Estimating Visual Saliency Through Single Image Optimization

Jia Li, *Member, IEEE*, Yonghong Tian, *Senior Member, IEEE*, Lingyu Duan, and Tiejun Huang

***Abstract*—This letter presents a novel approach for visual saliency estimation through single image optimization. Instead of directly mapping visual features to saliency values with a unified model, we treat regional saliency values as the optimization objective on each single image. By using a quadratic programming framework, our approach can adaptively optimize the regional saliency values on each specific image to simultaneously meet multiple saliency hypotheses on visual rarity, center-bias and mutual correlation. Experimental results show that our approach can outperform 14 state-of-the-art approaches on a public image benchmark.**

***Index Terms*—Quadratic programming, single image optimization, visual saliency.**

I. Introduction

VISUAL saliency is an useful tool to locate the attractive visual signals in images and videos. By focusing on the salient contents, images and videos can be analyzed as human vision system does. Consequently, the analysis results often demonstrate better capabilities to meet human perception. To that end, visual saliency estimation is now becoming one of the hottest yet challenging research area in signal processing, computer vision and multimedia analysis.

To estimate visual saliency, many approaches have been proposed in the past few decades. Among them, a widely accepted hypothesis is that *visual rarity* can work as a good criterion to quantize saliency. For example, the most famous approach proposed by Itti *et al.* [1] tried to estimate image saliency by fusing the multi-scale center-surround contrasts extracted from multiple preattentive features. In this manner, unique or rare image contents can pop-out from their surroundings and become salient. Similarly, Goferman *et al.* [2] incorporated the influence of spatial contexts to compute visual saliency. Harel *et al.* [3] represented images as undirected graphs whose edges were weighted by pixel-wise differences. A random walker was then used to find and pop-out the less-visited nodes (i.e., pixels). Chen *et al.* [4] proposed a Bayesian framework to jointly integrate the traditional low-level cues and the defocus prior from photographers for image saliency estimation.

Beyond the approaches that mainly extracted visual cues from the spatial domain, some approaches also tried to estimate image saliency in the transform domain. For example, Hou and Zhang [5] adopted the Fourier transform to detect irregular visual signals from the frequency domain. Similarly, many approaches first learned a visual dictionary consisting of various basis functions and then project the input signals onto the subspaces represented by these functions for visual saliency estimation. A fundamental assumption here is that certain subspaces may have better capabilities to distinguish salient targets from background distractors. For example, Bruce *et al.* [6] first learned a set of basis functions using independent component analysis and then estimated visual saliency through information maximization. Wang *et al.* [7] projected image signals onto these subspaces and then adopted the graph representations to find the salient locations. Yan *et al.* [8] proposed to learn a over-complete dictionary to characterize image patches and visual saliency was then estimated through matrix decomposition.

Usually, all these approaches can achieve promising performance using predefined or learned models that can map the explicit visual cues (e.g., contrasts and entropy) to saliency values. However, one drawback of these approaches is that they often adopt unified models to process various images, while these saliency hypotheses (e.g., rarity, center-bias and correlation hypotheses) may not always hold in different images. For instance, using the rarity hypothesis can easily detect small salient objects but may fail when processing images with large salient objects. Therefore, various saliency hypotheses should be adaptively taken into account to obtain the best saliency map on each specific image.

Toward this end, we propose a novel approach to estimate visual saliency through single image optimization. Instead of mapping visual features to saliency values with a unified model, we treat the saliency values of all regions as the optimization objective on each single image. In this process, the saliency value of each region is optimized when considering the influences of all the other image regions. By using a quadratic programming framework, these saliency values can be adaptively optimized on each image to simultaneously meet several saliency hypotheses on visual rarity, center-bias and mutual correlation. Experimental results show that our approach can outperform 14 approaches on a public image benchmark.

The remainder of this letter is organized as follows. Section II describes the details of the proposed approach. In Section III,

Manuscript received April 28, 2013; revised May 28, 2013; accepted June 12, 2013. Date of publication June 14, 2013; date of current version June 28, 2013. This work was supported in part by grants from the Chinese National Natural Science Foundation under Contracts 61035001 and 61072095, and National Basic Research Program of China under Contract 2009CB320906. The associate editor coordinating the review of this manuscript and aproving it for publication was Prof. Ce Zhu. *(Corresponding Author: Y. Tian.)*

The authors are with the National Engineering Laboratory for Video Technology, School of Electrical Engineering & Computer Science, Peking University, Beijing, China (e-mail: jia.li@pku.edu.cn; yhtian@pku.edu.cn; lingyu@pku.edu.cn; tjhuang@pku.edu.cn).

