Saliency Detection via Divergence Analysis: A Unified Perspective

Jia-Bin Huang and Narendra Ahuja
University of Illinois at Urbana-Champaign
{jbhuang1,n-ahuja}@illinois.edu

Abstract

A number of bottom-up saliency detection algorithms have been proposed in the literature. Since these have been developed from intuition and principles inspired by psychophysical studies of human vision, the theoretical relations among them are unclear. In this paper, we present a unifying perspective. Saliency of an image area is defined in terms of divergence between certain feature distributions estimated from the central part and its surround. We show that various, seemingly different saliency estimation algorithms are in fact closely related. We also discuss some commonly used center-surround selection strategies. Experiments with two datasets are presented to quantify the relative advantages of these algorithms.

1 Introduction

Visual saliency is the perceptual quality which makes some items in the scene pop out from their surround and immediately attract our attention. It is well-known that humans can detect salient areas effortlessly even in complex scenes and in clutter. An effective computational model for automatically generating saliency map from images is of great interests because it can facilitate many important computer vision and graphics applications, including adaptive image compression, object detection and recognition, thumbnail generation, content-aware image re-targeting, and non-photorealistic rendering, among many others.

Inspired by some principles of human visual attention supported by psychological studies, many saliency detection algorithms have been proposed in recent years. Here we list several principles that have been exploited in the literature:

- Rarity: less frequently-occurring features are salient [4, 27, 22]
- Local complexity: local unpredictability indicates high saliency [15]
- Contrast: High center-surround contrast draws visual attention [12, 9, 7, 1, 5]
- Priors: Special characteristics of high-level content of complex images, e.g., faces, learned from examples [19, 14, 21]

We refer readers to [24, 3] for a comprehensive survey of saliency detection algorithms. However, due to the difference in principles used and implementation details, the underlying relations among these methods remain hard to understand and their fundamental capabilities are unclear.

In this paper, we present a unifying perspective of *bottom-up* saliency detection algorithms. The saliency of a pixel is defined as the divergence between the probability distributions estimated using samples from center and surround, respectively. We explicitly show that most of the current bottom-up saliency models are in fact special cases within our formulation, corresponding to various assumptions and approximation (Section 2). Therefore, we provide a standardized interpretation of the quantities involved in them. Moreover, as divergence has well-known fundamental connections with well-established fields such as information theory and statistical decision theory [23], we can understand these methods in a more principled way. In addition, we discuss commonly used center-surround selection strategies (Section 3). Experimental results are shown to reveal the relative advantages of each algorithms and provide further insights (Section 4).

2 A Unifying Framework

2.1 Center-Surround Divergence

Denote $x_i \in \mathcal{M}, 1 \leq i \leq N$ as the i_{th} pixel location in an image with N pixels and spatial support $\mathcal{M}$ and $f_{x_i} \in \mathbb{R}^d$ as the features extracted at position x_i, e.g., luminance, color, orientation, texture, or motion. For a pixel located at x_i, we first define two disjoint spatial supports, namely center $\mathcal{C}_i$ and surround $\mathcal{S}_i$. We also denote their union as $\mathcal{A}_i = \mathcal{C}_i \cup \mathcal{S}_i$ and the patch centered at x_i as $\mathcal{N}_i$.

The saliency of x_i can thus be defined as the divergence between the two feature distributions estimated using samples from center and surround:

$$s_{x_i} = D(P_{\mathcal{C}_i} || P_{\mathcal{S}_i}), \tag{1}$$

where $D(\cdot||\cdot)$ is a contrast function which establishes the dissimilarity of one probability distribution to the other on a statistical manifold. The most frequently used class of divergences is the so-called f-divergence, which includes the well-known Kullback-Leibler divergence (KL divergence) as a special case.

