Saliency Filters: Contrast Based Filtering for Salient Region Detection

Federico Perazzi[1] Philipp Krähenbühl[2] Yael Pritch[1] Alexander Hornung[1]
[1]Disney Research Zurich [2]Stanford University

Abstract

Saliency estimation has become a valuable tool in image processing. Yet, existing approaches exhibit considerable variation in methodology, and it is often difficult to attribute improvements in result quality to specific algorithm properties. In this paper we reconsider some of the design choices of previous methods and propose a conceptually clear and intuitive algorithm for contrast-based saliency estimation.

Our algorithm consists of four basic steps. First, our method decomposes a given image into compact, perceptually homogeneous elements that abstract unnecessary detail. Based on this abstraction we compute two measures of contrast that rate the uniqueness and the spatial distribution of these elements. From the element contrast we then derive a saliency measure that produces a pixel-accurate saliency map which uniformly covers the objects of interest and consistently separates fore- and background.

We show that the complete contrast and saliency estimation can be formulated in a unified way using high-dimensional Gaussian filters. This contributes to the conceptual simplicity of our method and lends itself to a highly efficient implementation with linear complexity. In a detailed experimental evaluation we analyze the contribution of each individual feature and show that our method outperforms all state-of-the-art approaches.

1. Introduction

The computational identification of image elements that are likely to catch the attention of a human observer is a complex cross-disciplinary problem. Realistic, high-level models need to be founded on a combination of insights from neurosciences, biology, computer vision, and other fields. However, recent research has shown that computational models simulating *low-level* stimuli-driven attention [17, 20, 21] are quite successful and represent useful tools in many application scenarios, including image segmentation [14], resizing [5] and object detection [27].

Results from perceptual research [11,24,25] indicate that the most influential factor in low-level visual saliency is *contrast*. However, the definition of contrast in previous works is based on various different types of image features, including color variation of individual pixels, edges and gradients, spatial frequencies, structure and distribution of image patches, histograms, multi-scale descriptors, or combinations thereof. The significance of each individual feature often remains unclear [21], and as recent evaluations show [7] even quite similar approaches may exhibit considerably varying performance.

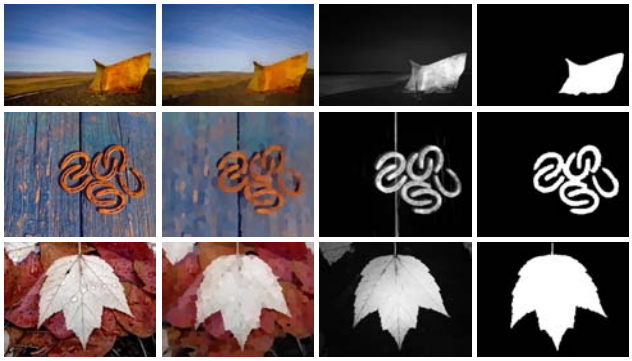


Figure 1: From left to right: input images, image abstraction into perceptually homogeneous elements, results of our saliency computation, ground truth labeling.

In this work we reconsider the set of fundamentally relevant contrast measures and their definition in terms of image content. Our method is based on the observation that an image can be decomposed into basic, structurally representative elements that abstract away unnecessary detail, and at the same time allow for a very clear and intuitive definition of contrast-based saliency.

Our first main contribution therefore is a concept and algorithm to decompose an image into perceptually homogeneous elements and to derive a saliency estimate from two well-defined contrast measures based on the uniqueness and spatial distribution of those elements. Both, local as well as the global contrast are handled by these measures in a unified way.

Central to the contrast and saliency computation is our second main contribution; we show that all involved operators can be formulated within a single high-dimensional Gaussian filtering framework. Thanks to this formulation, we achieve a highly efficient implementation with linear

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

complexity. The same formulation also provides a clear link between the element-based contrast estimation and the actual assignment of saliency values to all image pixels.

As we demonstrate in our experimental evaluation, each of our individual measures already performs close to or even better than existing approaches, and our combined method currently achieves the best ranking results on the public benchmark provided by [2, 21].

2. Related Work

Methods that model bottom-up, low-level saliency can be roughly classified into biologically inspired methods and computationally oriented approaches. Works belonging to the first class [15, 18] are generally based on the architecture proposed by Koch and Ullman [20], in which the low-level stage processes features such as color, orientation of edges, or direction of movement. One implementation of this model is the work by Itti *et al*. [18], which use a Difference of Gaussians approach to evaluate those features. However, as the evaluation by Cheng *et al*. [7] shows, the resulting saliency maps are generally blurry, and often overemphasize small, purely local features, which renders this approach less useful for applications such as segmentation, detection, etc.

In contrast, computational methods may also be inspired by biological principles, but relate stronger to typical applications in computer vision and graphics. For example, frequency space methods [13, 16] determine saliency based on the amplitude or phase spectrum of the Fourier transform of an image. The resulting saliency maps better preserve the high level structure of an image than [18], but exhibit undesirable blurriness and tend to highlight object boundaries rather than its entire area.

