Center-surround Divergence of Feature Statistics for Salient Object Detection

Dominik A. Klein and Simone Frintrop
Rheinische Friedrich-Wilhelms Universität Bonn
Institute of Computer Science III, Römerstr. 164, 53117 Bonn
{kleind,frintrop}@iai.uni-bonn.de

Abstract

In this paper, we introduce a new method to detect salient objects in images. The approach is based on the standard structure of cognitive visual attention models, but realizes the computation of saliency in each feature dimension in an information-theoretic way. The method allows a consistent computation of all feature channels and a well-founded fusion of these channels to a saliency map. Our framework enables the computation of arbitrarily scaled features and local center-surround pairs in an efficient manner. We show that our approach outperforms eight state-of-the-art saliency detectors in terms of precision and recall.

1. Introduction

Salient objects have the quality to visually stand out from their surroundings and are likely to attract human attention. A key property that makes an object salient is the visual difference to the background. A polar bear is salient on dark rocks, but almost invisible in snow. The detection of visual saliency is of high interest in many computer vision applications, ranging from general object detection in web images [3], over image thumbnailing [17], to computing a joint focus of attention in human robot interaction [20].

Visual saliency and, more general, visual attention have been widely investigated in neurobiology and psychophysics [18] and many computational models have been built based on such findings [22, 12, 6]. A survey on biologically-inspired attention systems can be found in [7]. Recently, several saliency approaches came up that are based on computational and mathematical ideas and usually less biologically motivated. These approaches range from the computation of entropy [13, 10], over determining features that best discriminate between a target and a null hypothesis [8], to learning the optimal feature combination with machine learning techniques [15, 3].

In this work, we present a new approach to compute visual saliency that combines the general structure of psychological attention models [21, 25] with a sound mathematical foundation, and additionally enables an efficient computational implementation. We define the saliency of an image region in an information-theoretic way by means of the Kullback-Leibler-Divergence (KLD). For a center and a surround region, we estimate the distributions of visual feature occurrences. Then, the KLD between these distributions expresses how much more capacity one can expect to require when events following the center distribution are coded according to the surround distribution. In other words, KLD measures how much the feature statistics in the center diverge from those in the surround.

This formulation of saliency has two advantages. First, it allows a consistent computation for all feature channels, in contrast to approaches that apply different feature extraction methods for each channel [15, 3]. Second and more important, it allows a well-founded fusion of feature channels. While absolute values of such channels quantify miscellaneous properties that are not necessarily unifiable in a straight-forward way, KLD abstracts them to a common entity. Additionally, we incorporate an efficient scale-space computation of center-surround pairs of arbitrary sizes.

We evaluate our approach on a standard benchmark database of salient objects [1] and compare the results with eight state of the art saliency detectors. It shows that our approach outperforms all other methods in terms of precision and recall. Our method shows its strength especially for small objects, for which good precision values are usually more difficult to obtain.

2. State of the Art

The concept of visual saliency comes from human perception and correlates with the ability of a region to attract attention [18]. While human attention can be attracted by bottom-up, data-driven as well as by top-down, knowledge-driven factors, saliency is associated with bottom-up attention that automatically attracts the human gaze.

Bottom-up attention has been widely studied in cognitive fields. A basis for many computational attention models are the Feature Integration Theory (FIT) [21] and the Guided Search model [25]. The FIT has introduced the structure

that still serves as basis for many computational attention systems: several feature channels (e.g. color or orientation), each divided into several feature types (e.g. red, yellow, horizontal, vertical), are investigated in parallel. Finally, the conspicuities are collected in a *master map of attention*. In later works, this map has been called *saliency map*.

Many computational models have been built according to this structure [12, 6, 24], among them one of the most popular systems, the iNVT of Itti et al. [12]. While these systems have obtained good results in simulating human eye movements and in applications ranging from object recognition to robotics [7], one problem is that the fusion of feature channels with per se not comparable properties is usually somewhat arbitrary.

