Adaptive Partial Differential Equation Learning for Visual Saliency Detection

Risheng Liu[†], **Junjie Cao**[†], **Zhouchen Lin** ✉[♯] and **Shiguang Shan**[‡]
[†]Dalian University of Technology
[♯]Key Lab. of Machine Perception (MOE), Peking University (zlin@pku.edu.cn)
[‡]Key Lab. of Intelligent Information Processing of Chinese Academy of Sciences (CAS)

Abstract

Partial Differential Equations (PDEs) have been successful in solving many low-level vision tasks. However, it is a challenging task to directly utilize PDEs for visual saliency detection due to the difficulty in incorporating human perception and high-level priors to a PDE system. Instead of designing PDEs with fixed formulation and boundary condition, this paper proposes a novel framework for adaptively learning a PDE system from an image for visual saliency detection. We assume that the saliency of image elements can be carried out from the relevances to the saliency seeds (i.e., the most representative salient elements). In this view, a general Linear Elliptic System with Dirichlet boundary (LESD) is introduced to model the diffusion from seeds to other relevant points. For a given image, we first learn a guidance map to fuse human prior knowledge to the diffusion system. Then by optimizing a discrete submodular function constrained with this LESD and a uniform matroid, the saliency seeds (i.e., boundary conditions) can be learnt for this image, thus achieving an optimal PDE system to model the evolution of visual saliency. Experimental results on various challenging image sets show the superiority of our proposed learning-based PDEs for visual saliency detection.

1. Introduction

As an important component for many computer vision problems (e.g., image editing [9], segmentation [18], compression [12], object detection and recognition [32]), saliency detection gains much attention in recent years and numerous saliency detectors have been proposed in the literature. According to their mechanisms of representing image saliency, existing work can be roughly divided into two categories: bottom-up and top-down approaches. The bottom-up methods [13, 7, 38, 36, 39, 34, 22, 15] are data-driven and focus more on detecting saliency from image features, such as contrast, location and texture. As one of the earliest work, Itti et al. [13] consider local contrast and define image saliency using center-surround differences of image features. Cheng et al. [7] also investigate the global contrast prior. Location is another important prior for modeling salient regions. The convex hull of interest points is employed in [38] to estimate the foreground location. The work in [39, 36] considers the image boundary as a background prior. Inspired by recent advances in machine learning, compressive sensing [34, 22] and operations research [15] are also utilized to detect salient image features. The work in [34, 22] assumes that a natural image can always be decomposed into a distinctive salient foreground and a homogenous background. So one can utilize low-rank and sparse matrix decomposition methods and their extensions for saliency detection. Very recently, Jiang et al. [15] formulate saliency detection as a semi-supervised clustering problem and use the well-studied facility location model to extract cluster centers for salient regions.

In contrast, the top-down approaches [26, 40] are often task-driven and incorporate more human perceptions for saliency detection. For example, Liu et al. [26] propose a supervised approach to learn to detect a salient region in an image. Yang et al. [40] use dictionary learning to extract region features and CRF to generate a saliency map.

In the past decades, Partial Differential Equations (PDEs) have shown their power of solving many low-level computer vision problems, such as restoration, smoothing, inpainting, and multiscale representation (see [5] for a brief review). This is mainly because theoretical analysis on these problems has already been accomplished in areas such as mathematical physics and biological vision. For example, scale space theory [23] proves that the multiscale representation of images are indeed solutions of heat equation with different time parameters.

Unfortunately, the existing PDE designing methodology (i.e., *define PDE with fixed formulation and boundary condition from general intuitive considerations*) is not suitable for complex vision tasks, such as visual saliency detection. This is because saliency is a kind of intrinsic information contained in the image and its description strongly depends on human perception. From the bottom-up view (i.e., lo-

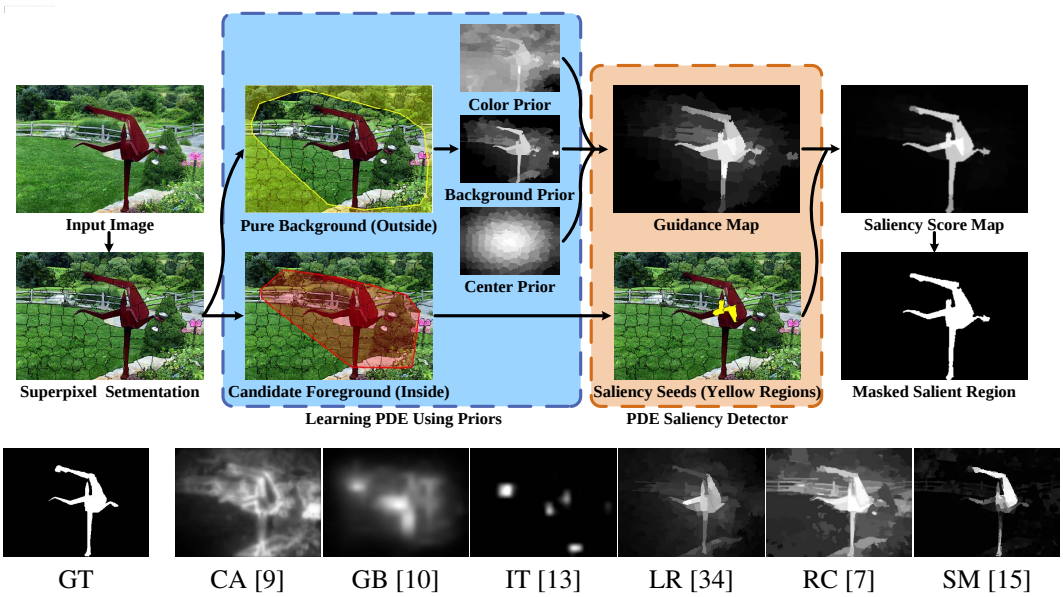


Figure 1. The pipeline of our learning-based LESD for saliency detection on an example image. The orange region illustrates the core components (i.e., guidance map and saliency seeds) of our PDE saliency detector, which will be formally introduced in Section 2. The blue region shows how to incorporate both bottom-up and top-down prior knowledge into our PDE system. The details of this PDE learning process will be presented in Section 3. The bottom row shows the ground truth (GT for short) salient region and saliency maps computed by some state-of-the-art saliency detection methods.

