A framework for visual saliency detection with applications to image thumbnailing report

Paper's Authors

Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, Junmo Kim

Advisor

Dr. Maryam Abedi

Student

Mohammad Shahpouri

October
2022

Contents

List of Figures

List of Equations

1 The framework

1.1 High level visual features

The Fisher Kernel [5] are employed as high level image descriptors.

$$\nabla \log p(X|\lambda) = \sum_{t=1}^{T} \nabla \log p(x_t|\lambda) \tag{1}$$

where $X = \{x_t, t = 1 \dots T\}$ and λ is the generative model parameters should be modified to best fit the data set X. A Gaussian mixture model (GMM) is employed to build a visual vocabulary. $\lambda = \{w_i, \mu_i, \sigma_i, i = 1 \dots N\}$, where N is the number of Gaussians and w_i, $_i$ and σ_i are respectively the weight, the mean vector and the variance vector. The GMM is trained using maximum likelihood estimation (MLE). The probability that it was generated by the GMM is $p(x_t|\lambda) = \sum_{i=1}^{N} w_i p(x_t|\lambda)$, where:

$$p(x_t|\lambda) = \frac{\exp\{-\frac{1}{2}(x_t - \mu_i)' \Sigma_i^{-1}(x_t - \mu_i)\}}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \tag{2}$$

The partial derivatives of $\log p(x_t|\lambda)$ according to the GMM parameters can be computed by the following formulas [5]:

$$\frac{\partial \log p(x_t|\lambda)}{\partial \mu_i^d} = \gamma_i(x_t) \left[\frac{x_t^d - \mu_i^d}{(\sigma_i^d)^2} \right] \tag{3}$$

$$\frac{\partial \log p(x_t|\lambda)}{\partial \sigma_i^d} = \gamma_i(x_t) \left[\frac{(x_t^d - \mu_i^d)^2}{(\sigma_i^d)^3} - \frac{1}{\sigma_i^d} \right] \tag{4}$$

where the superscript d denotes the d-th dimension of a vector and $\gamma_i(x_t)$ is the occupancy probability given by:

$$\gamma_i(x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{N} w_j p_j(x_t)} \tag{5}$$

The partial derivatives, equations (3) and (4), give us a $2 \times D \times N$ diminsional vector, where D is the dimension of the low level feature space. From (1) the Fisher Vector of the set of descriptors $X = \{x_t, t = 1 \dots T\}$ is the sum of individual Fisher vectors:

$$\nabla \log p(X|\lambda) = \mathbf{f}_X = \sum_{t=1}^{T} \mathbf{f}_t \tag{6}$$

1.2 Image indexation and retrieval

To thumbnail an image the K most similar images are retrieved as follows. First, a set of local image patches are extracted with their low level descriptors $Y = \{y_1, y_2, \ldots, y_M\}$. Visual vocabulary (GMM) is used and Fisher vector $\mathbf{f_Y}$ computed. To compute similarities between two images the following normalized L_1 similarity is employed:

$$sim(X,Y) = -\|\hat{\mathbf{f}}_X - \hat{\mathbf{f}}_Y\|_1 = -\sum_i |\hat{f}_X^i - \hat{f}_Y^i| \tag{7}$$

where $\hat{\mathbf{f}}$ is the vector $\mathbf{f}$ normalized to norm L_1 equal 1 ($\|\hat{\mathbf{f}}\|_1 = 1$) and $\mathbf{f}_X = \mathbf{f}_{X^+} + \mathbf{f}_{X^-}$ represents the global set of descriptors (salient and non salient) of image X.

1.3 Saliency detection

All Fisher vectors associated to the K similar images for salient and non-salient regions are summed:

$$\mathbf{f}_{FG} = \sum_{j=1}^{K} \mathbf{f}_{X_j^+} \quad \text{and} \quad \mathbf{f}_{FG} = \sum_{j=1}^{K} \mathbf{f}_{X_j^-} \tag{8}$$

A patch x_i then is considered salient, if its normalized L_1 distance to the foreground Fisher model is smaller than to the background Fisher model:

$$\|\hat{\mathbf{f}}_{x_i} - \hat{\mathbf{f}}_{FG}\|_1 < \|\hat{\mathbf{f}}_{x_i} - \hat{\mathbf{f}}_{BG}\|_1 \tag{9}$$

In order to increase the model's robustness, Fisher vectors of patches over a neighborhood $\mathcal{N}$ are summed:

$$\mathbf{f}_{\mathcal{N}} = \sum_{x_i \in \mathcal{N}} \mathbf{f}_i \tag{10}$$

The binary classifier is replaced with non-binary score which is a simple function of the normalized L_1 distances:

$$s(\mathcal{N}) = \|\hat{\mathbf{f}}_{\mathcal{N}} - \hat{\mathbf{f}}_{FG}\|_1 - \|\hat{\mathbf{f}}_{\mathcal{N}} - \hat{\mathbf{f}}_{BG}\|_1 \tag{11}$$