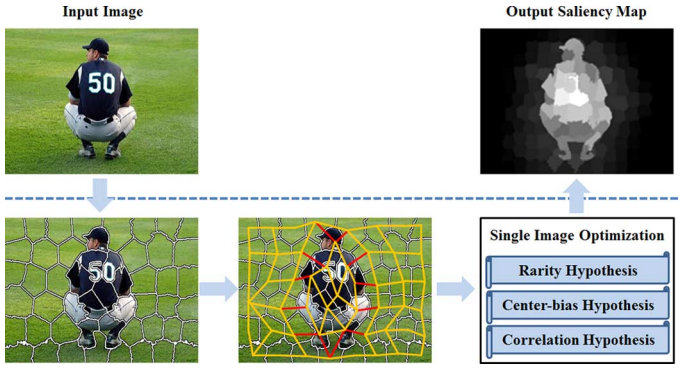

Fig. 1. The system framework of our approach. In our approach, the input image is first segmented into compact regions. After that, the correlations between regions are computed as the similarity between their visual appearance. Finally, the saliency values of all regions are simultaneously optimized to best meet the proposed hypotheses on visual rarity, center-bias and mutual correlation.

we conduct experiments to validate the effectiveness of our approach. Finally, the letter is concluded in Section IV.

II. Our Approach

To estimate visual saliency, we first segment an image $\mathcal{I}$ into N regions using the algorithm [9], denoted as $\{\mathcal{B}_i\}_{i=1}^N$. As shown in Fig. 1, these compact regions can well preserve object shapes and are much easier to obtain than the perfectly segmented objects. Moreover, such compact segmentation avoid the ambiguities around object boundaries which often arise when simply partitioning images into macro blocks with fixed sizes.

Given $\{\mathcal{B}_i\}_{i=1}^N$, visual saliency can be computed as a kind of *regional rarity*. That is, high saliency values should be assigned to unique or rare regions. To quantize such rarity, we have to derive the mutual correlations between all regions. Inspired by the idea that visually similar regions should have strong correlations, we represent $\mathcal{B}_i$ with its visual appearance $\mathbf{v}_i$, which is a column vector computed by averaging all the pixel-wise intensity, red/green opponency and blue-yellow opponency in $\mathcal{B}_i$. Here we use the approaches in [10] to compute the red-green opponency RG_υ and blue-yellow opponency BY_υ for pixel υ:

$$\begin{aligned} RG_\upsilon &= \frac{r_\upsilon - g_\upsilon}{\max(r_\upsilon, g_\upsilon, b_\upsilon)}, \\ BY_\upsilon &= \frac{b_\upsilon - \min(r_\upsilon, g_\upsilon)}{\max(r_\upsilon, g_\upsilon, b_\upsilon)}, \end{aligned} \tag{1}$$

where r_υ, g_υ, b_υ are the red, green and blue values of pixel υ. Similar to [10], we set $RG_\upsilon = BY_\upsilon = 0$ if $\max(r_\upsilon, g_\upsilon, b_\upsilon) < 0.1$ to avoid large fluctuations at low luminance.

After extracting $\{\mathbf{v}_i\}_{i=1}^N$, the correlation w_{ij} between $\mathcal{B}_i$ and $\mathcal{B}_j$ can be computed as a kind of visual similarity:

$$w_{ij} = \exp\left(-\frac{\|\mathbf{v}_i - \mathbf{v}_j\|_1}{3}\right). \tag{2}$$

From the definition in (2), we can see that the mutual correlation between any two image regions is symmetric (i.e., $w_{ij} = w_{ji}$). Note that the spatial distance between $\mathcal{B}_i$ and $\mathcal{B}_j$ is not considered in (2) to avoid the boundary effect. Usually, severe boundary effect may arise when incorporating the influence of such spatial distance. In this case, regions far from image centers will have relatively weak correlations with all the other regions. Consequently, high saliency values could be mistakenly assigned to the regions around image corners.

