Unsupervised Saliency Estimation based on Robust Hypotheses

Fei Xu, Min Xian, H. D. Cheng
Utah State University
{fei.xu, min.xian,hd.cheng}@aggiemail.usu.edu

Jianrui Ding, Yingtao Zhang
Harbin Institute of Technology
{jrding, yingtao_hit}@hit.edu.cn

Abstract

Visual saliency estimation based on optimization models is gaining increasing popularity recently. In this paper, we formulate saliency estimation as a quadratic program (QP) problem based on robust hypotheses. First, we propose an adaptive center-based bias hypothesis to replace the most common image center-based center-bias. It calculates the weighted center by utilizing local contrast which is much more robust when the objects are far away from the image center. Second, we model smoothness term on saliency statistics of each color. It forces the pixels with similar colors to have similar saliency statistics. The proposed smoothness term is more robust than the smoothness term based on region dissimilarity when the image has complicated background or low contrast. The primal-dual interior point method is applied to optimize the proposed QP in polynomial time. Extensive experiments demonstrate that the proposed method can outperform 10 state-of-art methods on three public benchmark datasets.

1. Introduction

Visual saliency estimation aims to locate objects in image which can capture human attention. It has earned popularity in many applications, such as image/video segmentation [1, 2], co-segmentation [3], and object recognition [4], etc.

One of the earliest saliency models proposed by Itti et al. [5] was based on the biological model proposed by Koch and Ullman [6]. The model generated an early bottom-up computational framework for saliency detection. Hou et al. [7] proposed the spectral residual method to detect salient objects based on Fourier Transform. The method calculates saliency values fast but is only sensitive to the boundary pixels of the salient objects. The phase spectrum based saliency detection is another frequency domain related method proposed by Guo et al. [8]. In order to address the problems of poor borders and low resolution map, Achanta et al. [9] proposed an efficient frequency-tuned approach to generate a full resolution saliency map. However, this method still cannot highlight full salient region.

In [10], Cheng et al. showed that the region-based global contrast method can achieve quite good saliency detection. They also demonstrated that the color space smoothing on the saliency map was an effective approach to reduce saliency noise. The method not only can just highlight object boundary, but also can highlight entire object regions. However, it failed when the objects had low contrast against background.

All the methods mentioned above are computational models: directly calculating saliency values based on image features: pixel or region color difference, spectral residual, phase information and gradient. The computational methods used the unified models to process various images and they will fail in many cases, such as big smooth objects, low image contrast, complicated background and not centered objects. Therefore, optimization models based visual saliency estimation method has attracted more and more attentions recently.

Chang et al. [11] modeled the saliency detection as a quadratic program (QP) optimization problem. They modeled three constraints in the objective function: saliency estimation term, objectness estimation term and the interaction term. Those constraints were heavily dependent on saliency priors and objectness priors; and the smooth term was only defined adjacent superpixels. Chen et al. [12] proposed an optimization method based on the low-rank matrix which modeled the center-bias prior, semantic prior and color prior in the objective function. The method treated the low rank problem as an approximation convex problem with equality constraints, which can only obtain local optimal. Jiang et al. [13] proposed a supervised method with regional contrast and background feature constraints. It used random forest regression to learn the saliency estimator based on richer descriptors. Kim et al. [14] proposed a supervised method based on random forest regression as well. They extracted high-dimensional features in color space and modeled the saliency detection as a linear program (LP) optimization problem. Li et al. [15] modeled the saliency detection as a QP problem. They modeled the region’s rarity, center-bias and regions’ correlation hypotheses. The method used the image center to model the center-bias hypotheses, which failed if the salient object is close to image borders; and the method generated inaccurate saliency map when the image has
complicated background. In [16] and [17], the saliency detection was formulated as a random walk problem in the absorbing Markov chain with boundary prior constraint.