Finally, to build a "saliency map" $\mathbf{S}$, value of each region center $s_{\mathcal{N}} = s(\mathcal{N})$ either can be interpolated the values between the centers or use a Gaussian propagation of the values. The Gaussian weigthed scores can be done as follows:

$$s(p) = \frac{\sum_{\mathcal{N}} s_{\mathcal{N}} w_{\mathcal{N}}(p)}{\sum_{\mathcal{N}} w_{\mathcal{N}}(p)} \tag{12}$$

where $w_{\mathcal{N}}$ is the value in pixel p of the Gaussian centered in the geometrical center of each the region $\mathcal{N}$.

1.4 Map refinement and thumbnail extraction

Two thresholds are chosen (one positive th_+ and one negative th_-) that separate the saliency map $\mathbf{S}$ into 3 different regions: pixels u labeled as salient ($\mathbf{S}(u) > th_+$), pixels v labeled as non-salient ($\mathbf{S}(v) < th_-$) and unknown (the others). Two Gaussian Mixture Models (GMMs) Ω_1 and Ω_0 are created, one using RGB values of salient (foreground) pixels and one using RGB values of non salient (background) pixels. Then the following energy is minimized:

$$E(L) = \sum_{u \in \mathcal{P}} D_u(u) + \sum_{(u,v) \in \mathcal{C}} V_{u,v}(u, v) \tag{13}$$

where the data penalty function $D_u(u) = -\log p(u|l_u, \Omega_{l_u})$ is the negative log likelihood that the pixel u belongs to Ω_{l_u}, with $l_u \in 0, 1$ and the contrast term:

$$V_{u,v}(u, v) = \gamma \sum_{(u,v) \in \mathcal{C}} \delta_{l_u,l_v} \exp(-\frac{\|u - v\|^2}{2 \times \beta}) \tag{14}$$

with $\delta_{l_u,l_v} = 1$ if $l_u = l_v$, $\mathcal{C}$ representing 4-way cliques, and $\beta = \mathbf{E}(\|u - v\|^2)$.

Min-cut/max-flow algorithm [6] is utilized to minimize the energy (13) leading to a binary annotation of the image. Using the new labels, To update the two GMM parameters and similarly to [7] iterate between energy minimization (13) and GMM updates until no modifications are made to the binary labels. This binary map can be considered as a new saliency map, denoted by $\mathbf{S}_G$.

In order to determine to keep $\mathbf{S}_B$ or $\mathbf{S}_G$ the following equation is applied:

$$\mathbf{S}^* = \begin{cases} \mathbf{S}_G & \text{if } \frac{\mathbf{S}_B \cap \mathbf{S}_G}{\mathbf{S}_B \cup \mathbf{S}_G} > th_d \\ \mathbf{S}_B & \text{otherwise} \end{cases} \tag{15}$$

with $0 < th_d < 1$, $th_d = 0.1$

Finally, two strategies are applied to extract a thumbnail: (1) select the biggest, most centered salient region as thumbnail or alternatively, (2) all the detected salient regions and re-target them into a single thumbnail as proposed in [8].

2 Experimental results

MRSA dataset [3] and PASCAL VOC dataset [4] are utilized to evaluate the proposed model.

