Cosaliency Detection from Superpixel-Scale Saliency Maps Fusion

Chung-Chi Tsai

chungchi@tamu.edu

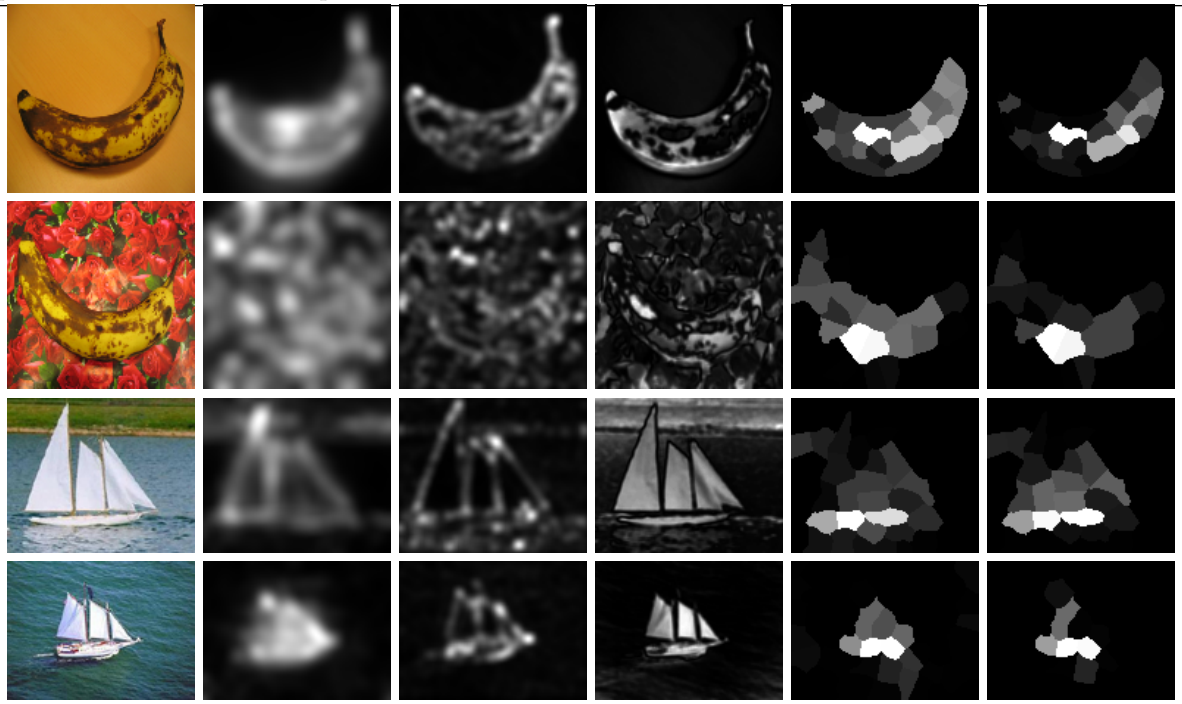
Department of Electrical and Computer Engineering, Texas A&M University.

July 3, 2016

1 Introduction

Salient object detection has gained great attentions in recent years because it allows us to extract the objects without prior knowledge on the image content. Many works have demonstrated its significance in fields such as image segmentation, object recognition and visual tracking, etc. These days, many efficient saliency detection methods has been proposed for single image based on various feature contrast measures. Different saliency measures catch different cues from images, thus they usually give different saliency detection performance. Inspired by primitive human visual system, Itti(IT)[1] proposed a center-surround differences across multiple-scale image features on intensity, colors and orientations to detect the saliency. Despite its novelty, this method poorly define the object border since it discards many high spatial frequency content. With purely computational points of view, Hou(SR)[2] defines the saliency through residual on the log-frequency domain. Despite its compational efficiency, this method mostly shows the object boundary rather than uniformly highlight the whole salient regions. Since Itti and Hou's methods need to resize the images, these process in-evidently sacrifice a lot of frequency content, Achanta(FT)[3] improved them with a full resolution method that able to uniformly highlight salient regions and well define the object boundaries by using the color difference from the average color of entire image. Later, Chang [4] proposed a more effective global contrast method that improves the drawbacks of Achanta's model by designing the contrast cues from comparing respective information with the data mean to comparing with every other pixels(HC) or regions(RC). However, with the advent of saliency detection, the complexity of the foregrounds and the backgrounds still make it difficult to accurately detect saliency regions. Beyond the single image saliency detection method, Li[5] proposed a cosaliency idea in early stage with focus on paired images to achieve more accurate saliency detection, since the salient objects can be more accurately located with reference to the information from the other image.