During the last decade, several approaches came up to model saliency with computational and mathematical methods that are mostly less biologically motivated. Kadir and Brady have introduced entropy-based saliency [13]. More recently, Hou and Zhang have computed the incremental coding length to measure the perspective entropy gain [10]. Entropy-based methods generally capture image regions with a lot of structure, which corresponds often but not always to salient regions. A problem occurs if the absence of structure makes an item salient, such as a person wearing white clothes in the jungle (cf. last row of Fig. 4).

Ma and Zhang have proposed a contrast-based method that uses fuzzy growing to extract regions from their saliency map [16]. Achanta et al. have introduced a simple approach that determines the difference of pixels to the average color and intensity value of the image [1, 2]. While their system has problems to detect saliencies for several classical pop-out experiments (cf. Sec. 4.1), it is fast and simple to implement.

Some groups have investigated alternative ways to compute saliency by applying different computer vision methods to obtain feature channels, which are finally fused by machine learning techniques. Liu et al. combined multi-scale contrast, center-surround histograms, and color spatial-distributions with conditional random fields [15]. Alexe et al. combined multi-scale saliency, color contrast, edge density, and superpixels in a Bayesian framework [3].

Information theory also has entered the field of saliency detection. Itti and Baldi have computed temporal saliency based on a Baysian notion of surprise [11]. Gao et al. have presented a decision-theoretic approach based on mutual information [8] and Chen has computed the co-saliency of two objects in different images with the KLD [5].

Bruce and Tsotsos have presented an interesting approach that computes the self-information of image regions with respect to their surround [4]. There are some parallels of this work to our approach. The differences are that while they base their feature detection on ICA coefficients that are learned from a large variety of images, we have specifically designed scalable feature detectors to represent the distributions in different feature channels. This enables us to compute features on any scale in a computationally feasible way and disengages us from the need of a training set. Furthermore, we compute the KLD instead of the self-information, apply local instead of global surround regions, and compute the saliencies on several scales.

Figure 1. Center-surround filter based on the Kullback-Leibler Divergence (KLD).

3. The Saliency Model

The main structure of our saliency system, called BITS (Bonn Information-Theoretic Saliency model), is based on the general layout of psychological attention models like the ones in [21, 25]: several feature channels are investigated in parallel and the conspicuities are fused to a single saliency map. The feature channels intensity, color, and orientation have been chosen since they belong to the basic features of the human attention system [26].[1]

The saliency computation itself is rather computationally than biologically motivated and consists of two steps. First, basic features analyze the occurrence of certain intensities, colors, and orientations on different scales. In a second step, the center-surround contrast is determined in an information-theoretic way. Two distributions of visual feature occurrences are determined for a center and a surround region and the Kullback-Leibler Divergence determines the difference between these distributions (cf. Fig. 1). An overview of the system is depicted in Fig. 2.

3.1. Basic Feature Cues

For our visual saliency system, we model the basic features of color, orientation, and intensity. From an input image in the HSL color space, integral layers are built in order to quickly compute pyramid representations from our scalable basic features. These integral layers enable the calculation of summed and averaged values of arbitrary sized rectangular regions in constant time [23].

The intensity feature is the average of the lightness layer within a rectangle of a certain scale. The color feature is also the average of a rectangular region, but a little trickier to compute in order to account for the saturation of the occurring colors. Hue and saturation that represent polar co-

[1] Another important feature is motion, but because here we concentrate on saliency detection in web images, motion is not required.

