} \max(0, 1-(y_i(\mathbf{x}^{(k)}) - y_j(\mathbf{x}^{(k)})))\end{aligned} \tag{4}$$

where $\mathbf{y}(\mathbf{x})$ is the object saliency map of 3 and $\delta_i^k = m_i(\mathbf{x}^(k))$ an indicator of the saliency of the i^{th} superpixel of the k^{th} image.

1.4 Pre-attentive saliency map

To obtained pre-attentive saliency map, images are first decomposed into patches $\{\mathbf{t}_i\}$ of 8×8 pixels and 3 RGB color channels. Then approximating the sparse coefficients of patch $\mathbf{t}$ with $\xi = \Psi^{-1}\mathbf{t}$. The sparse coefficients are used as features for the computation of saliency. ξ_i is well approximated by a generalized Gaussian distribution (GGD)

$$p(\xi_i;\alpha_i,\beta_i) = \frac{\beta_i}{2\alpha_i\Gamma(\frac{1}{\beta_i})}\exp\left(-\left|\frac{\xi_i}{\alpha_i}\right|^{\beta_i}\right) \tag{5}$$

where α_i is a scale parameter, β_i a shape parameter, and $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}dt$, $t>0$, the Gamma function. α_i is estimated by the maximum a posteriori (MAP) under the assumption of a conjugate (Gamma distributed) prior, is given by [5]

$$\hat{\alpha}_{i,MAP} = \left[\frac{1}{k_i}\left(\sum_{j=1}^{n}|\xi_i^{(j)}|^{\beta_i}+\nu\right)\right]^{1/\beta_i} \tag{6}$$

where $k_i = \frac{n+\eta}{\beta_i}$ and ν, η are fixed prior parameters ($\nu = \eta = 10^{-3}$). As in [6], β_i is learned for each image, using

$$\sigma^2 = \frac{\alpha^2 \Gamma(\frac{3}{\beta_i})}{\Gamma(\frac{1}{\beta_i})} \quad \kappa = \frac{\Gamma(\frac{1}{\beta_i})\Gamma(\frac{5}{\beta_i})}{\Gamma(\frac{3}{\beta_i})^2} \tag{7}$$

where σ^2 and κ are the variance and kurtosis of the i^{th} feature. As in [7], the saliency of image location l is

$$s(\xi_i(l); \alpha_i, \beta_i) = -\log p(\xi_i(l); \alpha_i, \beta_i) \tag{8}$$

Using (5) and (6) leads to

$$s(\xi_i(l)) = \frac{|\xi_i(l)|^{\beta_i}}{\frac{1}{k}(\sum_{j=1}^{n} |\xi_i^{(j)}|) + \nu} + Q \tag{9}$$

where Q is a constant that does not depend on $\xi_i(l)$. Finally, the saliency maps derived from the B feature channels are combined with

$$s_f(l) = \sum_{1}^{B} a_i s(\xi_i(l)) \tag{10}$$

and the pre-attentive saliency score of superpixel x_i set to the mean saliency score of its pixels. The weights a_i are learned.

1.5 Mid-level features

They proposed a bunch of mid-level features to evaluate the likelihood of a superpixel belonging to a generic object.

Element uniqueness measures the rarity of a superpixel color [8] with

$$\mathcal{U}_i = \sum_{j=1}^{N} \|c_i - c_j\|^2 w_{ij}^{(l)} \tag{11}$$

$$w_{ij}^{(l)} = \frac{1}{Z_i} \exp(-\frac{1}{2\sigma_l^2} \|l_i - l_j\|^2) \tag{12}$$

where l_i and c_i are the position and average CIELab color of the i^{th} superpixel, respectively. Z_i is a normalization factor to ensure that $\sum_{j=1}^{N} w_{ij}^{(l)} = 1$.

Element distribution measures the spatial variance of the color of a superpixel [8], according to

$$\mathcal{D}_i = \sum_{j=1}^{N} \|l_j - l_i^{(\mu)}\|^2 w_{ij}^{(c)} \tag{13}$$

where $w_{ij}^{(c)} = \frac{1}{Z_i}\exp(-\frac{1}{2\sigma_c^2}\|c_i - c_j\|^2)$ and $l_i^{(\mu)} = \sum_{j=1}^{N} w_{ij}^{(c)} l_j$ is the center of mass of color c_i. Z_i is a normalization constant such that $\sum_{j=1}^{N} w_{ij}^{(c)} = 1$.

