Salient Pattern Detection using W_2 on Multivariate Normal Distributions

Dominik Alexander Klein and Simone Frintrop

Department of Computer Science III, University of Bonn, Germany

Abstract. Saliency is an attribute that is not included in an object itself, but arises from complex relations to the scene. Common belief in neuroscience is that objects are eye-catching if they exhibit an anomaly in some basic feature of human perception. This enables detection of object-like structures without prior knowledge. In this paper, we introduce an approach that models these object-to-scene relations based on probability theory. We rely on the conventional structure of cognitive visual attention systems, measuring saliency by local center to surround differences on several basic feature cues and multiple scales, but innovate how to model appearance and to quantify differences. Therefore, we propose an efficient procedure to compute ML-estimates for (multivariate) normal distributions of local feature statistics. Reducing feature statistics to Gaussians facilitates a closed-form solution for the W_2-distance (Wasserstein metric based on the Euclidean norm) between a center and a surround distribution. On a widely used benchmark for salient object detection, our approach, named CoDi-Saliency (for Continuous Distributions), outperformed nine state-of-the-art saliency detectors in terms of precision and recall.

1 Introduction

The detection of salient objects that visually stand out from their surrounding and automatically attract attention has been intensely investigated during the last decade. It is of interest not only from a psychological perspective, but also from a computational one. Finding salient regions in images supports applications such as general object detection and segmentation in web images [1, 15], or steering a robots eyes and head [18, 19]. The proposed algorithm not only convincingly fulfills its main purpose to quantify saliency, but does this at low computational costs. We integrated efficient and innovative solutions for the most critical parts of saliency systems based on feature statistics: local estimation of distributions and calculation of their contrast. Therefore, CoDi-Saliency is especially suited for such large scale offline application or online usage on restricted mobile platforms.

Many computational approaches have been presented during the last two decades that compute visual saliency, ranging from the well-known Itti-Koch model [11] to approaches that learn optimal feature combinations with machine

learning techniques [15]. A survey on computational attention systems that determine saliency can be found in [5].

Since saliency is intrinsically based on the difference of a region with respect to its surround, it is clear that the computation of a feature (e.g. colors, gradients, entropy, etc.) is per se not sufficient to determine saliency. An example is a single bird that is salient in front of a blue sky but not among a swarm of other birds. Instead, computing the difference of some feature qualities in a region and in its surround is essential [3, 14, 12]. Most approaches determine the center-surround contrast by DoG-filters or approximations of these [11, 4]. Recently, some groups have represented the center and surround area by feature distributions to capture more information about the area [3, 6, 14, 12]. These approaches use discrete distributions in the form of histograms to represent the occurrences of features in an image patch. In contrast to this, we represent feature statistics by multivariate normal distributions that are compared with the Wasserstein distance based on the Euclidean norm. This metric is a well-known method to compare probability distributions and, in contrast to methods such as KLD, considers also the distance of feature entries. This is especially useful for computing saliency since there the similarity of feature values is an essential aspect.

This mathematically well-founded way to compute the saliency of a feature dimension is integrated into a complete framework that is based on findings from neuroscience and psychophysics. It computes several feature cues on multiple scales and finally fuses their conspicuities into a single saliency map. In contrast to most other saliency computation methods, our approach outputs fine-grained saliency maps in which the complete salient objects stand out. We show that our approach outperforms nine state-of-the-art saliency detectors in a segmentation task on the MSRA salient object database [15]. In addition, we show the biological validity of our approach on psychophysical test patterns.

2 The Saliency Model

The structure of our saliency system complies with the architecture of approved psychological visual attention models like those of Treisman and Gelade [20] or Wolfe [22]. Basic features of the human attention system [23] are processed independently from each other. Anomalous appearances of a feature with respect to surroundings are emphasized, resulting in one map of perceptional conspicuity per basic feature. Then, individual conspicuities are fused into a conjoint saliency map.