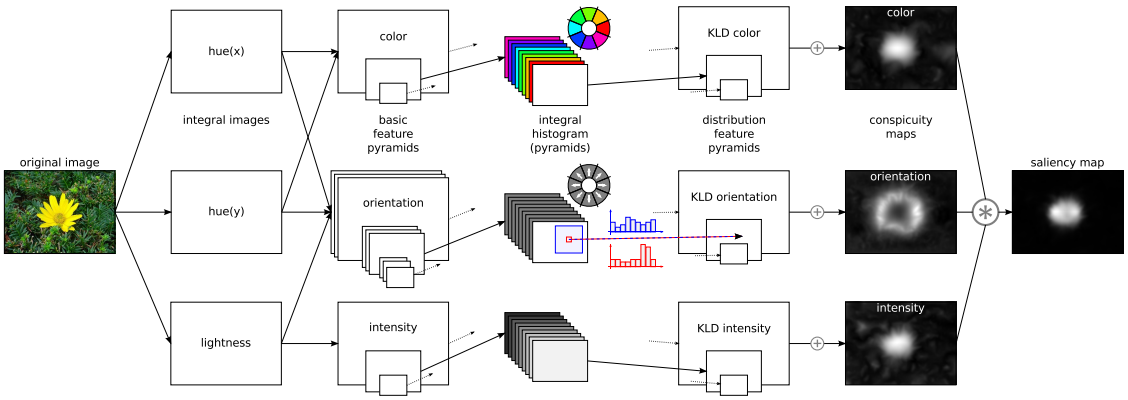

Figure 2. Schematic overview of our saliency system BITS.

ordinates in the HSL color space are converted into Cartesian coordinates, referred to as hue(x)- and hue(y)-layers (cf. Fig. 2). From those, average colors can be computed via integral layers, before the representation is transformed back into hue/saturation. The orientation feature computes partial derivatives from region-wise averages to determine the gradient direction on a given scale as in [14]. We apply the orientation feature separately for lightness-, hue(x)- and hue(y)-layer, because orientation should be observable from intensity as well as color contrasts. We compute pyramids of eight different feature scales in steps of factor $\sqrt{2}$. For this, we do not need to scale the image data, but the feature size, which is an advantage in terms of speed. The feature sampling rate depends on their scales, larger features are more coarsely sampled than smaller ones.

3.2. Center-Surround Distribution Feature based on Information Theory

Information theory is an area of statistics that is used to analyze signals and their transmission over channels. One of the main concepts is the notion of entropy, which quantifies the expected value of information that a signal of a given coding scheme contains. A coding scheme equates to a probability distribution over the occurrence of certain messages. The less predictable the occurrence of a message is, the higher the entropy. For instance the entropy of a uniform distribution is highest, while if one can predict the next message for sure, the entropy is zero.

As mentioned above, the difference of a region to its surroundings is essential to obtain visual saliency. One can convey this principle of difference to information theory by using the Kullback-Leibler Divergence,

$$\mathcal{D}_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x. \qquad (1)$$

KLD is a measure between two probability distributions P and Q, that meters the expected value how much longer a message must be to express events from P based on Q. The more P differs from Q, the higher the KLD.

For each pyramid layer of basic feature results, our contrast feature is computed. This feature is based on KLD and integrates a local center-surround mechanism to rate the conspicuity of a region (cf. Fig. 1). Here, the information-theoretic notion of a message is the parameter value of a visual feature. Therefore, we need to estimate local distributions of basic feature results: we split every basic feature layer into an integral histogram [19]. An integral histogram consists of layers that count the summed number of values top and left from a pixel that fall into a certain histogram bin. Here, sums of relative distances of the values on the corresponding basic feature map to centered values of the neighboring bins are counted. Thus, one can obtain bilinearly interpolated histograms for rectangular regions. For the periodic color and orientation feature, we build radial histograms with an additional center bin. In case of color, the saturation determines how much a feature sample counts for the center or a radial bin. For orientation and gradient magnitude we proceed correspondingly. This allows to contrast the absence or occurrence of orientation and color in the center with the surround distribution. Utilizing these integral histograms, we calculate the discrete KLD feature

$$D_{\mathrm{KL}}(C\|S) = \sum_{i=1}^{\mathrm{b}} C(i) \log \frac{C(i)}{S(i)}, \qquad (2)$$

in constant time, where b denotes the number of bins (here: 13), and C and S are distributions of center and surround regions with size ratios of 0.2 and 0.3. Increasing the number of bins did not substantially increase the quality of the system. The KLD feature maps are scale normalized corresponding to the ratio of the feature's surround region to the image size. Then they are rescaled and added per pixel on highest resolution to form one conspicuity map per channel. Fusion of conspicuity maps into a single saliency map is done by per element multiplication. This results in a fusion exponential proportional to geometric mean, but we omit calculation of the n^{th} root.