Table 1. A summary of saliency detection algorithms using divergence analysis

Basic form	Works	Assumptions and Notes	Center-surround selection strategies
$D_{KL}(P_{\mathcal{C}_i} \| P_{\mathcal{S}_i})$	[16]	Independence among feature channels	Multi-scale, patch-based
	[4, 27]	Self-information	$\mathcal{C}_i$: $\{x_i\}$, $\mathcal{S}_i$: $\mathcal{N}_i \setminus x_i$
	[22]	Difference of self-information	Single-scale patch-based
	[15]	Surround distribution $P_{\mathcal{S}_i} \sim P_U$	$\mathcal{C}_i$: adaptive, $\mathcal{S}_i$: $\mathcal{M} \setminus x_i$
$D_{KL}(P_{\mathcal{S}_i} \| P_{\mathcal{C}_i})$	[20]	Downsample image for speedup	$\mathcal{C}_i$: $\{x_i\}$, $\mathcal{S}_i$: $\mathcal{N}_i \setminus x_i$
	[26]	Luminance feature, look-up table for speedup	$\mathcal{C}_i$: $\{x_i\}$, $\mathcal{S}_i$: $\mathcal{M} \setminus x_i$
	[12, 9]	Contrast as center-surround difference	$\mathcal{C}_i$: fine scales, $\mathcal{S}_i$: coarse scales
	[1]	Replace all samples with its mean for speedup	$\mathcal{C}_i$: $\{x_i\}$, $\mathcal{S}_i$: $\mathcal{M} \setminus x_i$
	[2]	Maximum symmetric surround	$\mathcal{C}_i$: $\{x_i\}$, $\mathcal{S}_i$: adaptive
	[8]	K nearest neighbor for approximation	$\mathcal{C}_i$: $\{x_i\}$, $\mathcal{S}_i$: center-weighted
$D_{\lambda}(P_{\mathcal{C}_i} \| P_{\mathcal{S}_i})$	[7]	Discriminant center-surround hypothesis	Single-scale, patch-based
$D_{CS}(P_{\mathcal{C}_i} \| P_{\mathcal{S}_i})$	[5]	Sparse histogram comparison	$\mathcal{C}_i$: regions, $\mathcal{S}_i$: center-weighted

In the following subsections, we show that most of the saliency detection algorithms in the literature share Eqn. 1, which represents their common nature. Table 1 presents a summary of various saliency detection algorithms for different definitions of divergence.

2.2 From Center to Surround

We first consider saliency s_{x_i} as the KL divergence from center to surround, i.e., from $P_{\mathcal{C}_i}$ to $P_{\mathcal{S}_i}$:

$$s_{x_i} = D_{KL}(P_{\mathcal{C}_i}||P_{\mathcal{S}_i}) = \sum_f P_{\mathcal{C}_i}(f) \log \frac{P_{\mathcal{C}_i}(f)}{P_{\mathcal{S}_i}(f)}. \quad (2)$$

By assuming independence among the dimensions in f_{x_i}, one can compute the KL divergence in each feature channel and fuse them to form the final saliency map [16].

By shrinking the center support to a single pixel x_i, i.e., $C_i = \{x_i\}$, we have $P_{\mathcal{C}_i}(f_{x_i}) = 1$. Then, Eqn. 2 simplifies to

$$s_{x_i} = I(f_{x_i}) = -\log P_{\mathcal{S}_i}(f_{x_i}), \quad (3)$$

which yields the Shannon's self-information as used in AIM [4] and SUN [27] models.

The difference between the self-information of observing f_{x_i}evaluated using $P_{\mathcal{A}_i}$ and $P_{\mathcal{C}_i}$ has the form

$$-\log P_{\mathcal{A}_i}(f_{x_i}) - (-\log P_{\mathcal{C}_i}(f_{x_i})) = \log \frac{P_{\mathcal{C}_i}(f_{x_i})}{P_{\mathcal{A}_i}(f_{x_i})},$$

which gives rise to the saliency measure defined in [22].