cal image structure), it is challenging to exactly define a PDE system with fixed formulation and boundary conditions to describe all types of saliency due to the complexity of salient regions in real world images. From the top-down view (i.e., object-level structure), high-level human perceptions (e.g., color [34], center [31], and semantic information [16]) are important for saliency detection. But it is hard to automatically incorporate these priors into conventional PDEs. Moreover, the boundary conditions in most existing PDE systems are simply defined by some general understandings on the problem (e.g., well-posed guarantees [5] and initial values [23]), thus cannot handle complex (e.g., driven by both data and priors) vision tasks. Overall, traditional PDEs with fixed form and boundary conditions cannot efficiently describe complex visual saliency patterns quantitatively, thus may fail to solve the saliency detection problem.

1.1. Paper Contributions

In this paper, we provide a diffusion viewpoint to understand the mechanism and investigate the physical nature of saliency detection. Firstly, an adaptive PDE system, named Linear Elliptic System with Dirichlet boundary (LESD), is proposed to describe the saliency diffusion. Then we develop efficient techniques to incorporate both bottom-up and top-down information into saliency diffusion and learn the *specific formulation* and *boundary condition* of LESD from the given image,. Fig. 1 shows the pipeline of our learning-based PDE detector with comparisons on an example image. To our best knowledge, this is the first work that incorporates learning strategy into PDE technique for visual saliency detection. We summarize the contributions of this paper as follows:

- A novel PDE system is learnt to describe the evolution of visual attention in saliency diffusion. We prove that visual attention in our system is a monotone submodular function with respect to saliency seeds.
- We develop an efficient method to incorporate both bottom-up and top-down prior knowledge into the LESD formulation for saliency diffusion.
- We derive a discrete optimization model with PDE and matroid constrains to extract saliency seeds for LESD. By further proving the submodularity of the proposed model, the performance can be guaranteed.

1.2. Notations

Hereafter, we use lowercase bold letters (e.g., $\mathbf{p}$) to represent vector points and capital calligraphic ones (e.g., $\mathcal{S}$) to denote sets of points. $|\mathcal{S}|$ is the cardinality of $\mathcal{S}$. $\mathbf{1}$ is the all one vector. We denote the neighborhood set of $\mathbf{p}$ on a graph as $\mathcal{N}_{\mathbf{p}}$. $\|\cdot\|$ denotes the ℓ_2 norm. Suppose f is a real-value function on $\mathcal{V}$. For a given point $\mathbf{p}$ with neighbor $\mathcal{N}_{\mathbf{p}}$, we denote ∇f as the gradient of f and discretize it as $\nabla f = [f(\mathbf{p}) - f(\mathbf{q}_1), \cdots, f(\mathbf{p}) - f(\mathbf{q}_{|\mathcal{N}_{\mathbf{p}}|})]$. Similarly, let $\mathbf{v}$ be a vector field on $\mathcal{V}$ and denote $\mathbf{v}_{\mathbf{p}}$ as the vector at $\mathbf{p}$.

We denote the divergence of $\mathbf{v}$ as $\mathrm{div}(\mathbf{v})$ and discretize it at $\mathbf{p}$ as $\mathrm{div}(\mathbf{v_p}) = \frac{1}{2}\sum_{\mathbf{q}\in\mathcal{N}_\mathbf{p}}(\mathbf{v_p}(\mathbf{q}) - \mathbf{v_q}(\mathbf{p}))$, where $\mathbf{v_p}(\mathbf{q})$ is the vector element corresponding to $\mathbf{q}$[1].

2. Saliency Diffusion Using PDE System

This section proposes a diffusion viewpoint to understand visual saliency and establishes a PDE system to model saliency diffusion on an image. Numerical and theoretical analysis on our system is also presented accordingly.

2.1. Visual Attention Evolution

For a given visual scene, saliency detection is to find the regions which are most likely to capture human's attention. This paper tackles this task from a diffusion point of view. That is, we assume that our attention is firstly attracted by the most representative salient image elements (this paper names them as *saliency seeds*) and then the *visual attention* will be propagated to all salient regions.

Specifically, let $\mathcal{V}$ be the discrete image domain, i.e., a set of points corresponding to all image elements (e.g., pixels or superpixels). Then we define a real-value visual attention score function $f(\mathbf{p}) : \mathcal{V} \to \mathbb{R}$ to measure the saliency of $\mathbf{p} \in \mathcal{V}$. Suppose we have known a set of saliency seeds (denoted as $\mathcal{S}$) and its corresponding scores (i.e., $f(\mathbf{p}) = s_\mathbf{p}$ for $\mathbf{p} \in \mathcal{S}$). We can mathematically formulate saliency diffusion as an evolutionary PDE with Dirichelet boundary condition:

$$\frac{\partial f(\mathbf{p},t)}{\partial t} = F(f, \nabla f),\ f(\mathbf{g}) = 0,\ f(\mathbf{p}) = s_\mathbf{p},\ \mathbf{p} \in \mathcal{S},$$

where $\mathbf{g}$ is an environment point with 0 score (outside $\mathcal{V}$) and F is a function of f and ∇f.