4. Experiments and Results

We evaluated our saliency method on two kinds of data: psychological patterns (Sec. 4.1) and a database of salient objects [1] (Sec. 4.2). On both data sets, we compared our approach with eight state-of-the-art saliency models: the iNVT by Itti et al. [12], the Saliency Toolbox (ST) [24], two systems of Hou and Zhang (HZ07,HZ08) [9, 10], the AIM model of Bruce and Tsotsos [4], the system of Ma and Zhang (MZ) [16], and two versions of Achanta et al. (AC09,AC10) [1, 2]. For iNVT, ST, HZ08, AIM, AC09 and AC10 we used the code from the authors' web pages. For HZ07 and MZ we used the saliency maps provided online[2].

4.1. Psychological Patterns

Detecting outliers in "pop-out" images is an essential step for a saliency model, since the results clearly show the strengths and limitations of an approach. We designed intensity and color patterns with a gray background and items with the same intensity contrasts to the background. This allows to make sure that saliency really results from an item-item contrast and not from an item-background contrast.

Fig. 4 shows the results on these patterns for all saliency methods with available source code. Saliency maps that have their most salient region on the outlier are marked with a green bounding box, others with a red one. Except for our model, none of the models was able to detect all outliers. Some results can be explained by the system design: Achanta cannot detect orientation pop-outs, since it is a purely color/intensity based approach. AIM and AC09 cannot detect local pop-outs (row 4), since they use a global instead of local surround. The result of Hou (last row) shows that purely using entropy to compute saliency is not always sufficient: the uniform square on a textured background is not considered salient, since it shows low entropy compared to the high entropy of the background. On the other hand, the non-salient region in the result of AIM is due to the filter size and could be avoided by a scale-space extension.

4.2. Salient Object Database

Additionally, we performed quantitative experiments on the image set that was used in [1, 2]. It is a database of 1000 images, which is a subset of the MSRA salient object database [15]. The latter contains objects that were marked as salient by 2 out of 3 users. For the 1000 image subset, binary maps are available that show accurate contours of the salient objects. Fig. 5 shows some images of this database and the corresponding saliency maps for the saliency methods with available source code.

The saliency maps were evaluated according to [1]. A binary map was obtained from the saliency map by varying a threshold on the intensity values $[0, 255]$. Each of these 256 maps was compared to the ground truth binary map from the database and precision and recall were computed. This resulted in the precision-recall curves shown in Fig. 3, top. It should be noted that some of the methods (e.g., iNVT, ST, AIM) are designed rather for simulating human eye movements than for the detection of salient objects in web images. Therefore, these results should be regarded with caution. We have included them for completeness.