By assuming the surround distribution $P_{\mathcal{S}_i}$ to be uniform P_U, we can build connection with the local complexity-based methods [15], which uses entropy of a local region as a saliency measure:

$$H(P_{\mathcal{C}_i}) = \log |\mathcal{F}| - D_{KL}(P_{\mathcal{C}_i}||P_U), \quad (4)$$

where $\mathcal{F}$ is the set of the feature values.

2.3 From Surround to Center

As the KL divergence is not symmetric, one can compute the saliency as the KL divergence from the opposite direction:

$$s_{x_i} = D_{KL}(P_{\mathcal{S}_i}||P_{\mathcal{C}_i}) = \sum_f P_{\mathcal{S}_i}(f) \log \frac{P_{\mathcal{S}_i}(f)}{P_{\mathcal{C}_i}(f)}. \quad (5)$$

The meaning of Eqn. 5 and 2 can be better understood via the fundamental connection between the KL divergence and the likelihood theory [6].

$$D_{KL}(P_{\mathcal{S}_i}||P_{\mathcal{C}_i}) = -H(P_{\mathcal{S}_i}) - \sum_f P_{\mathcal{S}_i}(f) \log P_{\mathcal{C}_i}(f), \quad (6)$$

where the second term of the right hand side can be rewritten as the minus log-likelihood function:

$$-\sum_f P_{\mathcal{S}_i}(f) \log P_{\mathcal{C}_i}(f) = \frac{-1}{|\mathcal{S}_i|} \sum_{j:x_j \in \mathcal{S}_i} \log P_{\mathcal{C}_i}(f_{x_j}).$$

We can then interpret the quantity $D_{KL}(P_{\mathcal{S}_i}||P_{\mathcal{C}_i})$ as how well the probabilistic model of center $P_{\mathcal{C}_i}$ can explain the samples from surround. If $P_{\mathcal{C}_i}$ can provide a good fit of the surrounding samples, then the saliency s_{x_i} is small, and vice versa. On the other hand, Eqn. 2 measure saliency as how well the model of surround $P_{\mathcal{S}_i}$can explain samples from center.

We can view the likelihood model in Eqn. 6 as a generalization of many contrast-based methods [26, 1, 20, 2, 8], which makes different assumptions and approximations. For example, by shrinking the center support to a single pixel x_i and assuming the form of $P_{\mathcal{C}_i}$ as Gaussian distribution with mean f_{x_i} and variance σ^2, the minus log-likelihood in Eqn. 7 become

$$\frac{1}{|\mathcal{S}_i|} \sum_{j:x_j \in \mathcal{S}_i} \frac{(f_{x_i} - f_{x_j})^2}{\sigma^2} + const \quad (7)$$

Many of the contrast-based methods measure saliency by approximately evaluating Eqn. 7. As examples,

[1] replaces all f_{x_j} with its mean ($\frac{1}{|\mathcal{S}_i|}\sum_{j:x_j\in\mathcal{S}_i} f_{x_j}$), [12, 9] use difference between fine and coarse scales in Gaussian pyramids, [2] use adaptive surround $\mathcal{S}_i$, [26] assume that $P_{\mathcal{C}_i}$ follows Laplacian distribution, and [8] uses k nearest neighbors to achieve computational efficiency.

2.4 Symmetrised Divergence

In contrast to the non-symmetric KL divergence, some symmetrised divergences have also been proposed. One example is the λ divergence:

$$D_\lambda(P||Q) = \lambda D_{KL}(P||A)+(1-\lambda)D_{KL}(Q||A), \tag{8}$$

where P, Q, A are probability distributions and $A = \lambda P + (1-\lambda)Q$. By appropriately choosing λ as the prior probability of the center $\lambda = |\mathcal{C}_i|/|\mathcal{A}_i|$, the λ divergence between center and surround is

$$D_\lambda(P_{\mathcal{C}_i}||P_{\mathcal{S}_i}) = \lambda D_{KL}(P_{\mathcal{C}_i}||P_{\mathcal{A}_i})+(1-\lambda)D_{KL}(P_{\mathcal{S}_i}||P_{\mathcal{A}_i}),$$

which is the mutual information of feature distribution and center-surround label used in [7].