After computing the mutual correlations, we can now estimate the saliency value for each region. Instead of using a model that directly maps regional visual features to saliency values, we simultaneously optimize the saliency values of all regions in a specific image that can best meet several saliency hypotheses, including:

- **Rarity hypothesis**: an image region, which only has weak correlations (i.e., low visual similarities) with all the other regions, should probably be salient;
- **Center-bias hypothesis**: an image region, which appears near to image center, should probably be salient.
- **Correlation hypothesis**: two tightly correlated image regions should have similar saliency values if they are near to each other;

Following these hypotheses, we can formulate the problem of visual saliency estimation into an optimization framework. In this framework, saliency values of various regions can be optimized simultaneously to best meet the hypotheses. Let s_i be the saliency value of $\mathcal{B}_i$, we can optimize $\{s_i\}_{i=1}^N$ by solving:

$$\begin{aligned} \min_{\{s_i\}_{i=1}^N} & \sum_{i=1}^N s_i \sum_{j\neq i}^N w_{ij} + \lambda_c \sum_{i=1}^N s_i e^{d_i/d_D} \\ & + \lambda_r \sum_{i=1}^N \sum_{j\neq i}^N (s_i - s_j)^2 w_{ij} e^{-d_{ij}/d_D} \\ s.t.\ & 0 \le s_i \le 1, \quad \forall i, \\ & \sum_{i=1}^N s_i = 1. \end{aligned} \tag{3}$$

where d_D is half the image diagonal length. d_{ij} and d_i are the distances from $\mathcal{B}_i$ to $\mathcal{B}_j$ and image center, respectively. λ_c and λ_r are two weights to balance the influences of these terms, which can be automatically estimated as:

$$\lambda_c = \sqrt{N}, \quad \lambda_r = \log N. \tag{4}$$

From (3), we can see that high penalties will arise if: 1) assigning high saliency to a region that is tightly related to all the other regions(the first term); 2) allocating high saliency to regions near to image boundaries (the second term) or 3) assigning different saliency values to tightly related regions that are near to each other (the third term). Note that the optimization problem in (3) only has quadratic and linear terms with linear constraints, thus we can solve it with the active-set algorithm by iteratively searching for the active constraints and solving the equality problems with Lagrange multipliers.

After obtaining the saliency values for all image regions, another problem is how to extract the salient objects. As shown in Fig. 2, we adopt two intelligent thresholds to extract the salient

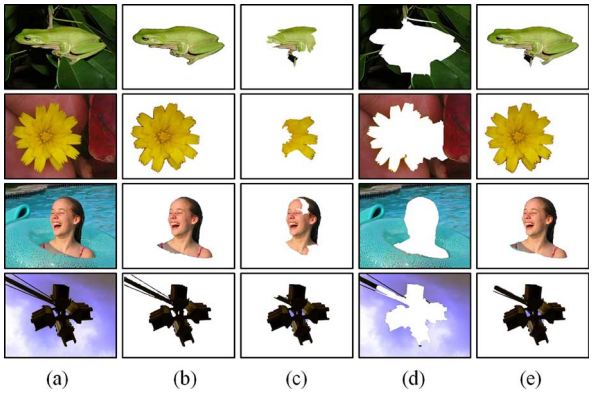

Fig. 2. Salient object extraction using two intelligent thresholds. These two thresholds are first calculated to selected reliable foreground and background regions, while other regions are then classified as foreground or background using the mutual correlations across regions. (a) images; (b) ground-truth salient objects; (c) reliable foreground regions; (d) reliable background regions; (e) extracted salient objects.

objects. First, we compute these two thresholds to select the most reliable foreground and background regions:

$$T_{low} = \frac{1}{N}\sum_{i=1}^{N} s_i, \quad T_{high} = \frac{2}{N}\sum_{i=1}^{N} s_i. \tag{5}$$

After that, regions with saliency values higher than T_{high} and lower than T_{low} are selected as reliable foreground and background regions, respectively. In particular, we directly select the top 5% regions as reliable foreground regions if $T_{high} > \max\{s_i\}_{i=1}^{N}$.

Given the reliable foreground and background regions, we then classify other regions according to their mutual correlations. Let $\mathbb{I}_{high}$ and $\mathbb{I}_{low}$ be the indices of the reliable foreground and background regions, we classify $\mathcal{B}_i$ with saliency value $T_{low} \leq s_i \leq T_{high}$ as a foreground region if:

$$\max\{w_{ij}, j \in \mathbb{I}_{high}\} > \max\{w_{ij}, j \in \mathbb{I}_{low}\}. \tag{6}$$

Otherwise, $\mathcal{B}_i$ will be classified as a background region. Different from [11] and [12] which directly binarize saliency maps using one threshold like T_{high}, we also select the reliable background regions using T_{low} and such background regions can help to recover the foreground regions whose saliency values are in $[T_{low}, T_{high}]$. In this manner, we can ensure that most salient regions can pop-out after the binarization, especially when the salient objects are very large (as shown in Fig. 2).