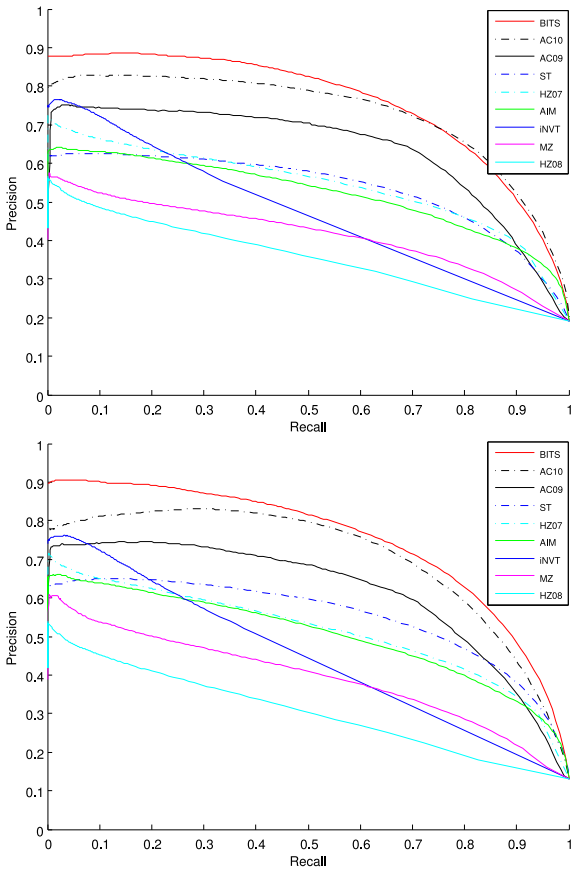

Figure 3. Precision-recall curves for the saliency maps of our system BITS and 8 other saliency detectors on the dataset of 1000 images from [1] (top) and of a subset of small objects (max. 20% of image) (bottom). See text for details.

As already pointed out in [15], obtaining high precision-recall values for images with large objects is not too difficult: if an object occupies 80% of the image, an algorithm that selects the whole image obtains 80% precision with 100% recall. Thus, it is more challenging to obtain high precision-recall curves for small objects. To test this, we determined a subset of the database containing small objects, similarly as in [15]. We selected 549 images with objects occupying at most 20% of the image area. The resulting precision-recall curves are shown in Fig. 3, bottom. Here it shows more clearly that our approach outperforms the other methods.