As the purpose of above PDE is to propagate visual attention from saliency seeds to other image elements, we adopt a linear diffusion term $\mathrm{div}(\mathbf{K_p}\nabla f(\mathbf{p}))$ for the score function, in which $\mathbf{K_p}$ is an inhomogeneous metric tensor to control the local diffusivity at $\mathbf{p}$. To incorporate our perception and/or high-level prior into the diffusion process, we further introduce a regularization term which is formulated as the difference between $f(\mathbf{p})$ and a guidance map $g(\mathbf{p})$ (will be discussed in Section 3), leading to the following form:

$$F(f, \nabla f) = \mathrm{div}(\mathbf{K_p}\nabla f(\mathbf{p})) + \lambda(f(\mathbf{p}) - g(\mathbf{p})),$$

where $\lambda \geq 0$ is a balance parameter.

2.2. Linear Elliptic System with Dirichlet Boundary

For saliency detection purpose, we only consider the situation when the saliency evolution is stable (i.e., no saliency attention can be further propagated). At this state, we omit the time t in our notation and only seek the solution to the following PDE:

$$F(f, \nabla f) = 0,\ f(\mathbf{g}) = 0,\ f(\mathbf{p}) = s_\mathbf{p},\ \mathbf{p} \in \mathcal{S}, \quad (1)$$

which is a Linear Elliptic System with Dirichlet boundary (LESD). Thus given an image, the saliency detection task reduces to the problem of solving an LESD.

Till now, we have established a general PDE system for saliency diffusion. Fig. 1 shows that our LESD (with properly learnt g and $\mathcal{S}$) can successfully incorporate image structure and high-level knowledge to model the saliency diffusion, thus achieves better saliency detection results than state-of-the-art approaches. Therefore, the main problem left for LESD is to develop an efficient learning framework to incorporate bottom-up image structure information and top-down human prior knowledge into (1). Before discussing this issue in Section 3, we first provide necessary numerical and theoretical analysis on LESD, which will significantly reduce the complexity of the learning process.

2.3. Discretization

Suppose $\mathcal{N}_\mathbf{p} = \{\mathbf{q}_1, \cdots, \mathbf{q}_{|\mathcal{N}_\mathbf{p}|-1}, \mathbf{g}\}$ is the neighborhood set of $\mathbf{p}$. Here the first $|\mathcal{N}_\mathbf{p}| - 1$ nodes are in the image domain $\mathcal{V}$ and will be specified in Section 3. The environment point $\mathbf{g}$ is connected to each node [37]. To measure the variance between $\mathbf{p}$ and its neighborhood $\mathcal{N}_\mathbf{p}$, we define an inhomogeneous metric tensor $\mathbf{K_p}$ as the following diagonal matrix[2]:

$$\mathbf{K_p} = \mathrm{diag}(k(\mathbf{p}, \mathbf{q}_1), \cdots, k(\mathbf{p}, \mathbf{q}_{|\mathcal{N}_\mathbf{p}|-1}), z_\mathbf{g}), \quad (2)$$

where $k(\mathbf{p}, \mathbf{q}) = \exp(-\beta\|h(\mathbf{p}) - h(\mathbf{q})\|^2)$ is the Gaussian similarity (with a strength parameter β) between the features of nodes, $h(\mathbf{p})$ is a feature vector at node $\mathbf{p}$, and $z_\mathbf{g}$ is a small constant to measure the dissipation conductance at $\mathbf{p}$. Then we can approximately discretize the LESD formulation as

$$f(\mathbf{p}) = \frac{1}{d_\mathbf{p} + \lambda}\Big(\sum_{\mathbf{q}\in\mathcal{N}_\mathbf{p}} \mathbf{K_p}(\mathbf{q})f(\mathbf{q}) + \lambda g(\mathbf{p})\Big), \quad (3)$$

where $\mathbf{K_p}(\mathbf{q})$ is the diagonal element of $\mathbf{K_p}$ corresponding to $\mathbf{q}$ and $d_\mathbf{p} = \sum_{\mathbf{q}\in\mathcal{N}_\mathbf{p}} \mathbf{K_p}(\mathbf{q})$. Based on this discrete scheme, our LESD can be reformulated as a linear system, thus can be easily solved.

2.4. Theoretical Analysis

It should be emphasized that the visual attention score f is indeed a *set function* on $\mathcal{V}$, i.e., $f(\mathcal{S}) : 2^\mathcal{V} \to \mathbb{R}$ as f

[1] Similar discretization scheme is also used for nonlocal total variation image processing [8].

[2] By anisotropic diffusion theory [37], $\mathbf{K_p}$ can also be chosen as a more general symmetric semi-positive definite matrix, which may lead to a more complex discretization scheme.

is the solution to (1) with respect to the saliency seed set $\mathcal{S}$. This implies that the solution to our LESD is inherently combinatorial, thus much more difficult to be handled than the PDEs in conventional low-level computer vision[3]. This is because the optimization of a combinatorial f without knowing any further properties can be extremely difficult (e.g., trivially worse-case exponential time and moreover inapproximable [21]). Fortunately, by proving the following theorem we can exploit some good properties, such as monotonicity (i.e., non-decreasing) and submodularity, of the solution to LESD. As shown in Section 3, these results provide good guarantees for our saliency detector.