III. Experimental Results

In this section, we conduct several experiments to validate the effectiveness of our approach. In the experiments, we adopt the image saliency benchmark proposed by Achanta *et al.* [11]. This benchmark contains 1,000 images with obvious salient objects. In each image, the salient objects are manually labeled with accurate masks. This benchmark has been used by many approaches such as [11]–[15] for evaluating visual saliency models.

On this benchmark, our approach is compared with 14 state-of-the-art approaches. These approaches can be roughly categorized into three groups, including:

- **Spatial group**: This group contains 7 approaches that detect salient locations in the spatial domain, including CS [1], CA [2], GB [3], SR [5], RARE [16], RAND [14] and HC [13];
- **Dictionary group**: this group consists of 3 approaches that learn visual dictionaries to assist visual saliency estimation, including AIM [6], SER [7] and ICL [17];
- **Region group**: this group contains 4 approaches that segment images into regions for visual saliency estimation, including FT [11], RC [13], CSP [15] and SF [12].

In the comparison, all approaches are evaluated from two perspectives. First, we use the Area Under the ROC Curve (**AUC**) to evaluate the performance of estimating saliency values[1]. Second, we use **Recall**, **Precision** and **FScore** to evaluate the performance of extracting salient objects. In the comparison, all the other saliency maps are binarized using the intelligent thresholds proposed in [12] and **FScore** is calculated by equally consider **Recall** and **Precision**:

$$\mathbf{FScore} = \frac{2 \times \mathbf{Recall} \times \mathbf{Precision}}{\mathbf{Recall} + \mathbf{Precision}}. \tag{7}$$

When calculate **FScore**, we equally treat **Recall** and **Precision** instead of emphasizing only **Precision** as in [11]–[13]. The reason is that for some applications such as mobile search and video retargeting, **Recall** is as important as **Precision** since these applications often rely on the features extracted from object boundaries. When only emphasizing **Precision**, some object parts will be wrongly suppressed and the "fake" boundaries will mislead these applications.

The performance of these approaches are shown in Table I and some representative examples are illustrated in Fig. 3. From Table I, we can see that our **FScore** and **AUC** are always among the best. In particular, our approach achieves the highest **FScore** in extracting salient objects, while the performance **AUC** is comparable with RC and SF in estimating saliency values. Actually, the main advantage of our approach is that it can *simultaneously optimizes* the saliency values of all regions in a specific image. In the optimization, the saliency value of a specific region is computed when the influences of all the other regions are taken into account. In this manner, the estimated saliency map can adaptively meet all the three saliency hypotheses on visual rarity, center-bias and mutual correlation.

Moreover, we can see that the framework of our approach is flexible. Suppose that we have specific prior knowledge on the salient targets (e.g., the salient targets that may appear in the image), we can simply add new penalty terms in (3) to bring in such prior knowledge to improve the estimated saliency maps. Moreover, the estimated saliency maps have the same size as the input image. As shown in Fig. 3, the boundaries of the salient objects can be well maintained. This is an useful characteristic since many object analysis techniques need to extract features from boundaries.

[1]The code for calculating **AUC** can be found at http://jdl.ac.cn/user/jiali/calcAUCjudd.m.