Existing mathematical model based saliency estimation methods still suffered from some drawbacks. First, these methods formulated the center-bias constraint using the image center, which will make them have low accuracy when the objects are near the image borders. Second, many region-based methods were based on the assumption that every small region is homogeneous. So they model the smoothness term based on region features. Such limitation make these methods cannot achieve good results when the images have complicated background.

In this work, we formulate the visual saliency estimation as a quadratic program (QP) problem with robust constraints. We do not need to assume that the closer the image region and the image center are, the high saliency value the region may have. Alternatively, we use the newly proposed adaptive center to model the center-bias constraint. The adaptive center is calculated based on local contrast and can locate object automatically. We also do not assume that small regions generated by the superpixels method or other segmentation methods are homogeneous. Therefore, the smoothness term is not defined on region features but on the saliency statistics (sum) of each color.

The rest of the paper is organized as follows: the proposed unsupervised saliency estimation approach is presented in section 2; the experimental results are discussed in section 3; and we give the conclusions in section 4.

2. Proposed approach

2.1. Problem formulation

We model the saliency estimation problem as a convex optimization problem, and the optimization of the model is to assign optimal saliency values for a set of image region \( \{R_i\}_{i=1}^N \). To facilitate the discussion, we define \( S = (s_1, s_2, \cdots, s_N)^T \) as a vector of saliency values for \( N \) image regions, where \( s_i \) denotes the saliency value of the \( i \)-th image region and \( s_i \in [0, 1] \).

The problem is formulated as

\[
\begin{align*}
\text{minimize} & \quad E(S) = E^{\text{data}}(S) + \beta E^{\text{smooth}}(S) \\
\text{subject to} & \quad 0 \leq s_i \leq 1, i = 1, 2, \cdots, N \\
& \quad \sum_{i=1}^N s_i = 1 \\
& \quad B^T S = 0, B = (b_1, b_2, \cdots, b_N)^T, b_i = \{0,1\}
\end{align*}
\] (1)

where the data term \( E^{\text{data}} \) models the region-based global contrast and the adaptive center-bias, \( E^{\text{smooth}} \) models the color-based smoothing, and \( \beta \) is defined to balance the influence of the two terms; \( b_i \) is assigned to 1 if the \( i \)-th region is adjacent to image border, and 0 otherwise.

We applied the method in [18] to segment the image into \( N \) regions. Similar to the method in [10], we extract regions’ histograms in \( \text{lab} \) color space as the region features and only keep \( n \) bins for each histogram. In the data term, the newly proposed adaptive center-bias hypothesis and the global contrast-based hypothesis are modeled. It forces the image region with high global contrast value and/or short distance to the adaptive center to have high saliency value. In the smoothness term, we model the robust correlation hypothesis based on color saliency.

2.2. Data term

\[
E^{\text{data}}(S) = S^T (D + \alpha C)
\] (2)

In Eq. (2), \( E^{\text{data}}(S) \) is a linear function of \( S \); the terms \( D = (d_1, d_2, \cdots, d_N)^T \) and \( C = (c_1, c_2, \cdots, c_N)^T \) are two \( N \times 1 \) vectors that denote the global contrast and adaptive center-bias of image regions, respectively. Large value of \( d_i \) indicates low global contrast between the \( i \)-th region and all other image regions, and small \( d_i \) indicates high global contrast (Eqs. (3) - (5)). The term \( S^T D \) defines the cost on the global contrast, and \( S^T C \) define the cost based on the newly proposed adaptive center-bias. The relative importance between the two terms is specified by \( \alpha \) which will be discussed in section 3.4.

\[ D = \text{the global contrast vector defined by} \]
\[
d_i = \sum_{j=1}^N w_{ij} \]
\[
w_{ij} = \exp(-ds(H_i, H_j)) \] (3)

where \( w_{ij} \) is the color contrast between the \( i \)-th region and the \( j \)-th region, and \( H_i \) and \( H_j \) are the color histograms of region \( i \) and \( j \), respectively. \( ds(H_i, H_j) \) defines the dissimilarity between two regions using the Bhattacharyya distance and is given as

\[
ds(H_i, H_j) = \sqrt{1 - \sum_k \sqrt{H_i(k)H_j(k)}} \] (5)

where \( H_i(k) \) and \( H_j(k) \) are the values of the \( k \)-th bin of histogram \( H_i \) and \( H_j \), respectively.