Many existing co-saliency model comprised of two kinds of maps, i.e., intra-image(local) saliency maps and inter-image(global) saliency maps. A local saliency map is constructed from salient region detection on single image as described above, while a global saliency map is obtained by detecting the similar regions shared by a group of images. In general, similar regions detection are composed of two steps; (1) feature extractions (2) feature matching or clustering. Several local features, such as, spatial location, color, and texture can be used to form different descriptors, and the region correspondence across images are achieved by a number of ways, for instance, Li[5] utilizes the SimRank algorithm on a co-multilayer superpixel tree to detect the color and texture similarity between superpixels across images. Meng[6] improves the SimRank matching method with consideration of geometric relationship by integrating the pairwise constraints[7]. Besides these paired images regional correspondence learning methods, several other extensions to multiple images inter-image saliency learning models are proposed [8][9]. Finally, the co-saliency map is learned by a fusion of both local and global saliency maps. In overall, most of the existing models fused the saliency maps by either fixed-weight summation or multiplication. Since it is hard to find the ranking of saliency detection methods that can always give superior results than the others, thus these simple fusion methods can not fully utilize the strengths while reduce the weakness among the saliency maps. Therefore, the investigation of saliency maps fusion became a need.



(a) (b) (c) (d) (e) (f)

Table 1: Local and global saliency maps; from left to right are (a)Images (b)IT (c)SR (d)FT (e)CC (f)CP; The global saliency maps are better than local saliency maps in banana images, but vise versa in the sailboat images.

Cao[10] proposed an approach using low-rank matrix recovery by robust principle component analysis (RPCA) technique to self-adaptively re-weight the significance of every saliency map. They assume that the salient regions in a group of images should contain coherent color information by constructing a feature matrix composed of a stack of color histogram from each image, then approximate it with the RPCA. If the approximated color histogram is largely different from the original, then they claim the corresponding saliency method must detect large false positive salient regions that contain non-similar color information with the objects, so that this map should be given a lower weight in the fusion step. However, this approach still contains some space for improvement. For instance, in the first sailboat image, we observe that the SR method gives good result on the object contour, thus, this advantage should be amplified in the contour regions; meanwhile, the FT method uniformly highlights the object area, thus this advantage should also be amplified in the object center area. Last but no the least, in both banana and sailboat images, the complexity of images confused the single image saliency detection method, which leads to some false positive regions. However, since the inter-image saliency map only highlight the regions appears in both images, therefore, it contains much less noise. As a result, we should give higher weight to inter-image saliency maps in the background regions. Based on the above observations, we intend to propose a model that can generate a cosaliency maps in a regional fusion way.

This article is organized as follows. Section II introduces our co-saliency fusion model and section III provides the validation of model with experimental results. Finally, we draw a conclusion in section IV.

2 Problem statement

Given a paired of image I^P and I^Q containing N_1 and N_2 superpixels and a total of K saliency detection methods , we aim at finding the weights $y_{i,j}^k$ ($i \in \{P, Q\}, j = [N_i]$) of the k-th saliency maps on a superpixel scale that able to uniformly highlight the foregrounds through unsupervised learning. Firstly, we choose Linear Spectral Clustering(LSC) method[11] to construct the superpixels for its efficiency and effectiveness.

By reasonably assume that the foregrounds share similar appearance and lies on distinct backgrounds, we design a model that encourage the following conditions;

1. similar superpixels inside or across images possess similar weighting vector $y_{i,j} = [y^1_{i,j}, y^2_{i,j}, \cdots, y^K_{i,j}]$.
2. Meanwhile, if i-th superpixel in image j, namely, $v_{i,j}$ is belonged to the foreground, we should assign higher weight to the saliency map with higher mean saliency value; on the contrary, if belonged to the background, a higher weight should be assigned to the saliency map with lower mean saliency value.

3 The proposed approach

Our approach formulates the task of image cosaliency fusion as an energy minimization problem on a graph. In this section, we introduce the graph structure, energy function, optimization process and the implementation details of our approach.

