A Co-Saliency Model of Image Pairs report

Paper's Authors

Hongliang Li, King Ngi Ngan

Advisor

Dr. Maryam Abedi

Student

Mohammad Shahpouri

October
2022

Contents

List of Figures

List of Equations

1 Proposed Method

Single-image saliency and multi-image saliency maps computed to obtain co-saliency.

1.1 Single-Image Saliency Map (SISM)

They propose a voting algorithm. Three types of saliency maps, Itti's model saliency [1], frequency-tuned saliency (FTA) [2], and spectral residual saliency (SRA) [3] are calculated to make saliency map. Assume I denotes an input image, while S_l represents the corresponding single-image saliency map.

$$S_l = \sum_{j=1}^{J} w_j \cdot \mathcal{N}(S_{l_j}) \tag{1}$$

where $\mathcal{N}(S_{l_j}$ denotes the jth normalized saliency map where each pixel has the salient value in the range $[0, 1]$. Here, w_j denotes the weight with $\sum_{j=1}^{J} w_j = 1$. If a pixel is identified as a salient pixel by most of algorithms, it will have a high single-image saliency value.

1.2 Multi-Image Saliency Map (MISM)

To caculate Multi-Image Saliency Map, if the two images contain a similar object, the object region in each image should be assigned high visual saliency values. Otherwise, low multi-image saliency values should be considered for the dissimilar regions. According to this principle, the multi-image saliency map of the image I_i is defined as:

$$S_g(I_i(p)) = \max_{q \in I_j} Sim(I_i(p), I_j(q)) \tag{2}$$

where p and q denote entities (e.g., pixels or regions) in images I_i and I_j, respectively. $Sim()$ represents a function that measures the similarity between two entities.

Multi-image saliency detection consists of four stages, namely pyramid decomposition, feature extraction, Sim-Rank optimization, and multi-image saliency computation.

1) Pyramid Decomposition of an Image Pair: This stage is used to obtain a pyramid of images with decreasing resolutions.

2) Region Feature Extraction: Color and texture descriptors are used as descriptors of regions. RGB, L*a*b* and YCbCr color spaces represent the color feature. using the k-means to cluster all pixels in the image pair into N clusters. Each cluster center is called a codeword. The histogram is computed for each region by counting the number of codewords at each bin (i.e., cluster).

The texture descriptor is created only from RGB color space. Consider an image pair, $p \times p$ patches are extracted from color image. Each patch is concatenated to form a single vector of size p^2. Then, to generate M clusters, the k-means clustering is performed over all vectors. To generate the final texture descriptor the component histograms are concatenated.

$$f^t(k) = [H_{3\times3}(k), H_{5\times5}(k), H_{7\times7}(k), \ldots] \tag{3}$$

where $H_{i\times i}(k)$ denotes the histogram computed for the kth region of size $i \times i$.

3) The Co-Multilayer Graph Representation: The similarity is measured in this step. First, a co-multilayer graph $G = (V, E)$ with nodes $v \in V$ and edges $e \in E$, where the nodes $V = \{V^I \cup V^{II}\}$ denote a set of regions. Two nodes v_i and v_j are connected by the directed links e_{ij} and e_{ji}, which have weights $w(e_{ij})$ and $w(e_{ji})$, respectively. the weight w_{ij} for the edge

$$e_{ij}$$

is defined as:

$$w_{ij} = \begin{cases} \exp(-\theta_f d(f_i, f_j)), & \text{if } l_i - l_j = -1 \text{ or } l_i - l_j = 0 \\ 0, & \text{if } \|l_i - l_j\| > 1 \text{ or } l_i > l_j \end{cases} \tag{4}$$

with

$$d(f_i, f_j) = \chi^2(f_i, f_j) = \sum_{z=1}^{Z_f} \frac{(f_i(z) - f_j(z))^2}{f_i(z) + f_j(z)} \tag{5}$$

where f_i and f_j denote the color or texture descriptor for regions i and j, respectively. Z_f denotes the dimensional number of the descriptor. θ_f is a constant that controls the strength of the weight. $\chi^2()$ denotes the chi-square distance.

4) Normalized Simrank Similarity Computation: SimRank [4] is utilized to measure the similarity of two region nodes. Let $s(a, b)$ denote the similarity score between objects a and b, which is defined as:

$$s(a, b) = \frac{C}{|In(a)||In(b)|} \sum_{i=1}^{|In(a)|} \sum_{j=1}^{|In(b)|} s(In_i(a), In_j(b)) \tag{6}$$

where C is a decay factor between 0 and 1. $|In(a)|$ and $|In(b)|$ denote the numbers of in-neighbors $In(a)$ and $In(b)$ for nodes a and b, respectively.