The term \( C \) models the proposed adaptive center-bias. Most methods used the image center to calculate center-bias values and will fail when the objects are far from the image center. Here we propose a new method based on an adaptive center to compute the center-bias value. \( C \) is defined by

\[
C = \exp(||RC_i - AC||_2/d_D) \] (6)

\[
AC = \sum_{i=1}^M p_i w'(p_i)/\sum_{i=1}^M w'(p_i) \] (7)

where \( RC_i \) is the coordinate of the \( i \)-th region’s center, \( AC \) is adaptive center, \( || \|_2 \) is the \( L_2 \) norm, \( d_D \) is the diagonal distance of the image, \( p_i = (x_i, y_i) \) is the coordinates of the \( i \)-th image pixel, \( M \) is number of image pixels, and \( w'(p_i) \) is the local contrast value of the \( i \)-th pixel. The local contrast map is calculated by using the local range filter. We first
transform color image into gray image, then the local range output \( w' (p_i) \) is computed by using \( \max\{(g(p)|p \in NH_{3 \times 3}^i\}) - \min\{(g(p)|p \in NH_{3 \times 3}^i\}) \), where \( g(p) \) is the gray value of the \( p \)th pixel, \( NH_{3 \times 3}^i \) is a set contains all the pixels (including \( p_i \)) of the 3-by-3 neighborhood around \( p_i \).

As shown in Eq. (7), the adaptive center \( AC \) is the weighted center of the local contrast map, e.g. if the energy of the local contrast map concentrated in the left side of the image, the adaptive center will move to the left. To overcome the effect of local contrast noise, we use the mean value of the local contrast as a threshold (discussed in section 3.4), and the local contrast values less than the threshold will be set as 0.

### 2.3. Smoothness term

Most region-based methods [11, 12, 15] model the smoothness term based on the region-based correlation hypothesis: two similar regions should have similar saliency values if they are close to each other. They calculate the similarity between two regions using the regions’ mean color values. These methods can achieve good results on smooth images, but if the image regions are inhomogeneous, the mean values cannot describe the region features precisely and the performance will drop dramatically. In [10], Cheng et. al demonstrated that replacing saliency value of each color using the weighted average saliency values of similar colors can reduce large amount noise. Inspired by their work, we model the smoothness term on color saliency which makes similar colors have similar saliency similar color saliency statistics (sum).

\[
E_{\text{smooth}}(S) = (S - A \times S)^T (S - A \times S) \tag{8}
\]

In Eq. (8), \( E_{\text{smooth}}(S) \) is a quadratic function in variable \( S \), and \( A \) is an \( N \)-by-\( N \) transition matrix which is proposed to smooth the current saliency map.

\[
A = H \times V \times H^T
\]

\[
H = \begin{pmatrix}
H_1^T \\
H_2^T \\
\vdots \\
H_n^T
\end{pmatrix}, \quad H_i = (h_1, h_2, \ldots, h_n)^T \tag{9}
\]

In Eq. (9), \( H \) is an \( N \)-by-\( n \) matrix contain histograms of \( N \) regions; and \( V \) is an \( n \)-by-\( n \) matrix saved the \( L_2 \) distance between each color pairs (Lab color); only top 1/4 nearest color distances are kept, and the others are set to zero; \( H_i \) is column vector saves the values of the \( i \)th region’s color histogram; the maximum number of histogram bins \( n \) is 1728 (12 \times 12 \times 12), only valid bins (\( h_i \neq 0 \)) are consider in computation.

