Learning optimal seeds for diffusion-based salient object detection

Song Lu
SVCL Lab, UCSD
sol050@ucsd.edu

Vijay Mahadevan
Yahoo Labs
vmahadev@yahoo-inc.com

Nuno Vasconcelos
SVCL Lab, UCSD
nuno@ucsd.edu

Abstract

In diffusion-based saliency detection, an image is partitioned into superpixels and mapped to a graph, with superpixels as nodes and edge strengths proportional to superpixel similarity. Saliency information is then propagated over the graph using a diffusion process, whose equilibrium state yields the object saliency map. The optimal solution is the product of a propagation matrix and a saliency seed vector that contains a prior saliency assessment. This is obtained from either a bottom-up saliency detector or some heuristics. In this work, we propose a method to learn optimal seeds for object saliency. Two types of features are computed per superpixel: the bottom-up saliency of the superpixel region and a set of mid-level vision features informative of how likely the superpixel is to belong to an object. The combination of features that best discriminates between object and background saliency is then learned, using a large-margin formulation of the discriminant saliency principle. The propagation of the resulting saliency seeds, using a diffusion process, is finally shown to outperform the state of the art on a number of salient object detection datasets.

1. Introduction

Saliency detection has been a topic of substantial research in the recent vision literature. Early efforts aimed to predict the locations of human eye fixations and introduced the fundamental principles of saliency detection. [24] equated saliency to a local center-surround operation and proposed a biologically inspired saliency network. [16] introduced discriminant saliency, implementing this center-surround operation with discriminant measures of contrast between feature responses at an image location and its neighborhood. [9] proposed a global definition, which equates saliency to uniqueness of the visual stimulus, by setting the saliency of a location to the inverse probability of its responses under the response distribution of the whole image. [22] investigated the role of scale in saliency, proposing a measure based on the frequency spectrum of the image. [20] formulated saliency as the stationary distribution of a random walk defined over a graph based representation of the image. Many papers have since been written by combining these principles or extending them to other tasks [7].

While these methods were purely *bottom-up* (stimulus driven) some researchers proposed models of *top-down* saliency. These use high-level information to guide the saliency computation, equating saliency to the detection of stimuli from certain object classes [14]. While bottom-up saliency predicts eye fixations, these methods are useful for high-level vision tasks, such as object recognition and localization. However, they require the definition of object classes of interest and are, thus, less generic. A compromise is to either consider the responses of object detectors as features for bottom-up saliency [27], or formulate saliency as a mid-level vision task. In this area, substantial attention has been devoted to the problem of *object saliency* [8]. The goal is no longer to predict eye fixations or the locations of specific object classes, but to identify the locations of salient objects, independently of what these may be [30].

Methods in this class are generally based on the earlier principles of bottom-up saliency, e.g. local measures of center-surround contrast [11, 30, 26] or global measures of stimulus uniqueness [11]. They differ from the eye-fixation models in three main aspects. First, rather than pixels, they process image primitives informative of object shape and segmentation, e.g. superpixels [11, 26] or image patches [18, 17, 19]. Second, to overcome the propensity of bottom-up saliency to respond more to edges than (homogeneous) object interiors, they include some form of *spatial propagation* of saliency information. This can be implemented various ways, including conditional random fields (CRFs) [30, 31], random walk models [19], energy models [10, 19], or diffusion processes [38]. Third, beyond the classical measures of bottom-up saliency, these models may also account for *objectness features*. These are features indicative of the likelihood of a superpixel belonging to a generic object, e.g. based on the distribution of edges in the image, geometric constraints, etc. These features are included in the spatial propagation model, aiding to focus the propagation of information on object regions.

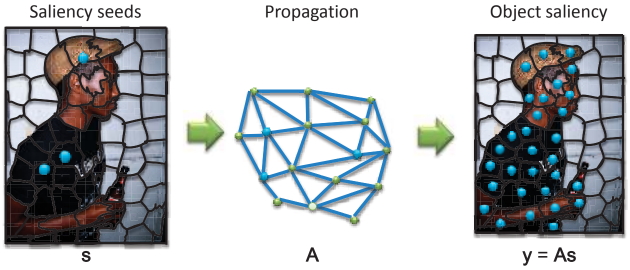

Figure 1: Salient object detection by graph diffusion. Images are decomposed into superpixel graphs. Given an initial set of saliency seeds $\mathbf{s}$ a diffusion process, characterized by matrix $\mathbf{A}$, is used to propagate saliency information throughout the graph. The optimal solution $\mathbf{y} = \mathbf{A}\mathbf{s}$ is an object saliency map.

In all these approaches, the image is mapped into a graph with image superpixels (or patches) as nodes and edge strength proportional to superpixel similarity. The spatial propagation model then performs inference on this graph, as illustrated in Figure 1. Under some models this corresponds to an energy minimization problem. For others, the process can be mapped into a random walk that converges to an equilibrium probability distribution. Finally, some models interpret the propagation as a manifold ranking problem, where nodes are ordered by similarity to a query consisting of a node subset. Interestingly, all these formulations have a simple closed-form solution, which is identical.