In CoDi, basic features are investigated in a multi-scale approach, utilizing a difference-of-Gaussian pyramid representation of the input image constructed as in [16]. We implemented the basic features of intensity and color, nevertheless it is possible to adopt the computational methods presented in this paper to further basic features as well. We express local feature occurrences by means of normal distributions. For each point in the image (scale-)space, two normal distributions are estimated: one characterizing the feature appearance closely centered around

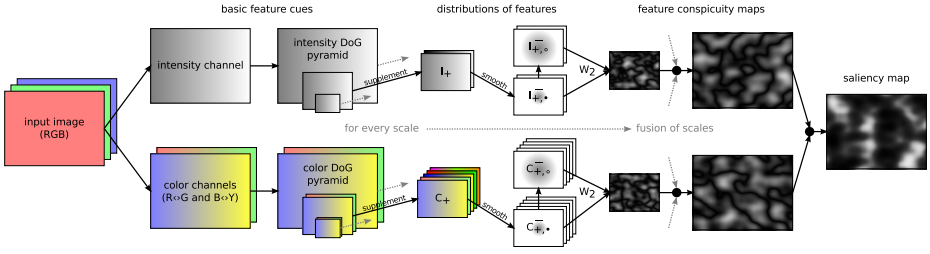

Fig. 1. Schematic intra-system view of the CoDi-Saliency computation.

the point, the other incorporating appearance of a wider surround. Then, the W_2-distance between those distributions is used as conspicuity measure determining the local center-surround contrasts. Figure 1 shows a flowchart of our system.

2.1 Basic Feature Cues

In a first step, the input image is transformed from RGB into a simple, but more psychologically motivated color space following the opponent-process theory [10]. From every pixel, the intensity and color features are computed as

$$\mathrm{I}(x,y) = \left(\frac{R+G+B}{3}\right)_{(x,y)} \quad \text{and} \quad \mathrm{C}(x,y) = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} R-G \\ B-\frac{R+G}{2} \end{pmatrix}_{(x,y)}. \quad (1)$$

From this image in opponent color space, a difference-of-Gaussian pyramid is computed. That way, we achieve a scale-separated representation $\mathrm{I}(x,y;t)$ of one-dimensional intensity feature and $\mathrm{C}(x,y;t)$ of two-dimensional color feature consisting of a red-green and blue-yellow contrast dimension. In theory, it would be best to use a perceptually normalized color space like CIELAB[1], but this requires additional knowledge about the illuminant, which is not given but varies heavily in image collections from unknown sources. Constant use of D_{65} standard illuminant factors for midday sunlight would probably introduce similar inaccuracies as using the proposed, efficiently computable space.

2.2 Local Feature Statistics

In our framework, local feature statistics are summarized by one-dimensional normal distributions for intensity, respectively two-dimensional distributions for color. To facilitate an efficient maximum likelihood estimation of the normal distribution parameters, supplementing layers are added to the feature maps, resulting in

$$\mathrm{I}_+(x,y;t) = \begin{pmatrix} i \\ i^2 \end{pmatrix}_{(x,y;t)} \quad \text{and} \quad \mathrm{C}_+(x,y;t) = (c_1, c_2, c_1^2, c_2^2, c_1 c_2)^{\mathrm{T}}_{(x,y;t)}. \quad (2)$$

[1] http://www.hunterlab.com/appnotes/an07_96a.pdf

Local occurrences of a feature are treated as weighted samples for the estimation process. The weights are determined by a Gaussian integration window (cf. Fig. 2), implemented by discrete convolution of the feature maps

$$\begin{aligned}\bar{\mathrm{I}}_+(x,y;t) &= \left(\mathrm{g}\left(\sigma^2\right) * \mathrm{I}_+(t)\right)(x,y)\\ \bar{\mathrm{C}}_+(x,y;t) &= \left(\mathrm{g}\left(\sigma^2\right) * \mathrm{C}_+(t)\right)(x,y)\,.\end{aligned} \tag{3}$$