In order to smooth region saliency values based on color saliency, we introduce the transition matrix \( A \). \( A \times S \) will output the smoothed saliency values of each region based on color saliency.

The histogram matrix \( H^T \) realizes the mapping from region saliency values \( S \) to color saliency values \( (H^T \times S) \). The value of the \( i \)th element in \( H^T \times S \) is the sum of saliency values of all pixels with same color value \( i \).

The color distance matrix \( V \) works as a weighted mean filter. It replaces the saliency values of each color to the average saliency values of its \( m \)-nearest neighbors. The value of the \( i \)th element in \( V \times H^T \times S \) is the smoothed saliency value of color \( i \). \( V \times H^T \times S \) maps the smoothed saliency value of each color to region saliency value.

The proposed smoothness term is a quadratic function, and the formulated problem is a QP problem with linear equality and inequality constraints. The original problem can be rewritten as following:

\[
\begin{align*}
\text{minimize} & \quad f_0(S) = S^T (D + \alpha C) + \\
& \quad \beta (S - A \times S)^T (S - A \times S) \\
\text{subject to} & \quad 0 \leq s_i \leq 1, \, i = 1, 2, \ldots, N \\
& \quad \sum_{i=1}^{N} s_i = 1 \\
& \quad B^T S = 0, \, B = (b_1, b_2, \ldots, b_N), \, b_i = \{0,1\} 
\end{align*} \tag{10}
\]

### 3. A primal-dual interior point method for optimization

We formulate the visual saliency estimation as a QP problem with linear equality and inequality constraints, and the primal-dual method can be applied to optimize the problem globally. The inequality constraints can be rewritten as a set of functions:

\[
\begin{align*}
& \quad f_k(S) = -S_k \leq 0, \, k = 1, 2, \ldots, N \\
& \quad f_k(S) = S_{k-N} - 1 \leq 0, \, k = N + 1, N + 2, \ldots, 2N 
\end{align*} \tag{11}
\]

In Eq. (11), \( N \) is the number of image regions and \( S_k \) is the saliency value of the \( k \)th region. We write all inequality constraints in a matrix (Eq. (12)).

\[
F(S) = \begin{bmatrix}
-\mathbb{I}_{2N-1} \\
f_1(S) \\
f_2(S) \\
\vdots \\
f_{2N}(S)
\end{bmatrix} = \begin{bmatrix}
-S \\
S - 1
\end{bmatrix} \tag{12}
\]

And the derivative matrix of Eq. (12) is

\[
Df(S) = \begin{bmatrix}
\nabla f_1(S)^T \\
\nabla f_2(S)^T \\
\n\vdots \\
\n\nabla f_{2N}(S)^T
\end{bmatrix} = \begin{bmatrix}
-\mathbb{I} \\
\mathbb{1}
\end{bmatrix} \tag{13}
\]

where \( I \) is the identity matrix.

In the primal-dual interior method, we update the dual residual (\( r_d \)), primal residual (\( r_p \)), and the centrality residual (\( r_c \)) in each iteration, and stop the optimization processing when the sum of the \( L_2 \) norms is less than \( 10^{-6} \).
The dual residual is
\[
r_d = \nabla f_0(S)^T + Df(S)^T \lambda + n(1)O + n(2)B
\]
\[
= D + \alpha C + 2\beta(1-A)^T(S - AS) + \left[ -I \right] \lambda + n(1)O + n(2)B
\]
where vectors \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{2N})^T \) and \( n = (n_1, n_2)^T \) are the dual feasible parameters, and \( O \) is a \( 2N \)-by-1 vector and all the values are 1s.

The primal residual is
\[
r_p = \left[ O^T S - I \right]
\]
The centrality residual is
\[
r_c = -\text{diag}(\lambda) f(S) - (1/t)O
\]
where \( t \) is the step size and initialized as 1.