[2] http://ivrg.epfl.ch/supplementary_material/RK_CVPR09

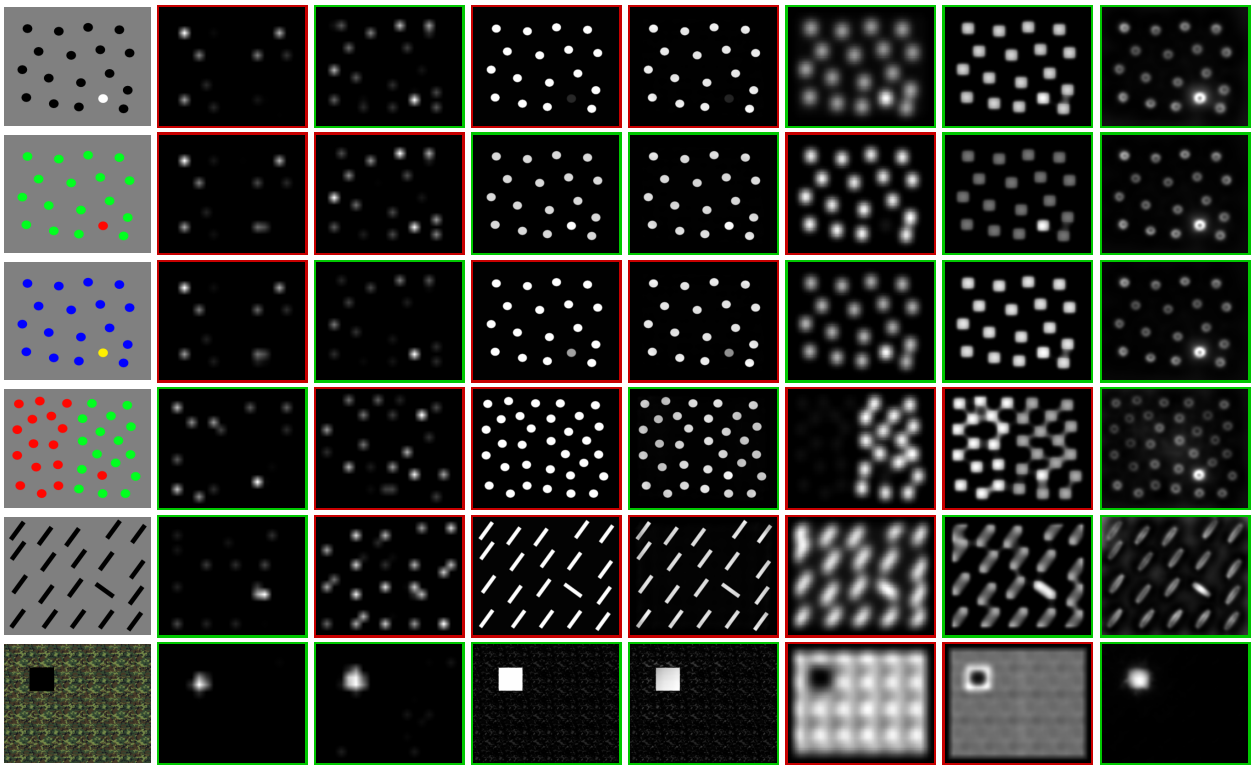

Figure 4. Comparison of saliency maps on psychological patterns. Saliency methods from left to right: iNVT [12], ST [24], AC09 [1], AC10, [2], HZ08 [10], AIM [4], our approach BITS. Green bounding boxes: outlier detected; red boxes: failure.

5. Conclusion

We presented a new approach to compute visual saliency in an information-theoretic way. By means of the Kullback-Leibler Divergence, we determine the contrast of the center and the surround distribution of features for the dimensions intensity, color, and orientation. This enables a well-founded fusion of channels based on a common entity. We have shown that the new approach outperforms eight other saliency computation methods, especially for small objects.

Since information-theoretic approaches are based on feature distributions, the computation is intrinsically more computationally expensive than the classical area-based center-surround filters. To obtain a system that is applicable in reasonable time, calculations are often restricted. For example, AIM uses a center patch with a fixed size and one global surround distribution that covers the complete image.

However, since the detection of salient structures relies essentially on center-surround pairs of different sizes, it is important to integrate scalable feature computations in a computationally still feasible way. Especially for applications on large image databases or on mobile robots, real-time performance is an essential requirement. With our integral image based framework, we found a good compromise between accuracy and speed. With less than 0.5 sec (320×240 pixel image, 2.66 GHz quad-core PC using double precision computations) the system is close to real-time performance. Since the code is not yet optimized, we are confident to obtain real-time performance easily by standard optimizations and/or more extensive parallelization.

Systems as the proposed one always include many parameters and design choices. The parameters used here have shown to be reasonable for the detection of salient objects in web images. We tested the approach also on other images, and it shows to be quite stable and not strongly dependent on parameter choices. Our combinations of center-surround pairs enable the detection of a wide range of sizes of salient regions. Nevertheless, for other applications such as modeling human eye movements, the parameters might have to be adapted to yield optimal performance.

References

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

[2] R. Achanta and S. Süsstrunk. Saliency detection using maximum symmetric surround. In *Proc. of Int'l Conf. on Image Processing (ICIP)*, 2010.

[3] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *Proc. CVPR*, 2010.

[4] N. Bruce and J. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24, 2009.