Sophisticated optimizations and approximations are essential to be applied in this step to achieve competitive performance: For Gaussians of a small standard deviation, separability of the Gaussian kernel in x- and y-dimension is exploited, resulting in run-time complexity $\mathcal{O}(\sigma n)$. For those of bigger σ, based on the central limit theorem, a Gaussian filter is approximated by repeated smoothing with b box-filters, choosing their width and height so that the result is a b^{th}-order approximation of a Gaussian of slightly smaller σ_{box}, inspired by [13]. The standard deviation of a box filter of extent w equals $\sqrt{(w^2-1)/12}$. Thus, applying b iterations, a filter of size $w_{\mathrm{ideal}} = \sqrt{(12\sigma^2)/b + 1}$ would provide the best approximation of this Gaussian. However, since $w \in \mathbb{N}$, we choose the next smaller and larger odd filter sizes

$$w_1 = 2\left\lfloor 0.5\sqrt{(12\sigma^2)/b + 1}\right\rfloor - 1 \quad \text{and} \quad w_2 = w_1 + 2 \tag{4}$$

instead. The closest approximation we can get showing a standard deviation lower than σ, is achieved by

$$m_1 = \left\lceil \frac{12\sigma^2 - bw_1^2 - 4bw_1 - 3b}{-4w_1 - 4}\right\rceil \quad \text{and} \quad m_2 = b - m_1 \tag{5}$$

repeated rounds of box-filtering with extents w_1 and w_2, respectively. This process results in an overall standard deviation of

$$\sigma_{\mathrm{box}} = \sqrt{(m_1 w_1^2 + m_2 w_2^2 - b)/12}. \tag{6}$$

The small defect of $\sigma_\Delta = \sqrt{\sigma^2 - \sigma_{\mathrm{box}}^2}$ is then smoothed as described before, resulting in overall run-time complexity of $\mathcal{O}(\sigma_\Delta n + bn)$. Furthermore, one can apply stepwise smoothing when computing center ($\bullet$) and surround ($\odot$) statistics, since w.l.o.g. for $\sigma_{\mathrm{s}} \geq \sigma_{\mathrm{c}}$

$$\bar{\mathrm{I}}_{+,\odot} = \mathrm{g}\left(\sigma_{\mathrm{s}}^2\right) * \mathrm{I}_+ = \mathrm{g}\left(\sigma_{\mathrm{s}}^2 - \sigma_{\mathrm{c}}^2\right) * \mathrm{g}\left(\sigma_{\mathrm{c}}^2\right) * \mathrm{I}_+ = \mathrm{g}\left(\sigma_{\mathrm{s}}^2 - \sigma_{\mathrm{c}}^2\right) * \bar{\mathrm{I}}_{+,\bullet} \tag{7}$$

and the same holds true for color.

After applying this local weighting of feature samples by means of smoothing the annotated feature maps, one can easily compute a center and a surround ML-estimate of normal distribution parameters with help of the images $\bar{\mathrm{I}}_{+,\{\bullet,\odot\}}$ and $\bar{\mathrm{C}}_{+,\{\bullet,\odot\}}$ utilizing the relations

$$\hat{\mu}_{\mathrm{I}} = \bar{i},\ \hat{\sigma}_{\mathrm{I}} = \overline{i^2} - \bar{i}^2,\ \hat{\mu}_{\mathrm{C}} = \begin{pmatrix}\bar{c}_1\\ \bar{c}_2\end{pmatrix}, \text{ and } \quad \hat{\Sigma}_{\mathrm{C}} = \begin{pmatrix}\overline{c_1^2} - \bar{c}_1^2 & \overline{c_1 c_2} - \bar{c}_1\bar{c}_2\\ \overline{c_1 c_2} - \bar{c}_1\bar{c}_2 & \overline{c_2^2} - \bar{c}_2^2\end{pmatrix} \tag{8}$$