The partial derivatives of \( r_d, r_p \), and \( r_c \) with respect to variables \( S, \lambda \), and \( n \) are as follows:
\[
\frac{\partial r_d}{\partial S} = 2\beta(1-A)^T(1-A),
\]
\[
\frac{\partial r_c}{\partial S} = \text{diag}(\lambda) \times \left[ -I \right], \quad \frac{\partial r_p}{\partial S} = \left[ O^T \right]
\]
\[
\frac{\partial r_d}{\partial \lambda} = \left[ -I \right]^T \text{diag}(f(S)), \quad \frac{\partial r_p}{\partial \lambda} = 0_{2 \times 2N}
\]
\[
\frac{\partial r_d}{\partial n} = [O \ B], \quad \frac{\partial r_c}{\partial n} = 0_{2 \times 2N}, \quad \frac{\partial r_p}{\partial n} = 0_{2 \times 2N}
\]
In each iteration, we get the Newton step \((\Delta S, \Delta \lambda, \Delta n)\) by solving Eq. (20) using the partial derivatives in Eqs. (17) - (19).

\[
\begin{bmatrix}
\frac{\partial r_d}{\partial S} & \frac{\partial r_d}{\partial \lambda} & \frac{\partial r_d}{\partial n} \\
\frac{\partial r_c}{\partial S} & \frac{\partial r_c}{\partial \lambda} & \frac{\partial r_c}{\partial n} \\
\frac{\partial r_p}{\partial S} & \frac{\partial r_p}{\partial \lambda} & \frac{\partial r_p}{\partial n}
\end{bmatrix}
\begin{bmatrix}
\Delta S \\
\Delta \lambda \\
\Delta n
\end{bmatrix}
= -
\begin{bmatrix}
r_d \\
r_c \\
r_p
\end{bmatrix}
\]

The variables \( S, \lambda \), and \( n \) are updated using the following equations.
\[
S^{k+1} = S^k + t^k \times \Delta S, \lambda' = \lambda + t^k \times \Delta \lambda, \quad n' = n + t^k \times \Delta n
\]

In Eq. (21), \( t^k \) is the step size and updated by using the line search method in each iteration; \( t^0 \) and \( S^0 \) are initialized as 1 and \( \frac{1}{n}(1,1,\ldots,1)^T \), respectively.

In order to deal with the salient regions connected to image border, the saliency value of each boundary region is replaced by the average saliency value of its three most similar (Eq. (5)) region.

4. Experiments

In this section, we validate the performance of the newly proposed method on three public datasets: ASD [9], SED1 [22], ECSSD [26]. The ASD dataset contains 1000 images selected from MSRA [23] dataset, and the manually labeled boundaries of salient objects are used for ground truth; the SED1 dataset has 100 images and pixel-wise masks; and each image in SED1 only has one object; the ECSSD dataset has 1000 images and their corresponding user labeled ground truth.

We use the precision-recall (P-R) curve, F-measure and mean absolute error (MAE) to evaluate the overall performance of saliency detection method. The P-R curve shows the mean precision and recall rate of all saliency maps on a dataset. For each method, the P-R curve is calculated by segmenting saliency map with threshold range from 0 to 255, and computing the precision and recall rates by comparing the thresholding result with the ground truth. To get the average precision and recall rates, we use an adaptive thresholding method [9], which chooses two times the mean saliency value as the threshold. The F-measure [6, 10] and MAE [25] are defined as
\[
F_y = \frac{(1+y^2)\text{Precision-Recall}}{\text{Precision} + \text{Recall}}
\]
\[
\text{MAE} = \frac{1}{M} \sum_{i=1}^{M} | S(p_i) - G(p_i) |
\]
where \( y^2 \) is set to 0.3 as suggested in [9], \( p_i \) is the coordinates of the \( i \)-th image pixel, \( S(p_i) \) is the saliency value of the \( i \)-th pixel, and \( G \) is the binary ground truth.