Theorem 1 [4] *Let $f(\mathbf{p};\mathcal{S})$ be the visual attention score of image element $\mathbf{p}$. Suppose the sources $\{s_{\mathbf{p}} \geq 0\}$ are attached to saliency seed set $\mathcal{S}$, i.e., $f(\mathbf{p}) = s_{\mathbf{p}}$ for all $\mathbf{p} \in \mathcal{S}$. Then f is a monotone submodular function with respect to $\mathcal{S} \subset \mathcal{V}$.*

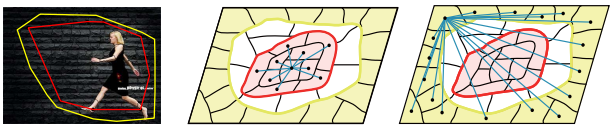
3. Learning LESD for Saliency Detection

This section discusses how to adaptively learn a specific LESD for saliency diffusion on a given image. For the given image, we first construct an undirected graph in the image feature space to model the neighborhood connections among image elements. Then we incorporate different types of human priors to establish the diffusion formulation (i.e., guidance map g). Based on the submodularity of the system, we also provide a discrete optimization model for boundary condition (i.e., saliency seeds $\mathcal{S}$) learning.

3.1. Feature Extraction and Graph Construction

For a given image, we generate superpixels to build the image elements set $\mathcal{V} = \{\mathbf{p}_1, \cdots, \mathbf{p}_{|\mathcal{V}|}\}$. Here any edge-preserving superpixel methods can be used and SLIC algorithm [3] is adopted in this paper. Then we define feature vectors $\{h(\mathbf{p}), \mathbf{p} \in \mathcal{V}\}$ as the means of the superpixels in the CIE LAB color space.

The image structure information is extracted as follows. Suppose the image domain $\mathcal{V}$ consists of two parts: the candidate foreground $\mathcal{F}_c$ (salient regions, may also contain some promiscuous image elements) and the pure background $\mathcal{B}_c$ (non-salient regions). We utilize a shift convex hull strategy to approximately estimate these two subsets from the input image. Specifically, we use Harris operator [35] to roughly detect the corners and contour points and estimate a convex hull $\mathcal{C}$ based on these points [38]. Then $\mathcal{F}_c$ can be obtained by collecting nodes *inside* $\mathcal{C}$. To further identify pure background nodes, we define an expended hull $\mathcal{C}'$ by adding adjacent nodes to $\mathcal{C}$. Then $\mathcal{B}_c$ is obtained by collecting all nodes *outside* $\mathcal{C}'$. Please see Fig. 2 (a) for an example of $\mathcal{C}$ and $\mathcal{C}'$.



(a) (b) (c)

Figure 2. Illustration of the shift convex hull strategy in (a) and connection relationship in (b)-(c). The red and yellow polygons in (a) denote $\mathcal{C}$ and $\mathcal{C}'$, respectively. The red and yellow regions in (b)-(c) represent $\mathcal{F}_c$ and $\mathcal{B}_c$, respectively. Lines in (c) indicate that all nodes in $\mathcal{B}_c$ are connected to each other.

Now we construct an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to reveal the connection relationships (i.e., $\mathcal{N}_{\mathbf{p}}$ for each $\mathbf{p}$) in the image domain, where $\mathcal{E}$ is a set of undirected edges corresponding to the nodes set $\mathcal{V}$[5]. We first define a k-regular graph structure to exploit local spatial relationship (Fig. 2 (b)). Then all the nodes in $\mathcal{B}_c$ are connected to each other to enforce the smoothness of background (Fig. 2 (c)). As there may exist promiscuous image elements, we do not further connect nodes in $\mathcal{F}_c$. Finally, all the nodes are connected to an environment node $\mathbf{g}$.

3.2. Learning Guidance Map Using Priors

This subsection shows how to incorporate different types of prior knowledge into the PDE system. For a given image, we first define a background diffusion to estimate the background prior. That is, we assume that the distribution of background is significantly different from that of foreground. Thus we perform a simplified LESD with $\lambda = 0$ to compute a background diffusion score f_b, i.e.,

$$\mathrm{div}(\mathbf{K}_{\mathbf{p}} \nabla f_b(\mathbf{p})) = 0,\ s.t.\ f_b(\mathbf{g}) = 0,\ f_b(\mathbf{p}) = 1,\ \mathbf{p} \in \mathcal{B}_c.$$

Here the boundary condition is defined by considering $\mathcal{B}_c$ as the background seed set with score 1 and adding an environment point $\mathbf{g}$ with score 0. It is easy to check that the solution to the background diffusion is a harmonic function, thus $f_b(\mathbf{p}) \in [0, 1]$[6]. So the elements in f_b can be viewed as probabilities of nodes belonging to the background. In this view, we have the probability of a node belonging to the foreground as $f_f(\mathbf{p}) = \mathbf{1} - f_b(\mathbf{p})$. By further incorporating high level prior knowledge (e.g., the color prior map f_c and the center prior map f_l[7], we define guidance map $g(\mathbf{p})$ as

$$g(\mathbf{p}) = f_f(\mathbf{p}) \times f_c(\mathbf{p}) \times f_l(\mathbf{p}), \tag{4}$$

and its value is normalized. To provide good boundary conditions for LESD, we also use g to define the scores of saliency seeds, i.e., $s_{\mathbf{p}} = g(\mathbf{p})$, for $\mathbf{p} \in \mathcal{S}$.

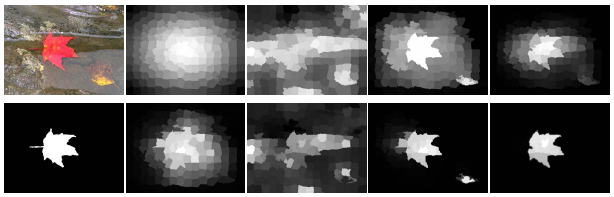
[3] In general, the solutions to PDEs with fixed formulation and boundary condition are continuous functions of space and/or time variables only, thus they are much easier to be handled.