This solution consists of the product $\mathbf{A}\mathbf{s}$ of a matrix $\mathbf{A}$, derived from the affinity matrix of the graph, and a vector $\mathbf{s}$, which contains saliency information. While $\mathbf{A}$ controls the propagation of information, $\mathbf{s}$ is the set of *saliency seeds* to be propagated. Although several seed mechanisms have been proposed in the literature, they tend to be heuristic in nature, e.g. selecting the super-pixels that most differ from those along the image border [38]. In this work, we introduce an alternative procedure, which learns *optimal saliency seeds* for object saliency. For this, we define $\mathbf{s}$ as a linear combination of feature responses and learn the vector of feature weights that maximizes the discriminant saliency criterion of [16]. The optimization is 1) based on a large margin definition of the contrast between feature responses to object and background and 2) solved by gradient descent. The algorithm is applied to a combination of bottom-up saliency and objectness features. The resulting saliency detector can be interpreted as an object saliency detector driven by an eye fixation model, but which also accounts for mid-level vision cues that are informative of objectness. An extensive experimental evaluation shows that the procedure obtains state-of-the-art results on various object saliency datasets.

2. Graph-based saliency

In graph-based approaches to object saliency, an image is mapped into a graph $G = (V, E)$ of N nodes. Node v_i corresponds to the i^{th} image location (which could be a pixel, image patch, or superpixel) and edge e_{ij} links node pair (i, j). Given a vector of N saliency observations $\mathbf{s}$, e.g. node saliencies obtained with a bottom-up saliency detector, the goal is to determine an *object saliency map* $\mathbf{y}$ by propagating information along the graph. Note that $\mathbf{y}$ is again a N dimensional vector, with one entry per location.

2.1. Conditional random fields

A popular approach is to rely on a CRF [30, 31]. This is a model of the form $P(\mathbf{y}|\mathbf{s}) = \frac{1}{Z}\exp\{-\mathcal{E}(\mathbf{y}|\mathbf{s})\}$ where $\mathbf{y} \in \{-1, 1\}^N$ is a binary vector, Z a normalization constant and $\mathcal{E}(\mathbf{y}|\mathbf{s})$ an energy function. The latter is of the form

$$\mathcal{E}(\mathbf{y}|\mathbf{s}) = \sum_i \mathcal{A}(y_i, s_i) + \sum_{ij} \frac{1}{\mathcal{N}_i} \sum_{j \in \mathcal{N}_i} \mathcal{I}(y_i, y_j, s_i, s_j). \tag{1}$$

where $\mathcal{A}(y_i, s_i)$ is the unary potential at node i, $\mathcal{I}(y_i, y_j, s_i, s_j)$ the pairwise potential between nodes i and j, and $\mathcal{N}_i$ the set of neighbors of node i according to G. Typical energy functions for saliency have unary potentials that assign low energy to locations of large s_i, e.g. $\mathcal{A}(y, s) = -ys$, and pairwise potentials that encourage spatial smoothness, e.g. $\mathcal{I}(y_i, y_j, s_i, s_j) = y_i y_j |s_i - s_j|$.

2.2. Quadratic energy models

The binary nature of the labels y_i makes CRF learning and inference fairly complex, typically requiring iterative belief propagation algorithms. Binary labels are also not strictly needed for the saliency, where a graded saliency map is frequently preferred. A popular alternative to the binary CRF is to minimize an energy function of the form

$$\mathcal{E}(\mathbf{y}|\mathbf{s}) = \sum_i k_i (y_i - s_i)^2 + \lambda \sum_{i,j} \frac{1}{2} w_{i,j} (y_i - y_j)^2 \tag{2}$$

where $y_i \in [0, 1]$, k_i are weighs, usually set to the identity ($k_i = 1$), and $w_{i,j}$ entries of the *affinity matrix* of G, i.e. $w_{i,j}$ is some measure of similarity between nodes i and j [19]. It can be shown that this problem has closed form solution

$$\mathbf{y}^* = (\mathbf{K} + \lambda\mathbf{L})^{-1}(\lambda\mathbf{K}\mathbf{s}) \tag{3}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacian, $\mathbf{D} = diag\{d_{11}, d_{22}, ..., d_{NN}\}$ the degree matrix of the graph, i.e. $d_{ii} = \sum_j w_{ij}$, and $\mathbf{K} = diag\{k_1, k_2, ..., k_N\}$. Quadratic energy models have been used for object saliency in [10].

2.3. Random walks

An alternative possibility is to formulate saliency as the equilibrium distribution of a random walk on the graph G. In this case, $\mathbf{y}$ is a probability distribution over the nodes, and a transition probability matrix is defined as

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}. \tag{4}$$

Method	Propagation matrix $\mathbf{A}$
Quadratic energy models	$(\mathbf{K} + \lambda(\mathbf{D} - \mathbf{W}))^{-1}\mathbf{K}$
Random walks	$(\mathbf{I} - \alpha\mathbf{W}\mathbf{D}^{-1})^{-1}$
Manifold ranking	$(\mathbf{I} - \alpha\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2})^{-1}$

Table 1: Diffusion matrices for graph based similarity propagation.