We compare the proposed method with most recent 10 state-of-the-art methods, including LC [24], spectrum residual (SR) [7], phase spectrum (PS) [8], frequency-tuned (FT) [9], histogram-based contrast (HC) [7], region-based contrast (RC) [10], fusing generic objectness and visual saliency (SVO) [11], context-aware (CA) [19], low rank (LR) [12], estimating visual saliency (EST) [15]. We use the authors’ original implementations for all methods except PS.

In addition, we validate the effectiveness of the newly proposed adaptive center-based constraint and the color saliency statistics-based smooth term on the datasets.

4.1. Effectiveness of adaptive center constraint

Similar to the procedures in section 3.2, we compare the methods with image center-based constraint with our method with adaptive center constraint. In our methods, \( \alpha \) and \( \beta \) are set to \( 2\sqrt{N} \) and \( 2N \), respectively. As shown in Figure 1, the proposed method with adaptive center-based data term is more robust than other center-bias methods.

4.2. Effectiveness of the newly proposed smoothness term

We compare the saliency estimation results of the method using the newly proposed smooth term with that of the method using region correlation-based smooth term [15]. In this experiment, we first run our method with the
newly proposed smooth term on the two datasets; and then run our method again by replacing the smooth term with the regions correlation-based one. The parameters are the same with those in [15]. The example results of 15 images are shown Figure 2.

The results in Figure 2 demonstrate that the newly proposed color saliency-based smooth term is more robust than the region correlation-based smooth term: our method with the newly proposed smooth term can generate accurate saliency map; however, our method with the region-correlation smooth term generates high background saliency values especially when the background is complicate.

![Image](image1)

**Figure 1.** Example results of adaptive center effectiveness. FC is fixed center. AC is adaptive center. From the last three columns we can see using the adaptive center can make the high saliency more concentrated on the objects than the fixed center.

![Image](image2)

**Figure 2.** The saliency maps of regions’ correlation smoothness term and our smoothness term. (a) the original images (b) the results of using regions’ correlation term (c) the results of using our smoothness term.

### 4.3. Parameter tuning

On SED1 dataset, we evaluate the performance of the proposed method with five thresholds: 0.3, 0.5, mean value, median value, and the tenth maximum value. The proposed method can generate better results if we use the mean or median value as the threshold. Therefore, we set the mean value of the local contrast map as the threshold in all our experiments. Parameter $\alpha$ is used to balance the influence of the adaptive center-based term ($S^T C$) and the global contrast based term ($S^D$). Parameter $\beta$ is applied to balance influence of data term and the smoothness term. On ASD dataset, we evaluate the performance of the proposed method with $\alpha$ and $\beta$ range from 0 to 10N. The step size is $0.2\sqrt{N}$, and the step size of $\beta$ is 0.4N. As shown in Figure 5, we obtain better P-R curve when $\alpha$ is equal to 1.8$\sqrt{N}$ and $\beta$ is equal to 1.2$\sqrt{N}$, and we will choose the two values in all experiments.

### 4.4. Comparison with state-of-the-art

In this section, we compare the overall performance of the proposed approach with 10 state-of-the art methods (SR [4], PS [5], FT [6], HC [7], RC [7], CA [10], LRS [11], EST [15], and LC [24]), on ECSSD, ASD and SED1 datasets. As shown in Figures 4 - 6, the proposed method outperforms 9 methods (SR, PS, FT, HC, RC, CA, LRS, EST and LC) on the three datasets; our MAE values, the F-measure values and P-R curve are better than all 10 methods on ECSSD and SED1 datasets.

### 5. Conclusion

In this paper, we propose a novel unsupervised saliency estimation method. The experimental results demonstrate that our method with the newly proposed adaptive center-based center-bias and the color saliency-based smoothness hypotheses is more accurate and robust than 10 state-of-the-art methods on three datasets. Furthermore, the proposed method can generate accurate and centralized saliency map, even when the salient object is far away from image center, image has complicate background and low contrast. In the future, we will generalize the proposed method for visual saliency estimation on images with multi-objects.

### References