[4] See supplemental materials for all proofs in this paper.

[5] As discussed in Section 2.3, the discretization of LESD is based on this connection relationship.

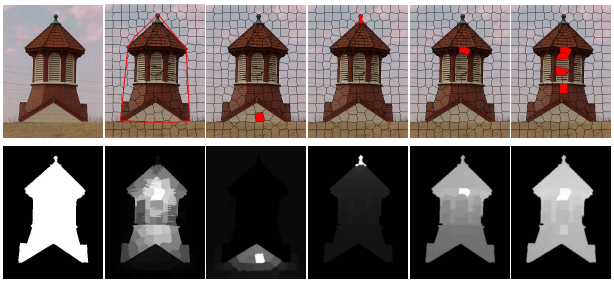
[6] Based on the maximum/minimum principles of harmonic functions.

[7] Please refer to [34] for detailed analysis on these two prior maps.



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

Figure 3. Saliency diffusion with different guidance maps. (a) input image and GT salient region. (b)-(e) center prior f_l, color prior f_c, background diffusion prior f_f, final guidance map g (top) and their corresponding saliency maps (bottom), respectively.



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

Figure 4. Saliency diffusion with different seeds. (a) input image and GT salient region. (b) $\mathcal{F}_c$ (inside red polygon) and g. (c)-(e) diffusion results using one candidate seed in $\mathcal{F}_c$: (c) background ($L = 10.6175$), (d) bad foreground ($L = 1.6818$) and (e) good foreground ($L = 31.7404$). (f) optimal seeds ($L = 43.8589$) and final saliency map. Here we report L values using the original saliency maps but normalize them for visual comparison.

3.3. Optimizing Saliency Seeds via Submodularity

Due to the following two reasons, we cannot choose all nodes in $\mathcal{F}_c$ as seeds for saliency diffusion. First, the convex hull may not adequately suppress background nodes in $\mathcal{F}_c$ (Fig. 4 (c)). Second and more importantly, it is observed that the seed with extremely high local contrast to its neighbors (e.g., nodes near object boundary and bright or dark nodes on the object) may also lead to a bad saliency map (Fig. 4 (d)). Therefore, it is necessary to search for *the most representative foreground nodes* in $\mathcal{F}_c$ to define boundary conditions for LESD. Note that the goal of LESD is to propagate the visual attention scores of seeds $\mathcal{S}$ to the whole image domain $\mathcal{V}$. So we would like to maximize the sum of scores f with respect to all image elements in $\mathcal{V}$ when the saliency diffusion is stable, that is, we solve the following discrete optimization problem:

$$\max_{\mathcal{S}\in\mathbb{M}^n} L(\mathcal{S}), \quad s.t. \begin{cases} f(\mathbf{p}) = \frac{1}{d_\mathbf{p}+\lambda}\left(\sum_{\mathbf{q}\in\mathcal{N}(\mathbf{p})} \mathbf{K}_\mathbf{p}(\mathbf{q}) f(\mathbf{q}) + \lambda g(\mathbf{p})\right), \\ f(\mathbf{g}) = 0,\ f(\mathbf{p}) = s_\mathbf{p},\ \mathbf{p}\in\mathcal{S}, \end{cases} \tag{5}$$

where $L(\mathcal{S}) = \sum_{\mathbf{p}\in\mathcal{V}} f(\mathbf{p};\mathcal{S})$ and $\mathbb{M}^n = \{\mathcal{S}|\mathcal{S} \subset \mathcal{F}_c, |\mathcal{S}| \le n\}$ is a uniform matroid [4] to enforce that the cardinality of $\mathcal{S}$ is no more than n. As visual attention scores can be considered as the relevances between nodes and the seeds, the above maximum criterion naturally tends to choose seeds in relatively larger connected subgraph (thus is more representative). Therefore, the nodes in $\mathcal{F}_c$ with high local contrast (i.e., less connections and paths to other nodes) will be removed from $\mathcal{S}$. One may concern that background nodes will also have a large L as they may connect to nodes outside $\mathcal{F}_c$. Fortunately, by learning a proper guidance map g, we can enforce very small saliency scores (in most case near zero) in background regions (g in Fig. 4 (b)). So background nodes in $\mathcal{F}_c$ still have a relatively small L value and cannot be included in $\mathcal{S}$ (Fig. 4 (c)).

In general, the performance of (5) is dependent on the maximum number of saliency seeds n (Fig. 5 (a)). Here we provide an adaptive way to identify n and further suppress background nodes in $\mathcal{F}_c$. We first define a background confidence function $w(\mathbf{p}) = 1/(1+g(\mathbf{p})^2)$ on $\mathcal{F}_c$, in which larger $w(\mathbf{p})$ implies that $\mathbf{p}$ has a higher probability of belonging to the background and should be suppressed. Therefore, we maximize another cost function $\hat{L}(\mathcal{S}) = L(\mathcal{S}) - \sum_{\mathbf{p}\in\mathcal{S}} w(\mathbf{p})$ in (5). Based on Theorem 1, we can prove the following corollary for L and $\hat{L}$.