The random walk is then characterized by the iteration $\mathbf{y}^{t+1} = \mathbf{P}^T\mathbf{y}^t$ and converges to an equilibrium distribution $\mathbf{y}^*$ which is the principal eigenvector of $\mathbf{P}$. This equilibrium distribution can be manipulated by introducing a vector $\mathbf{s}$ of probabilities of random jumps to the different nodes. In this case, the random walk is characterized by

$$\mathbf{y}^{t+1} = \alpha\mathbf{P}^T\mathbf{y}^t + (1-\alpha)\mathbf{s} \tag{5}$$

where $1-\alpha$ is the jump probability. It can be shown [12] that the resulting the equilibrium distribution is

$$\mathbf{y}^* = (1-\alpha)(\mathbf{I} - \alpha\mathbf{P}^T)^{-1}\mathbf{s}. \tag{6}$$

Random walk models have been proposed for object saliency detection in [19, 36].

2.4. Manifold ranking

In manifold ranking, the goal is to compute a rank y_i for each node in the graph with respect to a query $\mathbf{s}$. In this case the query is binary, $s_i \in \{0, 1\}$ and the optimal ranking minimizes an energy of the form

$$\sum_{i,j} w_{ij}\left(\frac{y_i}{\sqrt{d_{ii}}} - \frac{y_i}{\sqrt{d_{jj}}}\right)^2 + \mu\sum_i(y_i - s_i)^2 \tag{7}$$

where μ balances the unary (ensuring that the predicted rank matches the query) and pairwise (ensuring that the predictions are smooth over the graph) potentials. The optimal solution is [41]

$$\mathbf{y}^* = (\mathbf{I} - \alpha\mathbf{N})^{-1}\mathbf{s} \tag{8}$$

where $\mathbf{N} = \mathbf{D}^{-\mathbf{1/2}}\mathbf{W}\mathbf{D}^{-\mathbf{1/2}}$ is the normalized Laplacian of the graph and and $\alpha = 1/(1+\mu)$. This method has been used for object saliency detection in [38].

2.5. Saliency seeds

The methods above are quite similar. In all cases, the vector $\mathbf{s}$ is a set of *saliency seeds*, which are propagated throughout the graph, according to the node similarities defined by the affinity matrix $\mathbf{W}$. High affinities w_{ij} 1) encourage similar saliency for y_i and y_j under the energy model, 2) increase the probability of transitioning from node i to j under the random walk model, and 3) define a small geodesic distance between node i and j for manifold ranking. Nodes y_i^* with paths of high pairwise similarity to salient seed nodes in s_j^* receive high saliency values. In all cases, the optimal solution has the form

$$\mathbf{y}^* = \mathbf{A}\mathbf{s} \tag{9}$$

where $\mathbf{A}$ is a diffusion matrix, as defined in Table 1.

It should be noted that a larger set of diffusion processes has been studied in vision and learning. A review of these techniques, in the context of image retrieval, is presented in [12]. Most of these methods have the form above, using a different transition matrix $\mathbf{P}$. Much less attention has been devoted to the determination of the seed vector $\mathbf{s}$. This is not surprising, since in retrieval applications this is simply a binary vector that indicates query images. However, for saliency, the seeds s_i represent a preliminary[1] assessment of superpixel saliency and can play a significant role in the optimal solution. In the saliency context, they are best thought of as the outcome of a pre-attentive, purely stimulus driven, and mostly local perceptual process, which is extended into a spatially coherent saliency percept by the graph diffusion. This can be understood by noting that the seeds s_i define the unary potentials of (1), (2) and (7) and the probabilities of a random jump to each node in (5).

In the object saliency literature, saliency seeds have been mostly equated to bottom-up saliency. Different proposals include measures of pixel contrast [30], various implementations [30, 11, 25] of the discriminant center-surround saliency principle of [16], or variations [19] on the graph-based saliency model of [20]. This is usually complemented by a set of features that capture some properties of objects. These can be though as mid-level cues for object segmentation and are usually quite simple, e.g. measures of compactness of image regions of similar color [30, 36, 34, 25]. While some more sophisticated objectness measurement, such as the one proposed by [3], have been considered for object saliency [10], the impact of such mid-level cues is still poorly understood. In fact, there have been no attempts in the literature to produce optimal seeds for object saliency.

3. Optimal seeds for object saliency

In this section, we propose an algorithm for learning saliency seeds, by combining pre-attentive saliency maps and mid-level vision cues for object perception.

3.1. Feature based saliency diffusion

An image $\mathbf{x}$ is first segmented into N superpixels $\{\mathbf{x}_i\}$, $i = (1 \ldots N)$, using the algorithm of [2], and represented as a graph $G = (V, E)$ where each node corresponds to a

[1] Note that this is not to be confused with the initial state $\mathbf{y}^0$ of the random walk of (5), which does not have influence on the equilibrium solution of (6).