Therefore, the normalization of the SimRank score is employed to measure the similarity:

$$s^*(a, b) = \frac{s(a, b)}{\max(s(a, a), s(b, b))} \tag{7}$$

$s^*(a, b) = 1$ when the nodes a and b share the same sub-region nodes.

Substituting Equation 7 into Equation 2, the multi-image saliency map can be rewritten as:

$$S_g(I_i(p)) = \max_{q \in I_j} s^*(I_i(p), I_j(q)) \tag{8}$$

where p and q denote the region nodes in an image pair (I_i, I_j).

1.3 Co-Saliency Map

Two saliency maps, Equation 1 and Equation 8, are combined to compute co-saliency map:

$$\begin{aligned} SS(I_i(p)) &= \alpha_1 \cdot S_l(I_i(p)) + \alpha_2 \cdot S_g(I_i(p)) \\ &= \alpha_1 \cdot S_l(I_i(p)) + \alpha_2 \cdot (\alpha_3 \cdot S_g^c(I_i(p)) + \alpha_4 S_g^t(I_i(p))) \\ &= \beta_1 \cdot S_l(I_i(p)) + \beta_2 \cdot S_g^c(I_i(p)) + \beta_3 \cdot S_g^t(I_i(p)), \text{ for all } p \in R\{I_i\} \end{aligned} \tag{9}$$

where β_j is a constant with $\beta_1 + \beta_2 + \beta_3 = 1$ that is used to control the impact of the SISM and MISM on the image co-saliency. S_g^c and S_g^t denote the MISMs obtained by color and texture descriptors, respectively.

2 Experimental Results

The performance of the method is evaluated on several image pairs, which were used in [5] together with additional image pairs that we collected from various databases such as Microsoft Research Cambridge image database, the Caltech-256 Object Categories database, and PASCAL VOC dataset.

2.1 Quantitative Result

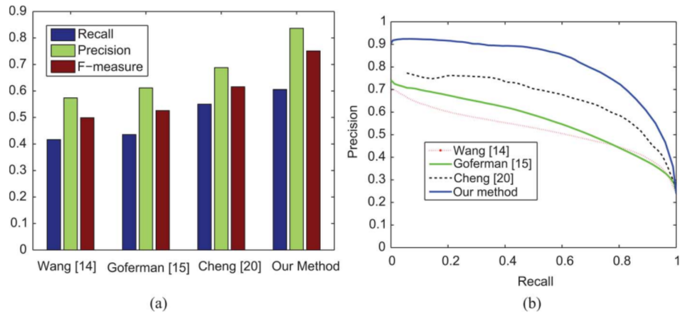

Figure 1. Evaluation results for 210 images. (a) Precision-recall bars for adaptive thresholds. (b) Precision-recall curves for varying thresholds.

2.2 Qualitative Result

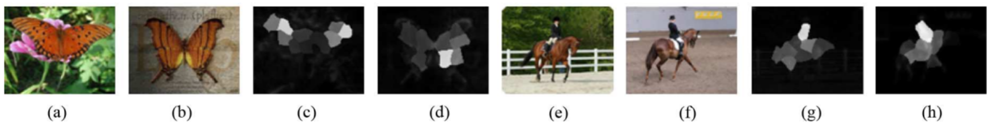

Figure 2. Experimental results for single objects. (a)-(b) and (e)-(f): Original image pairs. (c)-(d) and (g)-(h): Results by proposed method.

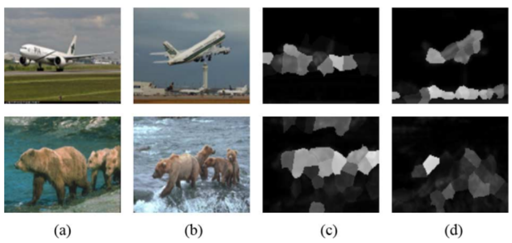

Figure 3. Experimental results for multiple objects. (a)-(b): Original image pairs. (c)-(d): Results by proposed method.

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998. 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

[3] 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. 1

[4] G. Jeh and J. Widom, "Simrank: A measure of structural-context similarity," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, (New York, NY, USA), p. 538–543, Association for Computing Machinery, 2002. 2

[5] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 269–276, 2009. 3