Pattern distinctness [9] measures the l_1 mean distance of patches in a superpixel to the mean patch, by principal component analysis (PCA). This is defined as $\mathcal{P}(\mathbf{x}_i) = \|\hat{\mathbf{x}}_i\|_1$, where $\hat{\mathbf{x}}_i$ contains the PCA coefficients of patch x_i.

Color distinctness same as pattern distinctness but for a PCA of the RGB color space of each patch.

Center bias distance between superpixel center and image center, normalized to $[0, 1]$.

Backgroundness similarity between a superpixel and the superpixels in the four image boundaries, using the similarity measure $\nu(x_i, x_j)$ of (1).

Local contrast measures [3] based on Chi-square distances between distributions of color and texton response [10] and geometric attributes such as size, and position.

2 Experiments

2.1 Eye fixation prediction

Table 2. Eye fixation prediction performance: Gaussian kernel/Shuffled AUC score

Gaussian kernel/Shuffled AUC score	Kootstra	Bruce	Judd	VOC2008_1000
RGB-Signature [11]	0.030/0.5869	0.040/0.6900	0.040/0.6547	0.065/0.6497
LAB-Signature [11]	0.040/0.6020	0.045/0.7115	0.040/0.6631	0.050/0.6595
Sparse [12]	0.015/0.6024	0.030/0.6956	0.020/0.6629	0.030/0.6491
Spectral [13]	0.040/0.5865	0.040/0.6898	0.040/0.6545	0.065/0.6527
SUN [14]	0.020/0.5609	0.030/0.6663	0.030/0.6565	0.050/0.6373
sparse-GGD	0.015/**0.6105**	0.030/**0.7140**	0.035/**0.6751**	0.050/**0.6681**

2.2 Salient object detection

Table 3. Object saliency detection performance: AUC/AP

AUC/AP	MSRA5000	SOD	SED1	SED2	VOC2008_1023
CB	0.9281/0.8289	0.7672/0.6235	0.9105/0.8380	0.8741/0.7767	0.7546/0.6158
FT	0.7605/0.5603	0.6078/0.4274	0.6699/0.5493	0.8205/0.7225	0.6071/0.4493
Gof	0.8622/0.6214	0.8027/0.5818	0.8513/0.6804	0.8617/0.6474	0.7847/0.5959
HC	0.8223/0.6452	0.6612/0.4646	0.7770/0.6311	0.8769/0.7773	0.6525/0.4756
RC	0.9200/0.7724	0.8133/0.6337	0.8881/0.7633	**0.9142/0.8272**	0.7965/0.6186
GBMR	0.9424/0.8614	0.8319/0.6759	0.9341/0.8841	0.8360/0.7548	0.7838/0.6442
PCA	0.9407/0.8057	0.8414/0.6423	0.9085/0.7862	0.9035/0.7905	0.8102/0.6451
SalseedProp	0.9058/0.8136	0.8175/0.6688	0.9176/0.8537	0.8806/0.7500	0.7908/0.6421
OptseedProp	**0.9615/0.8790**	**0.8684/0.7019**	**0.9530/0.8905**	0.9058/0.8062	**0.8181/0.6556**

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels," *École Polytechnique Fédéral de Lausssanne (EPFL)*, vol. 149300, p. 15, 2010. 1

[2] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3166–3173, 2013. 1

[3] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2083–2090, 2013. 1, 4

[4] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *2011 International Conference on Computer Vision*, pp. 1052–1059, 2011. 1

[5] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Comput*, vol. 21, pp. 239–271, Jan. 2009. 2

[6] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–6, 2007. 3

[7] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems* (Y. Weiss, B. Schölkopf, and J. Platt, eds.), vol. 18, MIT Press, 2005. 3

[8] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–740, 2012. 3, 4

[9] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1139–1146, 2013. 4

[10] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, pp. 29–44, Jun 2001. 4

[11] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012. 4

[12] X. Hou and L. Zhang, "Dynamic visual attention: searching for coding length increments," in *Advances in Neural Information Processing Systems* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), vol. 21, Curran Associates, Inc., 2008. 4

[13] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. 4

[14] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, “SUN: A Bayesian framework for saliency using natural statistics,” *Journal of Vision*, vol. 8, pp. 32–32, 12 2008. 4