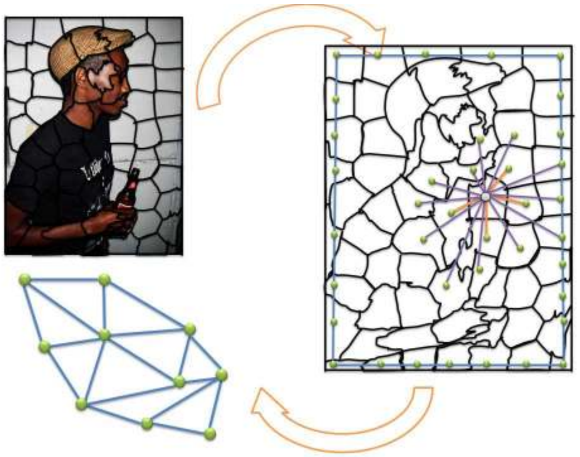

Figure 2: From image to graph: image is first decomposed into superpixels. A graph is then assembled by connecting 1) superpixel pairs that are either neighbors or "neighbors of neighbors," according to (10), and 2) boundary superpixels.

superpixel. Following [38], we adopt the affinity matrix

$$w_{ij} = \begin{cases} \nu(\mathbf{x}_i, \mathbf{x_j}) & \text{if } j \in \mathcal{N}_i \text{ or } \exists k \in \mathcal{N}_i | j \in \mathcal{N}_k \\ 0 & \text{otherwise} \end{cases} \qquad (10)$$

where $\mathcal{N}_i$ is the set of neighbors of superpixel $\mathbf{x}_i$. These are the superpixels that share an edge with $\mathbf{x}_i$. Furthermore, all superpixels in the boundary of the image are considered neighbors of each other. The construction of the graph is illustrated in Figure 1.

The function $\nu(\mathbf{x}_i, \mathbf{x}_j)$ is a measure of visual similarity of two superpixels. This is usually defined as a difference between the color distributions of $\mathbf{x}_i$ and $\mathbf{x}_j$. A discriminant alternative [26, 35], which we adopt, is to rely on a classifier to determine if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same object. In our implementation, this is a boosted decision tree, which operates on a feature space that accounts for color difference, texture difference, and geometric properties such as distance. Given the affinity matrix $\mathbf{W}$, saliency propagation is implemented with (3). While we have not investigated other diffusion schemes, we did perform some preliminary experiments to determine a good weighting matrix $\mathbf{K}$. The choice $\mathbf{K} = \mathbf{D}$ leads to the propagation matrix

$$\mathbf{A} = (\mathbf{I} - \mathbf{D}^{-1}\mathbf{W})^{-1}, \qquad (11)$$

which produced the best results in these experiments and is adopted in the remainder of the paper.

3.2. Seed representation

The emphasis of this work is on the seed vector $\mathbf{s}$. Since this contains a saliency value per graph node and nodes are image dependent, it is impossible to learn a universal $\mathbf{s}$. It is possible, however, to learn a universal combination of features informative of saliency. In this work we consider a linear combination of the form

$$\mathbf{s} = \mathbf{F}(\mathbf{x})\mathbf{w} \qquad (12)$$

where $\mathbf{F}$ is an $N \times K$ matrix, whose columns are the responses of K features to image $\mathbf{x}$. The weight vector $\mathbf{w}$ determines the contribution of the different features to the seed vector. Note that, with these seeds, the saliency map of (9) can be written as

$$\begin{aligned} \mathbf{y}(\mathbf{x}) &= \mathbf{A}(\mathbf{x})\mathbf{F}(\mathbf{x})\mathbf{w} && (13) \\ &= \sum_i w_i \mathbf{A}(\mathbf{x})\mathbf{f}_i(\mathbf{x}) && (14) \end{aligned}$$

where $\mathbf{f}_i(\mathbf{x})$ is the i^{th} column of $\mathbf{F}(\mathbf{x})$ and contains the saliency information derived from feature i. Hence, the diffusion process can be interpreted as a linear combination of *feature saliency diffusions* $\mathbf{y}_i(\mathbf{x}) = \mathbf{A}(\mathbf{x})\mathbf{f}_i(\mathbf{x})$, whose weights w_i encode the importance of the different features for saliency perception. This implies that the formulation can implement *feature-based attention* mechanisms, e.g. by varying $\mathbf{w}$ over time. This is a topic for future research.