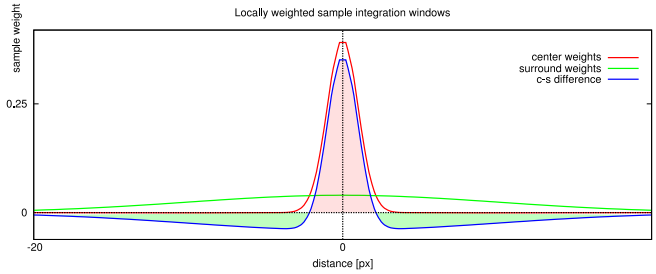

Fig. 2. Center and surround integration windows of feature samples are Gaussians of different standard deviations. Especially the large surround integration window requires fast approximation algorithms.

for every pixel and scale. Note that this computation scheme could also be applied in a similar way for multivariate normal distributions of more than two dimensions.

2.3 Center-Surround Difference of Feature Statistics

In the last section, we explained how to efficiently compute (multivariate) normal distributions of local, basic feature occurrences, so we can assume a center $P_\bullet$ and a surround distribution $P_\odot$ to be given for every pixel and scale. The next step is to determine how much the center appearance sticks out of the surround, in other words how different those two distributions are. Here, a plausible distance measure should not only score the similarity of probabilities between same feature manifestations, but also take into account the visual difference between manifestations. A white region within a black image is clearly more conspicuous than a gray one. The W_2-distance on the Euclidean norm, which is defined as

$$W_2(P_\bullet, P_\odot) = \left[\inf_{\gamma \in \Gamma(P_\bullet, P_\odot)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|_2 \, \mathrm{d}\gamma(x, y) \right]^{\frac{1}{2}} \tag{9}$$

with $\Gamma(P_\bullet, P_\odot)$ denoting the set of all couplings of $P_\bullet$ and $P_\odot$, meets this requirements, if the underlying feature space is defined reasonably. Most often it is imagined as a transport of mass problem: how much (probability) mass needs to be moved how far (wrt. Euclidean distance) to transform one probability density function into the other. For example, in computer vision the W_1-distance on discrete random variables is also referred to as earth mover's distance (EMD).

It would be intractable to evaluate the integral in equation 9 in case of arbitrary distributions. Thankfully, it can be solved algebraically for multivariate normal distributions, as established by Givens and Shortt [7], resulting in the closed form expression

$$W_2(P_\bullet, P_\odot) = \left[\|\mu_\bullet - \mu_\odot\|_2^2 + \operatorname{tr}(\Sigma_\bullet) + \operatorname{tr}(\Sigma_\odot) - 2 \operatorname{tr}\left(\sqrt{\Sigma_\bullet^{\frac{1}{2}} \Sigma_\odot \Sigma_\bullet^{\frac{1}{2}}} \right) \right]^{\frac{1}{2}}. \tag{10}$$

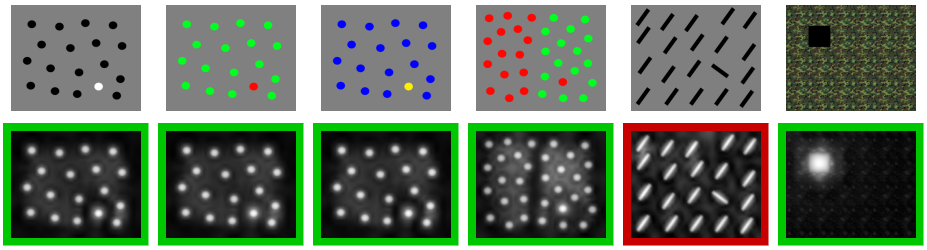

Fig. 3. Saliency maps of our system on psychophysical test patterns provided by Klein and Frintrop [12].

Applying equations 8 and 10 to our basic features $\bar{\mathrm{I}}_{+,\{\bullet,\odot\}}(x,y;t)$ and $\bar{\mathrm{C}}_{+,\{\bullet,\odot\}}(x,y;t)$ yields pyramids of intensity $\mathcal{I}(x,y;t)$ as well as color $\mathcal{C}(x,y;t)$ conspicuities.