Corollary 2 *Both $L(\mathcal{S})$ and $\hat{L}(\mathcal{S})$ are submodular functions. Furthermore, $L(\mathcal{S})$ is monotone with respect to $\mathcal{S}$.*

The monotonicity and submodularity of L together with the uniform matroid constraint in (5) imply that using a greedy algorithm to solve (5) yields a $(1-1/e)$-approximation [29]. Due to the non-monotone nature, we cannot have the same theoretical guarantee for $\hat{L}$. But in practice, by adding the stopping criterion $\hat{L}(\mathcal{S}\cup\{\mathbf{p}\}) \le \hat{L}(\mathcal{S})$, the maximization process for $\hat{L}$ can be automatically stopped and then the optimal seed set is obtained accordingly. We have experimentally found that a greedy algorithm with this stopping criterion is efficient for maximizing $\hat{L}$ in our saliency detector.

At the end of this section, we summarize the details for the learning-based LESD in Algorithm 1. The complete pipeline of our saliency detector on a test image is also illustrated in Fig. 1.

4. Discussions

In this section, we would like to discuss and highlight some aspects of our proposed PDE-based saliency detector.

4.1. Comparison to Existing Learning-Based PDE

Recently, Liu et al. [24, 25] utilize an optimal control technique to train PDEs for image processing. Although both [24, 25] and our work aim at learning PDEs for image analysis, the learning strategy in our work is different from theirs. In [24, 25], they adopt a nonlinear PDE formulation

Algorithm 1 Learning LESD for Saliency Detection

Input: Given an image I and necessary parameters.
Output: Saliency map for the given image.

Construct an image graph $\mathcal{G}$ on superpixels of I.
Calculate guidance map g using (4).
Initialize saliency seed set $\mathcal{S} \leftarrow \emptyset$.
while $|\mathcal{S}| \leq n$ **do**
  for $\mathbf{p} \in \mathcal{F}_c/\mathcal{S}$ **do**
    Solve (3) with saliency seeds $\mathcal{S} \cup \{\mathbf{p}\}$ for f.
    Obtain the gain $\Delta L(\mathbf{p}) = L(\mathcal{S} \cup \{\mathbf{p}\}) - L(\mathcal{S})$, or $\Delta \hat{L}(\mathbf{p}) = \hat{L}(\mathcal{S} \cup \{\mathbf{p}\}) - \hat{L}(\mathcal{S})$.
  end for
  $\mathbf{p}^* = \arg \max_{\mathbf{p} \in \mathcal{F}_c/\mathcal{S}} \Delta L(\mathbf{p})$ or $\arg \max_{\mathbf{p} \in \mathcal{F}_c/\mathcal{S}} \Delta \hat{L}(\mathbf{p})$.
  if $\hat{L}(\mathcal{S} \cup \{\mathbf{p}^*\}) \leq \hat{L}(\mathcal{S})$ (only for $\hat{L}$) **then**
    Break.
  end if
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathbf{p}^*\}$.
end while
Solve (3) with optimal g^* and $\mathcal{S}^*$ to obtain f^*.
Construct the final saliency map from f^*.

and learn the combination coefficients (i.e., the PDE form) from training image pairs (collected by hands). While our framework considers a linear elliptic system and learns *both the PDE form and its boundary conditions* to incorporate both bottom-up image structure and top-down human perception into our PDE system. Therefore, we can successfully handle the more complex saliency detection task.

4.2. Submodularity in Previous Vision Models

Submodularity is an important property for discrete set functions and has farreaching applications in operations research and machine learning [20]. It has also been applied to computer vision problems [19, 17, 15]. Although the work in [15] mentioned submodularity in their saliency detector, the mechanism of our work is very different from theirs. Specifically, the submodular optimization model in [15] is used to extract cluster centers[8] and graph clustering and saliency map computation steps are required in their framework. In contrast, we design a submodular optimization model to learn the Dirichlet boundary condition of the PDE system and directly extract the saliency map by solving the learnt PDE system (no further postprocessing is needed). Experimental results in the following section also show that our method achieves more accurate salient regions than [15].

[8] Similar clustering-based idea is also used in [17].

5. Experimental Results

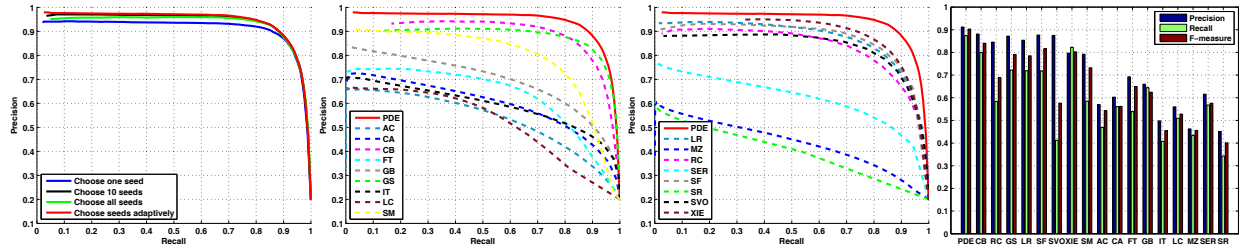
Experiments are performed on three image sets which are generated from two databases, i.e., MSRA [26] and Berkeley [28]. Firstly, we use a subset of MSRA with 1000 images provided by [2] (MRSA-1000). Then the comparison is performed on the whole MSRA database with 5000 images (MSRA-5000). Finally, we test algorithms on 300 more challenging images in the Berkeley image set. We set the number of superpixels as 200 for all the test images. We compare our methods (denoted as "PDE" in the comparisons) with seventeen state-of-the-art saliency detectors, such as IT [13], AC [1], CA [9], CB [14], FT [2], GB [10], GS [36], LC [41], LR [34], MZ [27], RC [7], SER [33], SF [30], SR [11], SM [15], SVO [6], and XIE [38]. For quantitative comparison, we report the precision, recall and F-measure values for the three image sets, respectively. We also present ground truth (GT) salient regions and the saliency maps for compared methods. For our method, we experimentally set $\beta = 10$ in the Gaussian similarity $k(\mathbf{p}, \mathbf{q})$ and $\lambda = 0.01$ in F for all test images.