3.3. Learning

In this work, $\mathbf{w}$ is a fixed vector learned from an annotated corpus $\mathcal{D} = \{(\mathbf{x}^{(1)}, \mathbf{m}^{(1)}), \ldots, (\mathbf{x}^{(n)}, \mathbf{m}^{(n)})\}$ of images $\mathbf{x}^{(i)}$ and object saliency maps $\mathbf{m}^{(i)}$. Optimality is defined in the discriminant saliency sense of [16], i.e. the optimal $\mathbf{w}$ is the one that best discriminates between object and background superpixels. This is implemented with the large-margin structured earning optimization [37]

$$\begin{aligned} \mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\alpha \parallel w \parallel^2 && (15) \\ + \sum_k \sum_{\{ij|\delta_i^k=1, \delta_j^k=0\}} \max(0, 1 - (y_i(\mathbf{x}^{(k)}) - y_j(\mathbf{x}^{(k)})) \end{aligned}$$

where $\mathbf{y}(\mathbf{x})$ is the object saliency map of (14) and $\delta_i^k = m_i(\mathbf{x}^{(k)})$ an indicator of the saliency of the i^{th} superpixel of the k^{th} image. Note that the optimization maximizes the object-background discrimination of the saliency maps at *equilibrium*, i.e. *after* the diffusion process of (9), by learning to rank the saliency of all pairs of object-background superpixels within each image. The optimization is carried out by gradient descent. The feature transformation $\mathbf{F}(\mathbf{x})$ includes both a pre-attentive bottom-up saliency map and mid-level features for object perception.

3.4. Pre-attentive saliency map

The bottom-up saliency map is based on a sparse image decomposition. Images are first decomposed into patches $\{\mathbf{t}_i\}$ of 8×8 pixels and 3 RGB color channels. A dictionary $\Psi \in \mathbb{R}^{192 \times B}$ of B basis functions is then learned from a collection of such patches [5]. The sparse representation is finally obtained with the fast decomposition of [23]. This consists of setting $B = 192$ and approximating the sparse coefficients of patch $\mathbf{t}$ with $\boldsymbol{\xi} = \Psi^{-1}\mathbf{t}$.

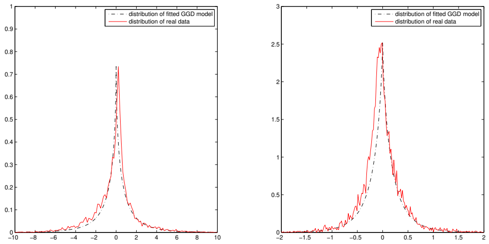

Figure 3: Histograms of two ξ_i and MAP GGD fits.

The sparse coefficients are used as features for the computation of saliency. Let $\xi_i(\boldsymbol{l})$, $i \in \{1, \ldots, B\}$, be the sparse coefficients of $\mathbf{t}(\boldsymbol{l})$, the patch centered at location $\boldsymbol{l}$ of image $\mathbf{x}$. When $\mathbf{x}$ is a natural image, the probability distribution of ξ_i is well approximated by a generalized Gaussian distribution (GGD)

$$p(\xi_i; \alpha_i, \beta_i) = \frac{\beta_i}{2\alpha_i \Gamma(\frac{1}{\beta_i})} \exp\left(-\left|\frac{\xi_i}{\alpha_i}\right|^{\beta_i}\right), \tag{16}$$

where α_i is a scale parameter, β_i a shape parameter, and $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} \mathrm{d}t$, $t > 0$, the Gamma function. Given a set of n i.i.d samples, $D_i = \{\xi_i^{(1)}, ..., \xi_i^{(n)}\}$, the maximum a posteriori (MAP) estimate of α_i, under the assumption of a conjugate (Gamma distributed) prior, is given by [16]

$$\hat{\alpha}_{i,MAP} = \left[\frac{1}{k_i}\left(\sum_{j=1}^{n} |\xi_i^{(j)}|^{\beta_i} + \nu\right)\right]^{1/\beta_i}, \tag{17}$$

where $k_i = \frac{n+\eta}{\beta_i}$ and ν, η are fixed prior parameters ($\nu = \eta = 10^{-3}$). As in [15], β_i is learned for each image, using

$$\sigma^2 = \frac{\alpha^2 \Gamma(\frac{3}{\beta_i})}{\Gamma(\frac{1}{\beta_i})} \qquad \kappa = \frac{\Gamma(\frac{1}{\beta_i})\Gamma(\frac{5}{\beta_i})}{\Gamma(\frac{3}{\beta_i})^2} \tag{18}$$

where σ^2 and κ are the variance and kurtosis of the i^{th} feature. Figure 3 shows typical examples of the resulting fits.

Following [9], the saliency of image location $\boldsymbol{l}$ is

$$s(\xi_i(\boldsymbol{l}); \alpha_i, \beta_i) = -\log p(\xi_i(\boldsymbol{l}); \alpha_i, \beta_i). \tag{19}$$

Using (16) and (17) leads to

$$s(\xi_i(\boldsymbol{l})) = \frac{|\xi_i(\boldsymbol{l})|^{\beta_i}}{\frac{1}{k}(\sum_{j=1}^{n} |\xi_i^{(j)}|^{\beta_i} + \nu)} + Q \tag{20}$$

where Q is a constant that does not depend on $\xi_i(\boldsymbol{l})$. This saliency measure can be computed efficiently using integral images [16]. In our implementation, the normalizing samples $\xi_i^{(j)}$ are the sparse coefficients in a neighborhood of the patch location $\boldsymbol{l}$. Finally, 1) the saliency maps derived from the B feature channels are combined with

$$s_f(\boldsymbol{l}) = \sum_{1}^{B} a_i s(\xi_i(\boldsymbol{l})) \tag{21}$$

and 2) the pre-attentive saliency score of superpixel $\mathbf{x}_i$ set to the mean saliency score of its pixels. The weights a_i are learned. In our experiments this is based on the eye fixation dataset of [9], from which we randomly selected 10 of the top 2% salient patches (according to the ground truth) as positive examples and 10 out of the bottom 40% as negative examples in each image. The associated saliency scores ($s(\xi_i)$ and ground truth) are used to learn the weights a_i using realboost [13].