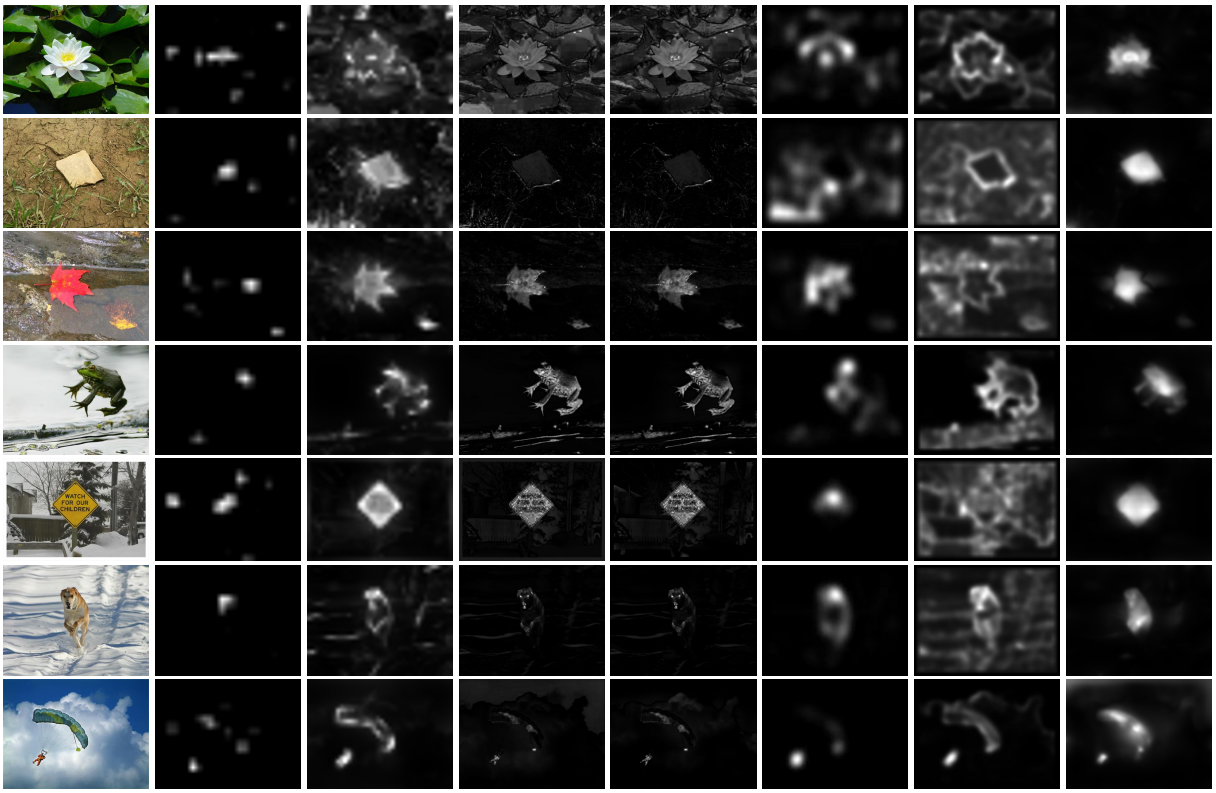

Figure 5. Comparison of saliency maps on natural images from the MSRA dataset [15]. Saliency methods from left to right: iNVT [12], ST [24], AC09 [1], AC10, [2], HZ08 [10], AIM [4], our approach BITS.

[5] H.-T. Chen. Preattentive co-saliency detection. In *Proc. of ICIP*, 2010.

[6] S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*, volume 3899 of *Lecture Notes in Artificial Intelligence (LNAI)*. Springer, 2006.

[7] S. Frintrop, E. Rome, and H. I. Christensen. Computational visual attention systems and their cognitive foundation: A survey. *ACM Trans. on Applied Perception*, 7(1), 2010.

[8] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. In *Proc. of ICCV*, 2007.

[9] X. Hou and L. Zhang. Saliency detection: a spectral residual approach. In *Proc. of CVPR*, 2007.

[10] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *Advances in Neural Information Processing Systems*, 2008.

[11] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009.

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

[13] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Proc. of ECCV*, 2004.

[14] D. Klein and A. Cremers. Boosting scalable gradient features for adaptive real-time tracking. In *Proc. of ICRA*, 2011.

[15] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Trans. on PAMI*, 2009.

[16] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM Int'l Conf. on Multimedia*, 2003.

[17] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Proc. of ICCV*, 2009.

[18] H. Pashler. *The Psychology of Attention*. MIT Press, Cambridge, MA, 1997.

[19] F. Porikli. Integral histogram: A fast way to extract histograms in Cartesian spaces. In *Proc. CVPR*, 2005.

[20] B. Schauerte and G. A. Fink. Focusing computational visual attention in multi-modal human-robot interaction. In *Proc. of Int'l Conf. on Multimodal Interfaces (ICMI)*, 2010.

[21] A. M. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[22] J. K. Tsotsos. Towards a computational model of visual attention. In *Early Vision and Beyond*. MIT Press, 1995.

[23] P. Viola and M. Jones. Robust real-time object detection. *Int. Journal of Computer Vision*, 57(2):137–154, 2002.

[24] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 2006.

[25] J. M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin and Review*, 1(2), 1994.

[26] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5:1–7, 2004.