2.4 Fusion of Scales and Feature Modalities

For every feature, the arithmetic mean over normalized scales is calculated as

$$\mathcal{I}(x,y) = \frac{1}{s} \bigoplus_{t=1}^{s} \sqrt{t}\,\mathcal{I}(x,y;t) \quad \text{and} \quad \mathcal{C}(x,y) = \frac{1}{s} \bigoplus_{t=1}^{s} \sqrt{t}\,\mathcal{C}(x,y;t) \tag{11}$$

with $\oplus$ denoting a rescale and add-per-pixel operator. Finally, the saliency map is given by the arithmetic mean across feature modalities

$$\mathcal{S}(x,y) = \frac{1}{2}\left(\mathcal{I}(x,y) + \mathcal{C}(x,y)\right). \tag{12}$$

3 Evaluation

We compared our saliency model, CoDi, with nine state-of-the-art saliency models: the iNVT by Itti et al. [11], the Saliency Toolbox (ST) [21], two systems of Hou and Zhang (HZ07,HZ08) [8, 9], the AIM model of Bruce and Tsotsos [3], the system of Ma and Zhang (MZ) [17], two versions of Achanta et al. (AC09,AC10) [1, 2], and the BITS system of Klein and Frintrop [12].

The evaluation was done first on the psychophysical test patterns used in [12] (Sec. 3.1) and second on the MSRA database of salient objects [1] (Sec. 3.2).

3.1 Psychophysical Soundness

The purpose of a saliency model usually is to mimic human behavior, thus it is crucial to obey the findings of neuroscience about the human attention mechanism. There, attention was studied testing the human ability to immediately detect outliers in so called "pop-out" images. A conform computational saliency model should likewise output the maximum saliency at the positions of such pop-outs.

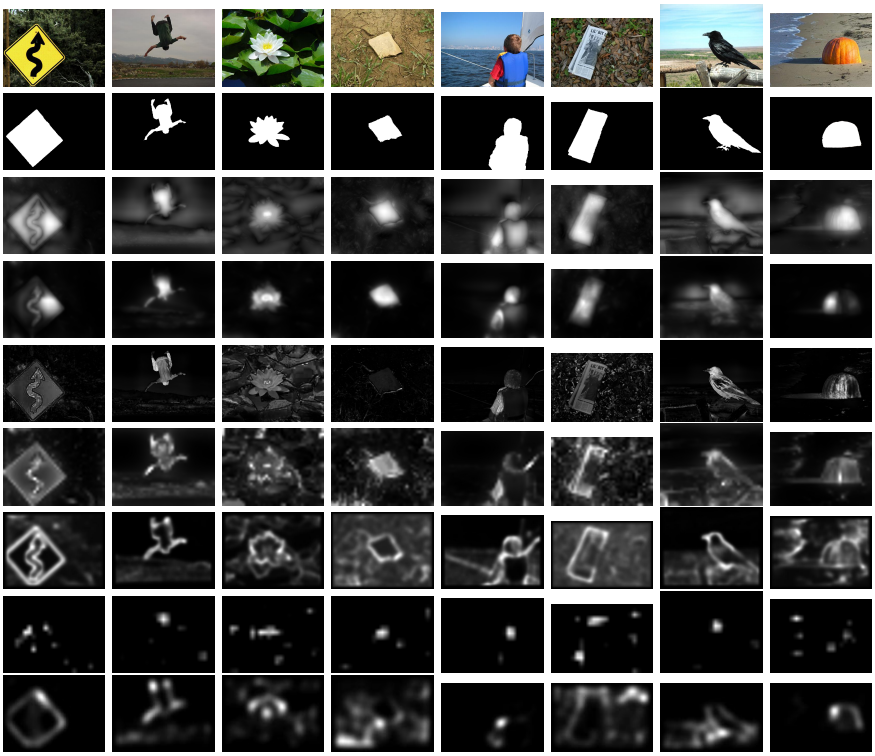

Fig. 4. Comparison of saliency maps on natural images from the MSRA dataset [15]. First row shows the original images, second row the corresponding ground truth. Next, results of the following saliency methods are listed from top to bottom: our approach CoDi, BITS [12], AC10 [2], ST [21], AIM [3], iNVT [11], HZ08 [9].