3.5. Mid-level features

A number of features have been proposed to evaluate the likelihood of a superpixel belonging to a generic object. The following were adopted to simulate mid-level saliency cues.
Element uniqueness: measures the rarity of a superpixel color [34] with

$$\mathcal{U}_i = \sum_{j=1}^{N} ||c_i - c_j||^2 w_{ij}^{(\boldsymbol{l})} \tag{22}$$

$$w_{ij}^{(\boldsymbol{l})} = \frac{1}{Z_i} \exp(-\frac{1}{2\sigma_{\boldsymbol{l}}^2} ||\boldsymbol{l}_i - \boldsymbol{l}_j||^2) \tag{23}$$

where $\boldsymbol{l}_i$ and c_i are the position and average CIELab color of the i^{th} superpixel, respectively. Z_i is a normalization factor to ensure that $\sum_{j=1}^{N} w_{ij}^{(\boldsymbol{l})} = 1$. We compute the element uniqueness scores using three values of $\sigma_{\boldsymbol{l}}$.
Element distribution measures the spatial variance of the color of a superpixel [34], according to

$$\mathcal{D}_i = \sum_{j=1}^{N} ||\boldsymbol{l}_j - \boldsymbol{l}_i^{(\mu)}||^2 w_{ij}^{(c)} \tag{24}$$

where $w_{ij}^{(c)} = \frac{1}{Z_i} \exp(-\frac{1}{2\sigma_c^2} ||c_i - c_j||^2)$ and $\boldsymbol{l}_i^{(\mu)} = \sum_{j=1}^{N} w_{ij}^{(c)} \boldsymbol{l}_j$ is the center of mass of color c_i. Z_i is a normalization constant such that $\sum_{j=1}^{N} w_{ij}^{(c)} = 1$. Again, we compute the element distribution scores for three different values of σ_c.
Pattern distinctness [32] measures the ℓ_1 mean distance of patches in a superpixel to the mean patch, by principal component analysis (PCA). This is defined as $\mathcal{P}(\mathbf{x}_i) = ||\hat{\mathbf{x}}_i||_1$, where $\hat{\mathbf{x}}_i$ contains the PCA coefficients of patch $\mathbf{x}_i$.
Color distinctness same as pattern distinctness but for a PCA of the RGB color space of each patch.
Center bias distance between superpixel center and image center, normalized to $[0, 1]$.

Backgroundness similarity between a superpixel and the superpixels in the four image boundaries, using the similarity measure $\nu(x_i, x_j)$ of (10).
Local contrast measures [26] based on Chi-square distances between distributions of color and texton response [29] and geometric attributes such as size, and position.

4. Experiments

Several experiments were performed to evaluate the proposed saliency detector. Although eye fixation prediction is not the primary objective of this work, we first study the effectiveness the bottom-up saliency map of Section 3.4 in predicting eye fixations. This is followed by an evaluation of the object saliency detector.

4.1. Eye fixation prediction

Protocol: We use four benchmark datasets of eye fixations. Bruce [9] contains 120 images, mostly of indoor and city scenes. Judd [27] consists of 1003 images, many of which contain human faces. This dataset has a high degree of photographer bias and few subjects per scene. Kootstra [28] contains 100 images from a wide variety of scenes. Finally, the subset of Pascal VOC2008 of [39] (denoted VOC2008_1000) contains 1000 images. The parameters of the proposed model were learned on VOC2OO8_1000 and the model tested on the remaining datasets. Performance is measured with the shuffled AUC score of [40].

The proposed saliency detector (denoted sparse-GGD) is compared to the image signature method of [21], the sparse method of [23], the spectral method of [22], and the SUN method [40]. As discussed in [6], the smoothness of the saliency map plays an important role in the performance of eye-fixation algorithms. We choose the optimal Gaussian smoothing kernel for each algorithm. The performance is summarized in Table 2. The proposed saliency model consistently outperforms all other models on the four datasets.

4.2. Salient object detection

Protocol: All models are evaluated by comparing the predicted object saliency map to the binary ground truth, using two metrics: AUC and AP (average precision) scores. Five datasets are used. MSRA5000 [30] contains 5000 images from the MSRA dataset, with refined manually segmented ground-truth. SED1 [4] contains 100 images, each with a single salient object. SED2 [4] contains 100 images with two salient objects each. Both SED1 and SED2 include pixel-wise salient object segmentation groundtruth. SOD [33] contains 300 images from the Berkeley segmentation dataset with the salient object boundary annotated. Since this dataset contains multiple objects per image, it can be challenging. Finally, VOC2008_1023 is a subset of the VOC2008 dataset, where each image contains one or more salient objects plus segmented groundtruth.