3.1 Graph construction

We construct a joint graph $G = (V_1 \cup V_2, E_1 \cup E_2)$. In G, each vertex $v_{i,j} \in V_i$ corresponds to a feature point of a superpixel, we extract different types of visual features around each superpixel, including:

Color. Three RGB, three Lab as well as three YCbCr color spaces are extracted together to produce 9 color feature for each pixel. Each color space is normalized to range $[0, 1]$. After then, all pixels in the image pair are quantized into M clusters by using the K-means algorithms. Each cluster center is called a codeword. For each superpixel, we compute the histogram by counting number of codewords at each bin. The color descriptor for a superpixel is represented by the M bins of the histogram.

Gabor filters. Gabor filter responses with 8 orientations, 3 scales and two phase offset, i.e., 0 and $\pi/2$ are extracted. The wavelength of the filters are chosen to be 3,5,7. All 48 features are stacked together and normalized to range $[0, 1]$ to form the texture feature of each pixel. After then, all pixels in the image pair are quantized into M clusters by using K-means algorithms. Following similar ways in constructing the color descriptor, we build the texture descriptor for a superpixel by counting the frequency of each bin in the histogram.

With $f_{i,j}$ and $f_{\hat{i},\hat{j}}$ denote the color or texture descriptor for the superpixel $v_{i,j}$ and $v_{\hat{i},\hat{j}}$ respectively, we use the chi-square distance to measure the feature dissimilarity between vertex.

$$d(f_{i,j}, f_{\hat{i},\hat{j}}) = \chi^2(f_{i,j}, f_{\hat{i},\hat{j}}) = \sum_{m=1}^{M} \frac{(f_{i,j}(m) - f_{\hat{i},\hat{j}}(m))^2}{f_{i,j}(m) + f_{\hat{i},\hat{j}}(m)}.$$

If $d_C(f_{i,j}, f_{\hat{i},\hat{j}})$ represents the color distance and $d_T(f_{i,j}, f_{\hat{i},\hat{j}})$ represents the texture distance, furthermore, let λ_C and λ_T be the weight coefficient controlling the influence and let σ_C and σ_T be the normalization factor, we define the affinity function as

$$A(f_{i,j}, f_{\hat{i},\hat{j}}) = \exp(-[\lambda_C * d_C(f_{i,j}, f_{\hat{i},\hat{j}})/\sigma_C + \lambda_T * d_T(f_{i,j}, f_{\hat{i},\hat{j}})/\sigma_T]).$$

An edge with edge weight $A(f_{i,j}, f_{\hat{i},\hat{j}})$ is added to link $v_{i,j}$ and $v_{\hat{i},\hat{j}}$ if $v_{\hat{i},\hat{j}}$ is one of the spatial neighbors in the same image or among the m nearest neighbors in feature distance. Finally, each vertex $v_{i,j}$ is associated with a weight vector $y_{i,j}$ that describes the significance ranking on each saliency map, where each $y_{i,j}$ component lies between $[0, 1]$ and the summation of weight $\sum_{k=1}^{K} y^k_{i,j} = 1$. We formulate the cosaliency fusion problem as a belief propagation problem on the graph G, and through an optimization model, we intend to encourage the vertex connected with an heavy edge share similar ranking $y_{i,j}$ on the saliency maps. More importantly, the scheme must automatically adjusts the weight distribution, such that the background regions become darker and the foreground regions become brighter after the saliency map fusion.

3.2 Energy Function

We seek a good weighting distribution $Y = [y_{P,1}; y_{P,2}; \cdots, y_{P,N_P}; y_{Q,1}; y_{Q,2}; \cdots; y_{Q,N_Q}]$ by minimizing the following energy function

$$J(Y) = ||Y||^2 + \lambda_1 * U_1(Z_1, Y) + \lambda_2 * U_2(Z_2, Y) + \lambda_3 * B(L, Y)$$
$$\text{s.t.} \sum_{k=1}^{K} y_{i,j}^k = 1, \text{and } 0 \leq y_{i,j}^k \leq 1$$

where λ_1,λ_2 and λ_3 are three non-negative constants used to control the influence of each term. In the objective function, the first unary term is regard as a basis, which give equal weight to all the saliency maps. The second unary term U_1 is considered as a biased solution for the fusion weight. If we set λ_2 and λ_3 equal 0, then the output Y will goes toward the biased solution. The third unary term corresponds to the second condition in section II, which tend to highlight the foreground from the background area. Lastly, the pairwise term corresponds to the first condition in section II, which make sure similar vertex contain similar weight distribution. The definitions of U_1, U_2 and B are given in the following.