Klein and Frintrop [12] introduced suitable test patterns and compared several approaches. Only the BITS system successfully passed all tests, the others ([11, 21, 1, 2, 9, 3]) each failed in at least two of the patterns. The CoDi-Saliency detector passes every pattern but the orientation (cf. Fig. 3), since this basic feature is not integrated. However, since the proposed framework is very generic, it should be possible to add further feature cues to our system.

3.2 Salient Object Detection

A quantitative performance analysis of our algorithm was done on the subset of 1000 images out of the MSRA salient object database [15] that was first used in [1]. The images contain objects that have been consistently marked as salient by several subjects. Pixel-precise binary maps are available that contain the ground truth shapes of the objects. Figure 4 shows exemplary results of the saliency maps of different saliency detectors with available source code.

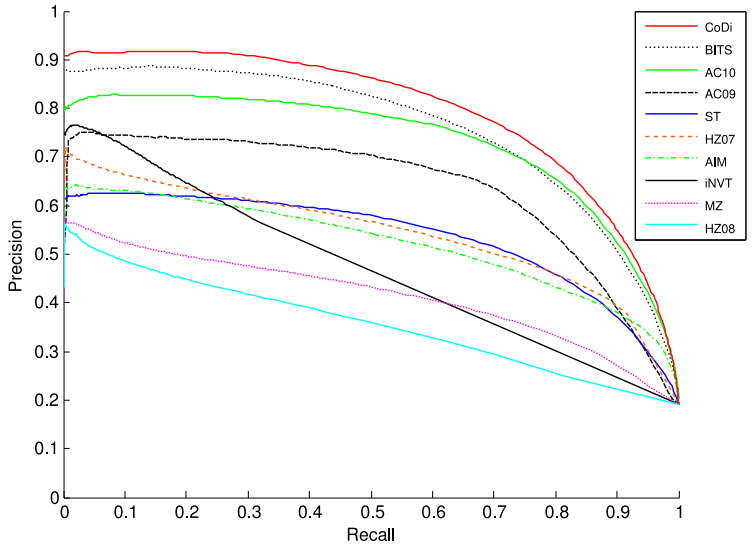

Fig. 5. Precision-recall curves for the salient object dataset of 1000 images from [1]. We compared our system CoDi against nine other saliency detectors.

The quantitative experiments were conducted as in [1]: all saliency maps are binarized by thresholding the intensity values between $[0, 255]$. Thereby, one achieves 256 possible segmentations for the dataset. Then, each is matched against the ground truth binary masks to obtain precision and recall. Finally, results are plotted together in the graph depicted in Figure 5. It can be seen that our approach outperforms the others. The second best, BITS, is also based on local distributions of basic features, but seems to be inferior because of two major points: first, the Kullback-Leibler divergence as a distance between distributions does not consider distances in the feature domain like the Wasserstein metric. Second, feature space discretization in histograms can be problematic, especially for 2D features such as color. Instead, our algorithm works on continuous distributions. The precision value for threshold 255 using our approach is ≈ 0.91 (cf. left boundary of Fig. 5), stating that in more than 9 out of 10 cases the most salient point was located on the object of interest. Thus, it should serve as a good starting point for a subsequent object segmentation algorithm.

With parameters chosen as used for this evaluation, the computation time per image executed by an Intel Core i7-2600 CPU is 82ms on average. It mainly depends on the number and resolution of evaluated scales. If necessary, it can be tuned to even less computation time without disproportionally downgrading the results. Besides that, the framework is well suited for parallelization and should benefit much if one makes use of a modern GPU.