The proposed saliency detector is compared to eight of the best performing methods in the literature - Gof [17], CB [25], HC [11], RC [11], GBMR [38], PCA [32], FT [1], and SF [34]. For the proposed approach, we considered two variants. *SalseedProp* has no learning, using bottom-up saliency alone to determine seeds and (11) for propagation. *OptseedProp* uses the optimal seeds learned by (15) and (11) for propagation. Table 3 summarizes the performance of all algorithms on all datasets. Note that SOD and VOC2008_1023 are most challenging, MSRA5000 and SED1 easier, and SED2 has intermediate difficulty. Among the previous methods, the recently proposed GBMR and PCA achieved best performance. The proposed SalseedProp is better than FT, Gof and HC in all datasets, and than CB in all but the relatively easy MSRA5000 dataset. This is likely because CB is well suited to scenes with a simple background and a single object. SalseedProp performs on par with RC but is inferior to GBMR and PCA. On the other hand, OptseedProp performs consistently better than all of Gof, CB, HC, RC, GBMR, PCA, FT, SF, in all datasets except SED2 (where RC performs slightly better). The result of applying the various algorithms to representative images from VOC2008_1023 and MSRA5000 are shown in Figure 4. Note how learning optimal salient seeds can improve saliency performance.

5. Conclusion

In this work, we presented an approach for salient object detection. The image is represented as a graph over the set of its superpixels. The saliency of salient seed locations is propagated through the graph via a diffusion process. Unlike previous heuristic approaches to seed selection, an optimal set of salient seeds is learned using a large margin formulation of the discriminant saliency principle. An extensive experimental evaluation shows that the procedure obtains state of the art results on a number of object saliency datasets.

Acknowledgements

This work was partially funded by the NSF awards IIS-1208522 and CCF-0830535.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604. IEEE, 2009. 6

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels. In *École Polytechnique Fédéral de Lausssanne (EPFL), Tech. Rep*, volume 149300. Citeseer, 2010. 3

Table 2: Eye fixation prediction performance: Gaussian kernel/Shuffled AUC score

Gaussian kernel/Shuffled AUC score	Kootstra	Bruce	Judd	VOC2008_1000
RGB-Signature [21]	0.030/0.5869	0.040/0.6900	0.040/0.6547	0.065/0.6497
LAB-Signature [21]	0.040/0.6020	0.045/0.7115	0.040/0.6631	0.050/0.6595
Sparse [23]	0.015/0.6024	0.030/0.6956	0.020/0.6629	0.030/0.6491
Spectral [22]	0.040/0.5865	0.040/0.6898	0.040/0.6545	0.065/0.6527
SUN [40]	0.020/0.5609	0.030/0.6663	0.030/0.6565	0.050/0.6373
sparse-GGD	0.015/**0.6105**	0.030/**0.7140**	0.035/**0.6751**	0.050/**0.6681**

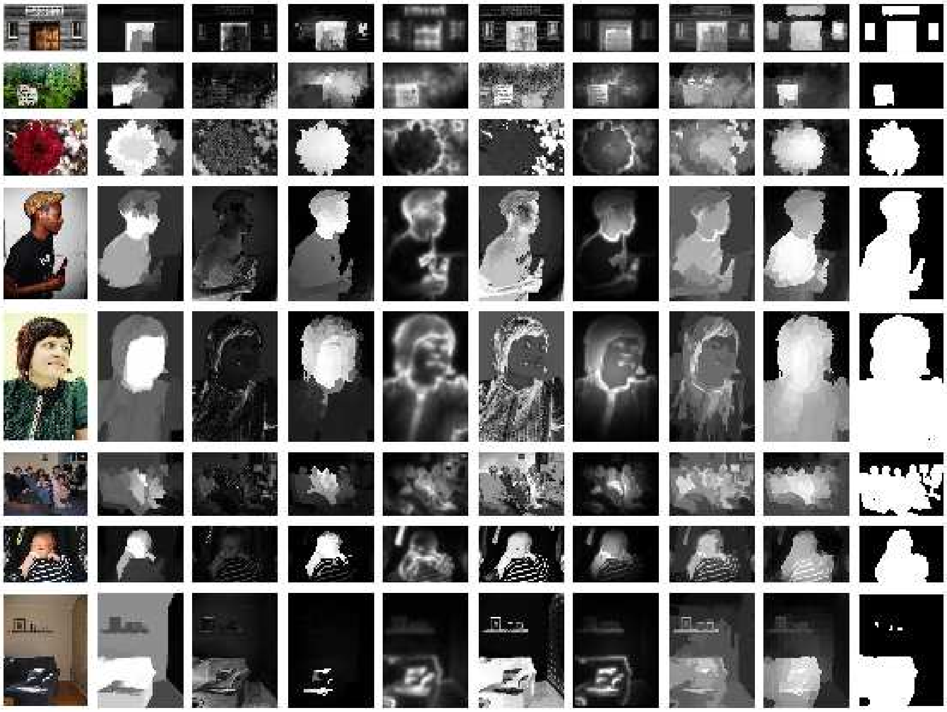

Figure 4: Comparison of results on representative images of VOC2008_1023 and MSRA5000. The original image is shown on the leftmost column. The other columns, from left to right, are the outputs of: 'CB','FT','GBMR','Gof', 'HC','PCA','RC','OptSeedProp (proposed)'. The binary ground truth is shown in the rightmost column.