2.1 Quantitative evaluation

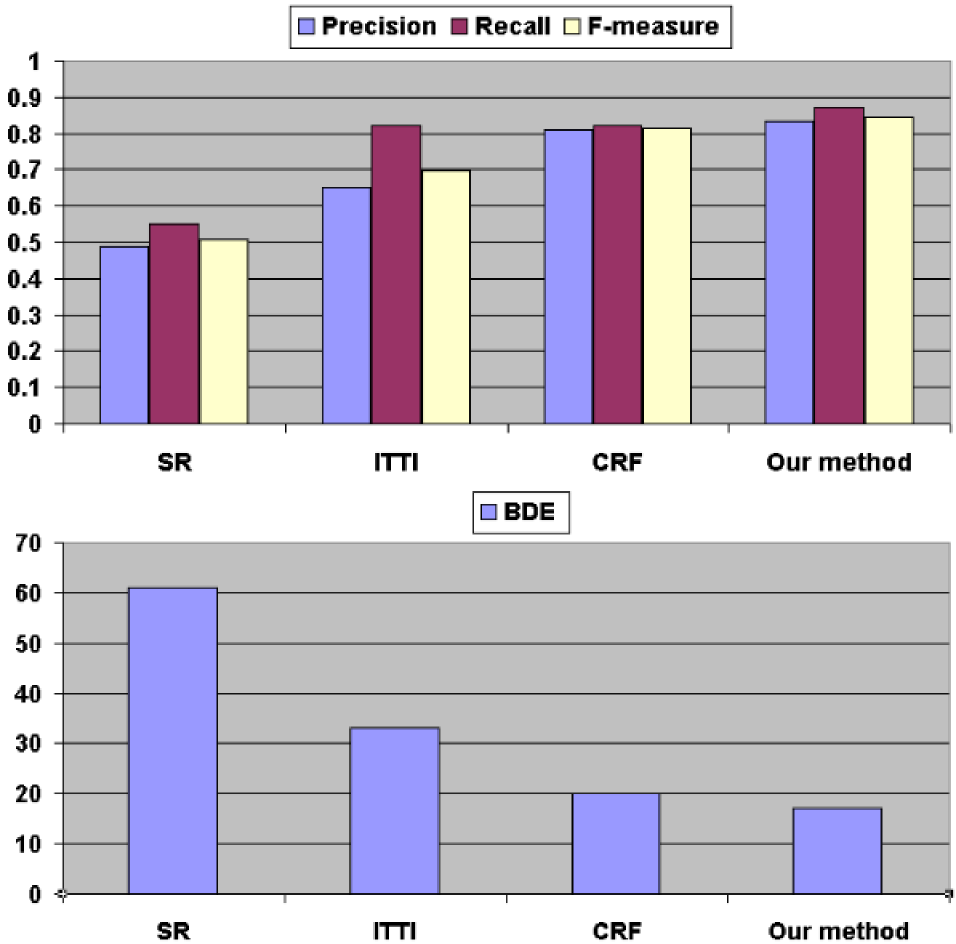

Figure 1. Above, comparison of our method with (SR) [1], (ITTI) [2] and (CRF) [3] using precision, recall, F_α measures. Below, comparison of the same methods using BDE (Bounding Box Displacement Error).

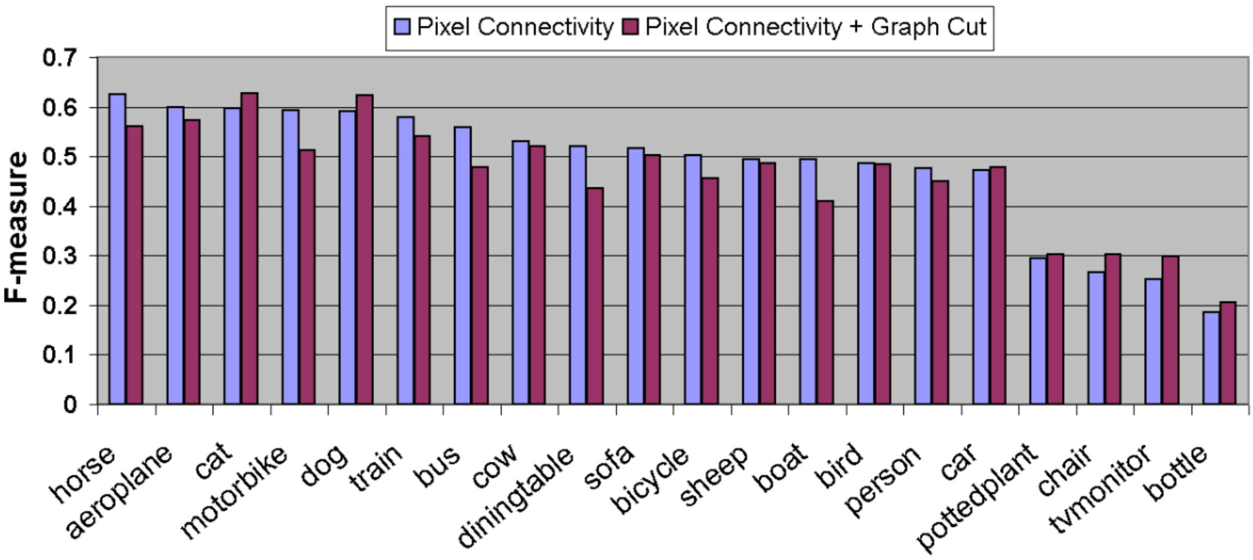

Figure 2. Performances (F_α) in the PASCAL dataset [4] with and without Graph-Cut.

2.2 Qualitative evaluation

Figure 3. Some qualitative results collected in MRSA dataset obtained using Graph-Cut refinement.

Figure 4. Some qualitative results collected in MRSA dataset where the decision mechanism rejected Graph-Cut refinement.

Figure 5. Some qualitative results collected in PASCAL dataset.

References

[1] 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. ii, 4

[2] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10, pp. 1489–1506, 2000. ii, 4

[3] 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. ii, 4

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, Jun 2010. ii, 4, 5

[5] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. 1

[6] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004. 3

[7] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, p. 309–314, aug 2004. 3

[8] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch, "Automatic image retargeting," MUM '05, (New York, NY, USA), p. 59–68, Association for Computing Machinery, 2005. 3