5.1. Quantitative Comparisons

The quantitative comparisons between our method and other state-of-the-art approaches are performed on MSRA-1000, MSRA-5000, and Berkeley, respectively. The average precision, recall, and F-measure values are computed in the same way as in [2, 7, 38, 15].

We first compare the performance of our two objective functions (i.e., L and $\hat{L}$) on the MSRA-1000 image set and show the results in Fig. 5 (a). It can be seen that the $\hat{L}$-strategy performs well (red curve) because this non-monotonic model can adaptively determine the optimal $\mathcal{S}$. When we properly define a seed number ($n = 10$ in this case) for L, this monotone model can also achieve good performance (black curve). But it can be seen that the results of L-based strategy are dependent on the number of saliency seeds (blue and green curves). This is because a too small n may lead to insufficient diffusion, while a too large n may introduce incorrect nodes to the seed set. Based on this observation, we always utilize the $\hat{L}$-strategy in the following experiments.

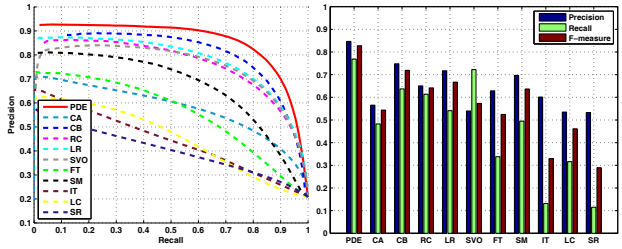
The precision-recall curves of all seventeen methods on MSRA-1000 are presented in Fig. 5 (b) and (c). The average precision, recall and F-measure values using an adaptive threshold [2] are shown in Fig. 5 (d). We also perform experiments on all 5000 images in the MSRA database. To achieve more reasonable comparison results, here we use accurate human-labeled masks rather than the bounding boxes used in the previous work to evaluate the saliency detection results. The results are presented in Fig. 6. The Berkeley image set is more challenging than MSRA as many images in this set contain multiple foreground objects with different sizes and locations. We report the comparison



(a) (b) (c) (d)

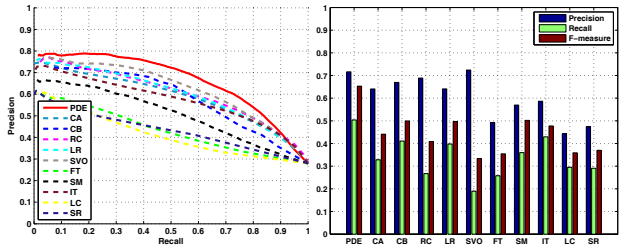
Figure 5. Results on the MSRA-1000 image set. (a) Precision-recall curves of our method with different design options. (b)-(c) Precision-recall curves of all test methods. (d) Average precision, recall, and F-measure values.

results in Fig. 7.



(a) (b)

Figure 6. Results on the MSRA-5000 image set. (a) Precision-recall curves. (b) Average precision, recall, and F-measure values.



(a) (b)

Figure 7. Results on the Berkeley image set. (a) Precision-recall curves. (b) Average precision, recall, and F-measure values.

The center-surround contrast based methods, such as IT [13], GB [10] and CA [9], can only detect parts of boundaries of salient objects. Using superpixels, recent approaches, such as CB [14] and RC [7], are capable of detecting salient objects. But they usually fail to suppress background regions and also lead to lower precision-recall curves. In Fig. 5 (b), we observe that GS [36] shares a similar precision with ours when the recall is larger than 0.96. However, the geodesic distance to boundary strategy in that method tends to recognize background parts as salient regions when their colors are significantly different from the boundary. So in most cases, their precision is much lower than ours at the same recall level. It can be seen that overall our PDE saliency detector achieves the best performance on all the three challenging image sets. These results also verify that the proposed learning strategy can successfully incorporate both bottom-up and top-down information into saliency diffusion.

5.2. Qualitative Comparisons

We show example saliency maps computed by some typical saliency detectors in Fig. 8. As an eye fixation prediction based method, IT [13] can only identify center-surround differences but misses most of the object information. The simple low-rank assumption in LR [34] may be invalid when images contain complex structures. RC [7] explores superpixels to highlight the object more uniformly, but the complex background always challenges such methods [9, 10, 7]. In SM [15], regions inside a salient object which share a similar color with the background will be regarded as part of the background. As a result, they may share the same saliency value with the background region. In contrast, our method can successfully highlight the salient regions and preserve the boundaries of objects, thus producing results that are much closer to the ground truth.

6. Conclusions

This paper develops a PDE system for saliency detection. We define a Linear Elliptic System with Dirichlet boundary (LESD) to model the saliency diffusion on an image and prove the submodularity of its solution. We then solve a submodular maximization model to optimize the boundary condition and incorporate high-level priors to learn the PDE formulation. We evaluate our PDE on various challenging image sets and compare with many state-of-the-art techniques to show its superiority in saliency detection. In the future, we plan to extend the submodular PDE learning technique to incorporate more complex human perception and high-level priors for other challenging problems in computer vision.