[3] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, pages 73–80. IEEE, 2010. 3

[4] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *CVPR*, pages 1–8. IEEE, 2007. 6

[5] A. J. Bell and T. J. Sejnowski. The independent components of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997. 4

[6] A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, pages 478–485. IEEE, 2012. 6

[7] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012. 1

[8] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV*, pages 414–429. Springer, 2012. 1

[9] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2005. 1, 5, 6

[10] 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*, pages 914–921. IEEE, 2011. 1, 2, 3

[11] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416. IEEE, 2011. 1, 3, 6

[12] M. Donoser and H. Bischof. Diffusion processes for retrieval revisited. In *CVPR*, 2013. 3

[13] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002. 5

[14] D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):989–1005, 2009. 1

[15] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. In *ICCV*, pages 1–6. IEEE, 2007. 5

[16] D. Gao and N. Vasconcelos. Decision-theoretic saliency: Computational principles, biological plausibility, and im-

Table 3: Object saliency detection performance: AUC/AP

AUC/AP	MSRA5000	SOD	SED1	SED2	VOC2008_1023
CB	0.9281/0.8289	0.7672/0.6235	0.9105/0.8380	0.8741/0.7767	0.7546/0.6158
FT	0.7605/0.5603	0.6078/0.4274	0.6699/0.5493	0.8205/0.7225	0.6071/0.4493
Gof	0.8622/0.6214	0.8027/0.5818	0.8513/0.6804	0.8617/0.6474	0.7847/0.5959
HC	0.8223/0.6452	0.6612/0.4646	0.7770/0.6311	0.8769/0.7773	0.6525/0.4756
RC	0.9200/0.7724	0.8133/0.6337	0.8881/0.7633	**0.9142/0.8272**	0.7965/0.6186
GBMR	0.9424/0.8614	0.8319/0.6759	0.9341/0.8841	0.8360/0.7548	0.7838/0.6442
PCA	0.9407/0.8057	0.8414/0.6423	0.9085/0.7862	0.9035/0.7905	0.8102/0.6451
SalseedProp	0.9058/0.8136	0.8175/0.6688	0.9176/0.8537	0.8806/0.7500	0.7908/0.6421
OptseedProp	**0.9615/0.8790**	**0.8684/0.7019**	**0.9530/0.8905**	0.9058/0.8062	**0.8181/0.6556**

plications for neurophysiology and psychophysics. *Neural Computation*, 21(1):239–271, 2009. 1, 2, 3, 4, 5

[17] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1915–1926, 2012. 1, 6

[18] V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs to model saliency in images. In *CVPR*, pages 1698–1705. IEEE, 2009. 1

[19] V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs for salient object detection in images. *Image Processing, IEEE Transactions on*, 19(12):3232–3242, 2010. 1, 2, 3

[20] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006. 1, 3

[21] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):194–201, 2012. 6, 7

[22] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1–8. IEEE, 2007. 1, 6, 7

[23] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *Advances in neural information processing systems*, pages 681–688, 2008. 4, 6, 7

[24] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998. 1

[25] 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*, volume 3, page 7, 2011. 3, 6

[26] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013. 1, 4, 6

[27] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113. IEEE, 2009. 1, 6

[28] G. Kootstra and L. R. Schomaker. Prediction of human eye fixations using symmetry. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society (CogSci09)*, pages 56–61. Cognitive Science Society, 2009. 6

[29] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001. 6

[30] T. Liu, J. Sun, and H.-Y. Tang, Xiaoou aTang Shum. Learning to detect a salient object. In *CVPR*. IEEE, 2007. 1, 2, 3, 6

[31] L. Mai, Y. Niu, and F. Liu. Saliency aggregation: A data-driven approach. In *CVPR*, 2013. 1, 2

[32] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *CVPR*, 2013. 5, 6

[33] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, volume 2, pages 416–423. IEEE, 2001. 6

[34] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740. IEEE, 2012. 3, 5, 6

[35] E. Rahtu, J. Kannala, and M. Blaschko. Learning a category independent object detection cascade. In *ICCV*, pages 1052–1059. IEEE, 2011. 4

[36] Z. Ren, Y. Hu, L.-T. Chia, and D. Rajan. Improved saliency detection based on superpixel clustering and saliency propagation. In *Proceedings of the international conference on Multimedia*, pages 1099–1102. ACM, 2010. 3

[37] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, page 104. ACM, 2004. 4

[38] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 1, 2, 3, 4, 6

[39] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Studying relationships between human gaze, description, and computer vision. In *CVPR*, pages 739–746. IEEE, 2013. 6

[40] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008. 6, 7

[41] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. *Advances in neural information processing systems*, 16:169–176, 2003. 3