Another alternative is the Cauchy-Schwarz divergence [13], which is given by

$$D_{CS}(P||Q) = -\log\frac{\int P(x)Q(x)\mathrm{d}x}{\sqrt{\int P(x)^2\mathrm{d}x\int Q(x)^2\mathrm{d}x}}.$$

When estimating the probabilistic density P, Q using non-parametric density estimation (known as Parzen windowing), the CS divergence can be easily evaluated. Specifically, we estimate the pdf of $P_{\mathcal{C}_i}$ using

$$\hat{P}_{\mathcal{C}_i}(f_x) = \frac{1}{|\mathcal{C}_i|}\sum_{j:x_j\in\mathcal{C}_i} W_{\sigma^2}(f_x, f_{x_j}), \tag{9}$$

where $W_{\sigma^2}(\cdot,\cdot)$ is a Gaussian kernel with parameter σ^2. Then the CS divergence $D_{CS}(P_{\mathcal{C}_i}||P_{\mathcal{S}_i})$ has the form

$$-\log\frac{\sum_{l:x_l\in\mathcal{C}_i}\sum_{j:x_j\in\mathcal{S}_i}K_{l,j}}{\sqrt{\sum_{l,l':x_l,x'_l\in\mathcal{C}_i}K_{l,l'}\sum_{j,j':x_j,x'_j\in\mathcal{S}_i}K_{j,j'}}},$$

where $K_{l,j}$ denotes $W_{2\sigma^2}(f_{x_l}, f_{x_j})$. This gives rise to the histogram contrast saliency measure in [5].

3 Center-Surround Support Selection

The center-surround hypothesis for saliency detection is inspired by the center-surround mechanisms occurring in the early stages of biological vision [11, 25]. However, the selection of the center and surround is not trivial. We investigate various strategies used for selecting support of center and surround.

3.1 Selection of Center Support

The simplest choice for center is to use a single pixel [4, 27, 8, 2]. However, estimating the distribution of center with a single observation clearly introduces high variance. As ways of increasing the sample size, patch or window-based approaches have been proposed [22, 16]. Yet, without knowing image discontinuities in the vicinity, the optimal patch/window size of center cannot be estimated, which can be only partly mitigated by added complexity of multi-scale computation. Region-based approaches emerge as a good choice for spatial support estimation of center. Region-based methods provide appropriate spatial scales and directly involve potential object boundaries. Note that region-based saliency is different from region-enhanced saliency, which amounts to post-processing of pixel/patch-based saliency by averaging them over segments [18].

3.2 Selection of Surround Support

The selection of the surround is closely related to the notion of local and global saliency. For example, by choosing surround as the whole image, the algorithm predicts globally salient regions. For local saliency, finite support or center-weighted kernels can be used.

4 Experimental Results

In this section, we quantitatively evaluate these bottom-up saliency detection algorithms to provide a comparative study.

Dataset and evaluation metric: We show performance comparison on two publicly available datasets: the MSRA dataset [19] and the McGill dataset [17]. For the MSRA dataset, we use a subset of 1,000 images where groundtruth segmentations are available [1]. McGill dataset contains 235 natural images with rough categorization based on difficulty.

To quantitatively evaluate the performance of these saliency detection algorithms, we use binary (thresholded) saliency masks derived from the saliency map and compare them with human segmentation to compute the precision and recall curve.

State-of-the-art saliency detection methods: We conduct a comparative study on the following state-of-the-art methods and show the results in Fig. 1.