For colorspace techniques one can distinguish between approaches using local or global analysis of (color-) contrast. Local methods estimate the saliency of a particular image region based on immediate image neighborhoods, *e.g*., based on dissimilarities at the pixel-level [22], using multi-scale Difference of Gaussians [17] or histogram analysis [21]. While such approaches are able to produce less blurry saliency maps, they are agnostic of global relations and structures, and they may also be more sensitive to high frequency content like image edges and noise [2].

Global methods take contrast relations over the complete image into account. For example, there are different variants of patch-based methods which estimate dissimilarity between image patches [12, 21, 28]. While these algorithms are more consistent in terms of global image structures, they suffer from the involved combinatorial complexity, hence they are applicable only to relatively low resolution images, or they need to operate in spaces of reduced dimensionality [10], resulting in loss of small, potentially salient detail. The method of Achanta *et al*. [2] also works on a per-pixel basis, but achieves globally more consistent results by computing color dissimilarities to the mean image color. They use Gaussian blur in order to decrease the influence of noise and high frequency patterns. However, their method does not account for any spatial relationship inside the image, and may highlight background regions as salient.

Related to our definition of contrast is the work of Liu *et al*. [21] which combines multi-scale contrast, local contrast based on surrounding, context, and color spatial distribution to learn a conditional random field (CRF) for binary saliency estimation. However, the significance of features in the CRF remains unclear. Ren *et al*. [26] and Cheng *et al*. [7] employ image segmentation as part of their saliency estimation. In [26] the segmentation serves solely to alleviate the negative influence of highly textured regions, noise and outliers during their subsequent clustering. Cheng *et al*. [7], who generate 3D histograms and compute dissimilarities between histogram bins, reported the best performing method among global contrast-based approaches so far. However, due to the use of larger-scale image segments in both approaches [7, 26], contrast measures involving spatial distribution cannot easily be formulated. Moreover, such methods have problems handling images with cluttered and textured background.

Despite many recent improvements, the varying evaluation results in [7] indicate that the actual significance of individual features and contrast measures in existing methods is difficult to assess. Our work reduces the set of contrast measures to just two, which can be intuitively defined over abstract image elements, while still producing pixel-accurate saliency masks.

3. Overview

As motivated before, we propose an algorithm that first decomposes the input image into basic elements. Based on these elements we define two measures for contrast that are used to compute per-pixel saliency. Hence, our algorithm consists of the following steps (see Figure 2).

1. Abstraction. We aim to decompose the image into basic elements that preserve relevant structure, but abstract undesirable detail. Specifically, each element should locally abstract the image by clustering pixels with similar properties (like color) into perceptually homogeneous regions. Discontinuities between such regions, *i.e*., strong contours and edges in the image, should be preserved as boundaries between individual elements. Finally, constraints on shape and size should allow for compact, well localized elements.

One approach to achieve this type of decomposition is an edge-preserving, localized oversegmentation based on color (see Figure 2 b). Thanks to this abstraction, contrast between whole image regions can be evaluated using just those elements. Furthermore, we show that the quality

(a) Source image. (b) Abstraction. (c) Uniqueness. (d) Distribution. (e) Saliency. (f) Ground truth.

Figure 2: Illustration of the main phases of our algorithm. The input image is first abstracted into perceptually homogeneous elements. Each element is represented by the mean color of the pixels belonging to it. We then define two contrast measures per element based on the uniqueness and spatial distribution of elements. Finally, a saliency value is assigned to each pixel.

of saliency maps is extremely robust to the number of elements. We can then define our two measures for contrast.

2. Element uniqueness. This first contrast measure implements the commonly employed assumption that image regions, which stand out from other regions in certain aspects, catch our attention and hence should be labeled more salient. We therefore evaluate how different each respective element is from all other elements constituting an image, essentially measuring the "rarity" of each element.

In one form or another, this assumption has been the basis for most previous algorithms for contrast-based saliency. However, thanks to our abstraction, variation on the pixel level due to small scale textures or noise is rendered irrelevant, while discontinuities such as strong edges stay sharply localized. As discussed in Section 2, previous multi-scale techniques often blur or lose this information.

3. Element distribution. While saliency implies uniqueness, the opposite might not always be true [19]. Ideally colors belonging to the background will be distributed over the entire image exhibiting a high spatial variance, whereas foreground objects are generally more compact [12, 21].

The compactness and locality of our image abstracting elements allows us to define a corresponding second measure, which renders unique elements more salient when they are grouped in a particular image region rather than evenly distributed over the whole image. Techniques based on larger-scale image segmentation such as [7] lose this important source of information.

An example showing the differences between element uniqueness and element distribution is shown in Figure 3.