Acknowledgements

Risheng Liu would like to thank Gunhee Kim and Guangyu Zhong for useful discussions. Risheng Liu is supported by the NSFC (Nos. 61300086, 61173103,

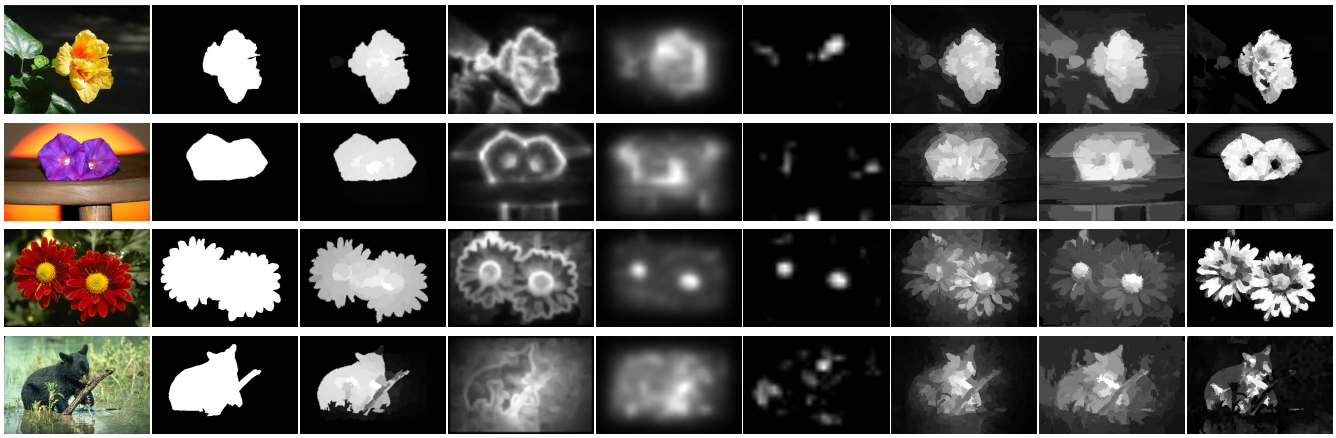


Figure 8. Qualitative comparisons of different approaches. The top three rows are examples in MSRA and the bottom is in Berkeley.

U0935004) and the China Postdoctoral Science Foundation. Junjie Cao is supported by the NSFC (No. 61363048). Zhouchen Lin is supported by the NSFC (Nos. 61272341, 61231002, 61121002). Shiguang Shan is supported by the NSFC (No. 61222211).

References

[1] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. In *ICVS*, 2008.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE T. PAMI*, 34(11):2274–2282, 2012.

[4] G. Calinescuy, C. Chekuri, M. Pal, and J. Vondrak. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM J. Computing*, 40(6):1740–1766, 2011.

[5] T. Chan and J. Shen. *Image processing and analysis: variational, PDE, wavelet, and stochastic methods*. SIAM, 2005.

[6] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *ICCV*, 2011.

[7] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, 2011.

[8] G. Gilboa and S. Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2008.

[9] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE T. PAMI*, 34(10):1915–1926, 2012.

[10] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006.

[11] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[12] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE T. IP*, 13(10):1304–1318, 2004.

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE T. PAMI*, 20(11):1254–1259, 1998.

[14] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. Automatic salient object segmentation based on context and shape prior. In *BMVC*, 2011.

[15] Z. Jiang and L. S. Davis. Submodular salient region detection. In *CVPR*, 2013.

[16] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[17] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011.

[18] B. C. Ko and J.-Y. Nam. Object-of-interest image segmentation based on human attention and semantic region clustering. *JOSA A*, 23(10):2462–2470, 2006.

[19] V. Kolmogorov and R. Zabin. What energy functions can be minimized via graph cuts? *IEEE T. PAMI*, 26(2):147–159, 2004.

[20] A. Krause and D. Golovin. Submodular function maximization. *Tractability: Practical Approaches to Hard Problems*, 3, 2012.

[21] A. Krause and C. Guestrin. Beyond convexity: Submodularity in machine learning. In *ICML Tutorials*, 2008.

[22] C. Lang, G. Liu, J. Yu, and S. Yan. Saliency detection by multitask sparsity pursuit. *IEEE T. IP*, 21(3):1327–1338, 2012.

[23] T. Lindeberg. *Scale-space theory in computer vision*. Springer, 1993.

[24] R. Liu, Z. Lin, W. Zhang, and Z. Su. Learning PDEs for image restoration via optimal control. In *ECCV*, 2010.

[25] R. Liu, Z. Lin, W. Zhang, K. Tang, and Z. Su. Toward designing intelligent PDEs for computer vision: An optimal control approachn. *Image and Vision Computing*, 31(1):43–56, 2013.

[26] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE T. PAMI*, 33(2):353–367, 2011.

[27] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM Multimedia*, 2003.

[28] V. Movahedi and J. H. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPR Workshops*, 2010.

[29] G. L. Nemhauser and L. A. Wolsey. Best algorithms for approximating the maximum of a submodular set function. *Mathematics of Operations Research*, 3(3):177–188, 1978.

[30] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012.

[31] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[32] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *CVPR*, 2004.

[33] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12), 2009.

[34] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, 2012.

[35] J. Van De Weijer, T. Gevers, and A. D. Bagdanov. Boosting color saliency in image feature detection. *IEEE T. PAMI*, 28(1):150–156, 2006.

[36] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV*. 2012.

[37] J. Weickert. *Anisotropic diffusion in image processing*, volume 1. Teubner Stuttgart, 1998.

[38] Y. Xie, H. Lu, and M.-H. Yang. Bayesian saliency via low and mid level cues. *IEEE T. IP*, 22(5):1689–1698, 2013.

[39] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[40] J. Yang and M.-H. Yang. Top-down visual saliency via joint CRF and dictionary learning. In *CVPR*, 2012.

[41] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *ACM Multimedia*, 2006.