- Rarity-based: AIM [4], SUN [27], SW [22] from section 2.2
- Contrast-based: IT [12], GB [9], CA [8], AC [2], FT [1], LC [26] from section 2.3 and HC [5], RC [5] from section 2.4.
- Spectrum-based: SR [10]

Quantitative results: Figure 1 shows the mean precision-recall curves on MSRA and McGill datasets of all methods listed above. (We use two separate plots to avoid confusion.) Several observations can be seen in the comparative experiments. For examples, self-information methods AIM [4] and SUN [27] have similar performance regardless of using image-specific or

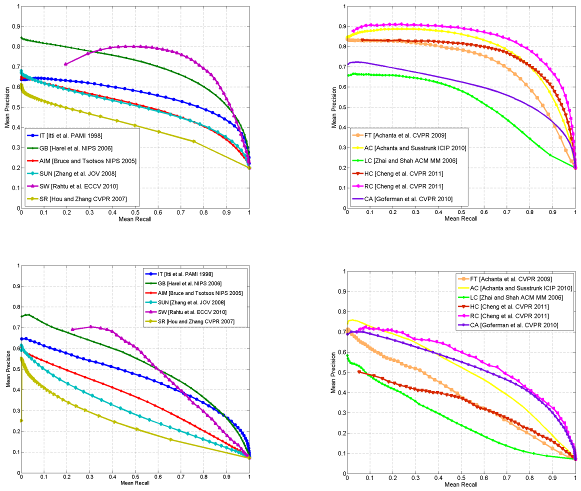

Figure 1. Quantitative comparison on MSRA (first row) and McGill (second row) datasets.

generic statistics of natural images. From LC [26] to FT [1], the performance improves as the feature sets are richer. AC [2] improves upon FT [1] with adaptive surround support. RC [5] improves upon HC [5] via better center support selection.

5 Conclusions

We have shown theoretical connections among various bottom-up saliency detection algorithms. The unified perspective sheds new light on current methods in the literature, by providing a standardized interpretation. We also discuss several center-surround selection strategies. Comparative evaluation on two publicly available datasets brings out the relative strengths of each method.

The list of saliency detection methods in this paper is not exhaustive, e.g., spectrum-based approaches are missing. In the future, we plan to extend our framework to include those methods as well as build a common ground for detailed comparison, for example, by systematically controlling parameters when comparing different divergence measures, features, and center-surround selection schemes.

Acknowledgment

The support of the Office of Naval Research under grant N00014-12-1-0259 is gratefully acknowledged.

References

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

[2] R. Achanta and S. Susstrunk. Saliency detection using maximum symmetric surround. In *ICIP*, 2010.

[3] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *PAMI*, 2012.

[4] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, 2005.

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

[6] T. M. Cover and J. A. Thomas. *Elements of information theory*. 2006.

[7] D. Gao, V. Mahadevan, and N. Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *NIPS*, 2007.

[8] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, 2010.

[9] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, 2006.

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

[11] L. Itti and C. Koch. Computational modeling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

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

[13] R. Jenssen, J. Principe, D. Erdogmus, and T. Eltoft. The cauchy-schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

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

[15] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, 45(2):83–105, 2001.

[16] D. A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. In *ICCV*, 2011.

[17] J. Li, M. Levine, X. An, and H. He. Saliency detection based on frequency and spatial domain analyses. In *BMVC*, 2011.

[18] F. Liu and M. Gleicher. Region enhanced scale-invariant saliency detection. In *ICME*, 2006.

[19] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H. Y. Shum. Learning to detect a salient object. In *CVPR*, 2007.

[20] Y. Ma and H. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM MM*, 2003.

[21] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson. Top-down control of visual attention in object detection. In *ICIP*, 2003.

[22] E. Rahtu, J. Kannala, M. Salo, and J. Heikkil. Segmenting salient objects from images and videos. In *ECCV*, 2010.

[23] M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *JMLR*, 12:731–817, 2011.

[24] A. Toet. Computational versus psychophysical image saliency: A comparative evaluation study. *PAMI*, 99(1), 2011.

[25] R. Wurtz et al. Visual receptive fields of striate cortex neurons in awake monkeys. *J Neurophysiol*, 32(5):727–742, 1969.

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

[27] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20, 2008.