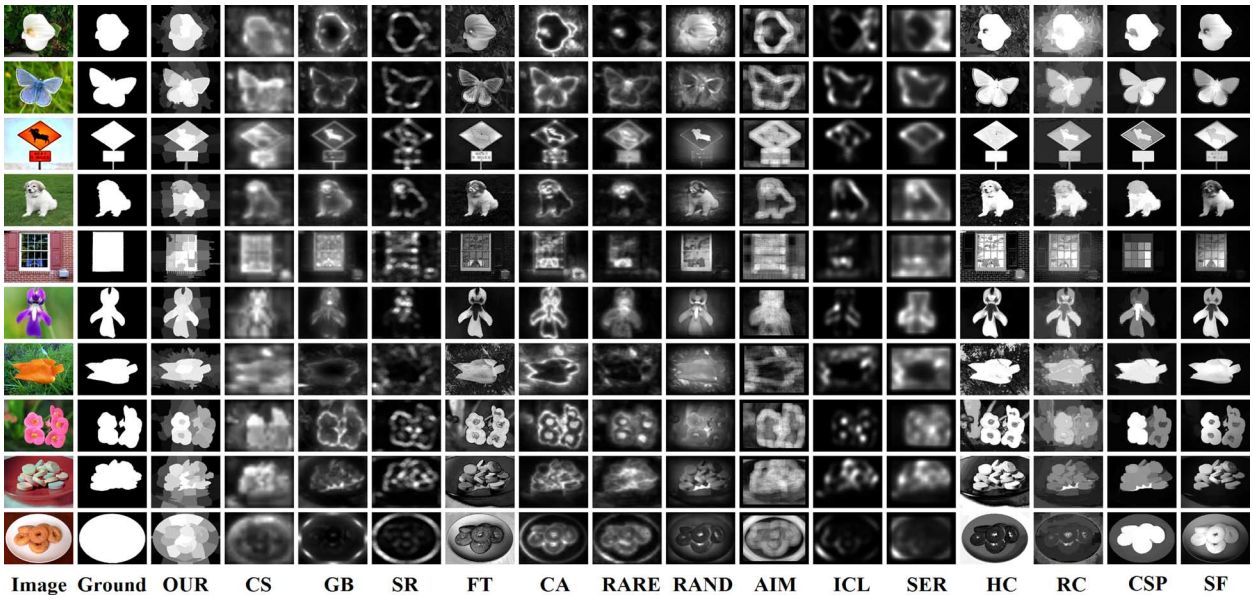

Fig. 3. Representative saliency maps of our approach and the other 14 approaches. Each saliency map is normalized into $[0, 255]$ to get an intuitive view.

TABLE I
COMPARISON BETWEEN OUR APPROACH AND OTHER 14 APPROACHES

Algorithm		Recall	Precision	FScore	AUC
Spatial Group	CS	0.29	0.69	0.41	0.81
	GB	0.38	0.67	0.49	0.80
	SR	0.36	0.50	0.42	0.72
	CA	0.47	0.62	0.54	0.81
	RARE	0.53	0.68	0.59	0.83
	RAND	0.46	0.70	0.56	0.83
	HC	0.66	0.75	0.70	0.84
Dictionary Group	AIM	0.37	0.53	0.44	0.78
	ICL	0.41	0.60	0.49	0.78
	SER	0.50	0.63	0.56	0.82
Region Group	FT	0.51	0.76	0.61	0.81
	RC	0.47	0.87	0.61	**0.87**
	CSP	0.76	0.89	0.82	0.85
	SF	0.71	**0.90**	0.79	**0.87**
	OUR	**0.81**	0.86	**0.83**	0.86

IV. CONCLUSION

In this letter, we propose an approach that can automatically pop-out the salient regions without predefining or learning models to map visual features to saliency values. Given the region-wise correlations, the saliency values of various regions can be adaptively optimized on each image to simultaneously meet multiple saliency hypotheses such as visual rarity, center-bias and mutual correlation. Experimental results show that our approach outperforms 14 state-of-the-art approaches in extracting the salient objects from images. In the future work, we will try to improve our approach by bringing in more saliency hypotheses. Moreover, we will also try to incorporate some task-dependent factors into the optimization framework to generate top-down saliency maps.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[3] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Neural Inf. Process. Syst.*, pp. 545–552, 2006.

[4] Z. Chen, J. Yuan, and Y.-P. Tan, "Hybrid saliency detection for images," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 95–98, 2013.

[5] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[6] N. D. Bruce and J. K. Tsotsos, "Saliency based on information maximization," *Neural Inf. Process. Syst.*, 2005.

[7] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Visual saliency detection by spatially weighted dissimilarity," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[8] R. Shi, Z. Liu, H. Du, X. Zhang, and L. Shen, "Region diversity maximization for salient object detection," *IEEE Signal Process. Lett.*, vol. 19, no. 4, pp. 215–218, 2012.

[9] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Patt. Analy. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[10] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.

[11] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

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

[13] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[14] T. N. Vikram, M. Tscherepanow, and B. Wrede, "A saliency map based on sampling an image into random rectangular regions of interest," *Pattern Recognit.*, pp. 3114–3124, 2012.

[15] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, "Automatic salient object segmentation based on context and shape prior," in *Proc. Brit. Machine Vision Conf.*, 2011, pp. 110.1–110.12.

[16] N. Riche, M. Mancas, B. Gosselin, and T. Dutoit, "Rare: A new bottom-up saliency model," in *IEEE Int. Conf. Image Processing*, 2012.

[17] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Neural Inf. Process. Syst.*, 2008.