4 Conclusion

We introduced CoDi-Saliency, a new method to compute visual saliency in a probabilistic fashion. The overall framework follows the conventional structure of cognitive visual attention systems, computing the conspicuity for each basic feature cue individually before fusing them to a common saliency map. Local normal distributions of basic features were aggregated and estimated employing an efficient approximation algorithm for Gaussian image convolution on intelligently supplemented feature maps. This enabled the computation of W_2-distance in constant time per pixel.

The presented approach outperformed nine other saliency detection methods on the widely used Achanta subset of the MSRA salient object database. Although systems based on feature distributions and local contrasts are usually slower than those directly using basic features and global contrasts, our method is sufficiently fast to be applied on mobile systems or for large scale datasets.

In future work, we will investigate how the feature distributions computed for saliency can be reused for graph-based segmentation of the object located around the most salient point. Furthermore, we want to enhance the framework with further feature cues such as orientation or symmetry. Additionally, it would be interesting to adopt the system also for other domains, e.g. the prediction of human eye movements.

References

1. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: CVPR (2009)
2. Achanta, R., Süsstrunk, S.: Saliency detection using maximum symmetric surround. In: Proc. of Int'l Conf. on Image Processing (ICIP) (2010)
3. Bruce, N., Tsotsos, J.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision 9(3), 1–24 (2009)
4. Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-directed Search, Lecture Notes in Artificial Intelligence (LNAI), vol. 3899. Springer (2006)
5. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundation: A survey. ACM Trans. on Applied Perception 7(1) (2010)
6. Gao, D., Vasconcelos, N.: Bottom-up saliency is a discriminant process. In: Proc. of ICCV (2007)
7. Givens, C.R., Shortt, R.M.: A class of wasserstein metrics for probability distributions. Michigan Math. J. 31(2) (1984)
8. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: Proc. of CVPR (2007)
9. Hou, X., Zhang, L.: Dynamic visual attention: Searching for coding length increments. In: Advances in Neural Information Processing Systems (2008)
10. Hurvich, L., Jameson, D.: An opponent-process theory of color vision. Psychological review 64(6) (1957)
11. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on PAMI 20(11) (1998)

12. Klein, D.A., Frintrop, S.: Center-surround divergence of feature statistics for salient object detection. In: Proc. of Int'l Conf. on Computer Vision (ICCV) (2011)
13. Kovesi, P.: Arbitrary gaussian filtering with 25 additions and 5 multiplications per pixel. Tech. Rep. UWA-CSSE-09-002, School of Computer Science, University of Western Australia (2009)
14. Lin, Y., Fang, B., Tang, Y.: A computational model for saliency maps by using local entropy. In: Proc. of AAAI (2010)
15. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE Trans. on PAMI (2009)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int'l J. of Computer Vision (IJCV) 60(2), 91–110 (2004)
17. Ma, Y.F., Zhang, H.J.: Contrast-based image attention analysis by using fuzzy growing. In: ACM Int'l Conf. on Multimedia (2003)
18. Ruesch, J., Lopes, M., Bernardino, A., Hörnstein, J., Santos-Victor, J., Pfeifer, R.: Multimodal saliency-based bottom-up attention: A framework for the humanoid robot icub. In: Proc. of Int'l Conf. on Robotics and Automation (ICRA) (2008)
19. Schauerte, B., Kühn, B., Kroschel, K., Stiefelhagen, R.: Multimodal saliency-based attention for object-based scene analysis. In: Proc. of Int'l Conf. on Intelligent Robots and Systems (IROS) (2011)
20. Treisman, A.M., Gelade, G.: A feature integration theory of attention. Cognitive Psychology 12, 97–136 (1980)
21. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks (2006)
22. Wolfe, J.M.: Guided search 2.0: A revised model of visual search. Psychonomic Bulletin and Review 1(2) (1994)
23. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience 5, 1–7 (2004)