3.2.1 Unary term $U_1(Z_1, Y)$

Each saliency map is a nonlinear transformation from the image, thus it can be treated as a feature representation of the image in the saliency feature space. Li[12] proposed a method using the saliency feature plus RPCA for a fusion weight on the superpixel scale. By dividing the image into N_i superpixels, the saliency map k can be represented by a N_i dimensional vector $X_{i,k} = [x_{i,1,k}, x_{i,2,k}, \cdots, x_{i,N_i,k}]$, where the j-th element of the vector corresponds to the mean saliency value of the j-th superpixel $v_{i,j}$ on image i. By stacking K different vectors generated from various saliency detection on a given image, X is seen as a feature representation for image I. By assuming that the background regions usually lie in a low dimensional space, Li use the RPCA to decompose X into a low rank approximate plus a residual matrix E. After then, the magnitude of each column $E_j = (e_{j,1}, e_{j,2}, \cdots, e_{j,K})^T$ is transformed into the strength of saliency on each superpixel, and the magnitude of each component in the column vector can be regarded as the contribution of each saliency map to that region. Here, we let $Z_1 = 1 - E$, and define

$$U_1(Z_1, Y) = tr(Z_1 Y').$$

3.2.2 Unary term $U_2(Z_2, Y)$

In the joint graph, each vertex $v_{i,j}$ of a image is connected with m nearest neighbors on the other image. Let $s_{i,j}$ represent the highest weight(similarity value) among the m candidates. In this way, it can be regarded as an indicator of a good match between superpixels in the paired images. Suppose there is a superpixel $v_{\hat{i},\hat{j}}$ that is very similar to a superpixel $v_{i,j}$. The value of $s_{i,j}$ must be large, and the value of $1 - s_{i,j}$ is small. We design the penalty assigning weight $y_{i,j}$ to the k-th saliency map to be $(1 - s_{i,j})x_{i,j,k}y_{i,j}$. According to this way, assigning a large saliency value does not cause large energy. On the contrary, if $s_{i,j}$ is small, then assigning a large saliency value will induce large energy. By letting the matrix $(Z_{2,i})_{j,k} = (1 - s_{i,j})x_{i,j,k}$, and $Z = [Z_{2,1}; Z_{2.2}]$, we define

$$U_2(Z_2, Y) = tr(Z_2 Y').$$

3.2.3 Pairwise term $B(L, Y)$

Let L be the normalized Laplacian matrix for joint graph G. This pairwise term encourages the smoothness of weight distribution between connected superpixels in the graph, and we define

$$B(L, Y) = YLY'.$$

4 Experiments

References

[1] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998.

[2] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[3] Ravi Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*, pages 1597–1604. IEEE, 2009.

[4] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *IEEE CVPR*, pages 409–416. Citeseer, 2011.

[5] Hongliang Li and King Ngi Ngan. A co-saliency model of image pairs. *Image Processing, IEEE Transactions on*, 20(12):3365–3375, 2011.

[6] Fanman Meng, Hongliang Li, and Guanghui Liu. A new co-saliency model via pairwise constraint graph matching. In *Intelligent Signal Processing and Communications Systems (ISPACS), 2012 International Symposium on*, pages 781–786. IEEE, 2012.

[7] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1482–1489. IEEE, 2005.

[8] Hongliang Li, Fanman Meng, and King Ngi Ngan. Co-salient object detection from multiple images. *IEEE Transactions on Multimedia*, 15(8):1896–1909, 2013.

[9] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. Cluster-based co-saliency detection. *Image Processing, IEEE Transactions on*, 22(10):3766–3778, 2013.

[10] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Wei Feng. Self-adaptively weighted co-saliency detection via rank constraint. *IEEE Transactions on Image Processing*, 23(9):4175–4186, 2014.

[11] Zhengqin Li and Jiansheng Chen. Superpixel segmentation using linear spectral clustering. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1356–1363. IEEE, 2015.

[12] Junxia Li, Jundi Ding, and Jian Yang. Visual salience learning via low rank matrix recovery. In *Asian Conference on Computer Vision*, pages 112–127. Springer, 2014.