4. Saliency assignment. The two above contrast measures are defined on a per-element level. In a final step, we assign the actual saliency values to the input image to get a pixel-accurate saliency map. Thanks to this step our method can assign proper saliency values even to fine pixel-level detail that was excluded, on purpose, during the abstraction phase, but for which we still want a saliency estimate that conforms to the global saliency analysis.

4. Algorithm

In the following we describe the details of our method, and we show how the contrast measures as well as the saliency assignment can be efficiently computed based on N-D Gaussian filtering [4].

4.1. Abstraction

For the image abstraction we use an adaptation of SLIC superpixels [3] to abstract the image into perceptually uniform regions. SLIC superpixels segment an image using K-means clustering in RGBXY space. The RGBXY space yields local, compact and edge aware superpixels, but does not guarantee compactness. For our image abstraction we slightly modified the SLIC approach and instead use K-means clustering in geodesic image distance [8] in CIELab space. Geodesic image distance guarantees connectivity, while retaining the locality, compactness and edge awareness of SLIC superpixels. See Figures 2 and 7 for examples.

4.2. Element uniqueness

Element uniqueness is generally defined as the rarity of a segment i given its position $\mathbf{p}_i$ and color in CIELab $\mathbf{c}_i$ compared to all other segments j:

$$U_i = \sum_{j=1}^{N} \|\mathbf{c}_i - \mathbf{c}_j\|^2 \cdot \underbrace{w(\mathbf{p}_i, \mathbf{p}_j)}_{w_{ij}^{(p)}}. \qquad (1)$$

By introducing $w_{ij}^{(p)}$ we effectively combine global and local contrast estimation with control over the influence radius of the uniqueness operator. A local function $w_{ij}^{(p)}$ yields a local contrast term, which tends to overemphasize object boundaries in the saliency estimation [22], whereas $w_{ij}^{(p)} \approx 1$ yields a global uniqueness operator, which cannot represent sensitivity to local contrast variation.

Moreover, evaluating Eq. (1) globally generally requires $O(N^2)$ operations, where N is the number of segments. This is why some related works down-sample the image to a resolution where quadratic number of operations is feasible. As discussed in previous sections, saliency maps computed on down-sampled images cannot preserve sharply localized

contours and generally exhibit a high level of blurriness (see comparison in Section 5). Cheng *et al*. [7] approximate Eq. (1) using a histogram. Achatan *et al*. [2] approximate it as the distance to mean color. Both approximations are completely global with $w_{ij}^{(p)} = 1$.

We will show that for a Gaussian weight $w_{ij}^{(p)} = \frac{1}{Z_i} \exp(-\frac{1}{2\sigma_p^2}\|\mathbf{p}_i - \mathbf{p}_j\|^2)$ Eq. (1) can be evaluated in linear time $O(N)$. σ_p controls the range of the uniqueness operator and Z_i is the normalization factor ensuring $\sum_{j=1}^{N} w_{i,j}^{(p)} = 1$. We decompose Eq. (1) by factoring out the quadratic error function:

$$\begin{aligned} U_i &= \sum_{j=1}^{N} \|\mathbf{c}_i - \mathbf{c}_j\|^2 w_{ij}^{(p)} \\ &= \mathbf{c}_i^2 \underbrace{\sum_{j=1}^{N} w_{ij}^{(p)}}_{1} - 2\mathbf{c}_i \underbrace{\sum_{j=1}^{N} \mathbf{c}_j w_{ij}^{(p)}}_{\text{blur } c_j} + \underbrace{\sum_{j=1}^{N} \mathbf{c}_j^2 w_{ij}^{(p)}}_{\text{blur } c_j^2}. \end{aligned} \quad (2)$$

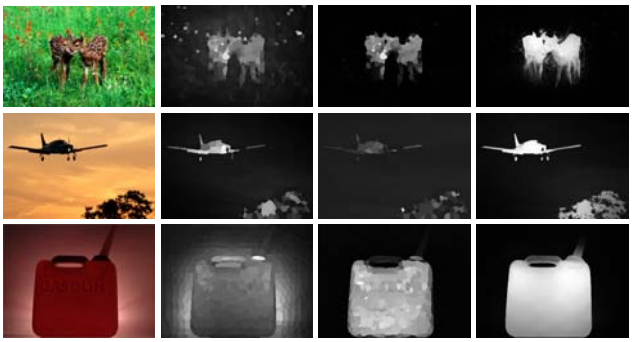
Both terms $\sum_{j=1}^{N} \mathbf{c}_j w_{ij}^{(p)}$ and $\sum_{j=1}^{N} \mathbf{c}_j^2 w_{ij}^{(p)}$ can be evaluated using a Gaussian blurring kernel on color $\mathbf{c}_j$ and the squared color $\mathbf{c}_j^2$. Gaussian blurring is decomposable along x and y axis of the image and can thus be evaluated very efficiently.

In our implementation we use the permutohedral lattice embedding presented in Adams *et al*. [4], which yields a linear time approximation of the Gaussian filter in arbitrary dimensions. The permutohedral lattice exploits the band limiting effects of Gaussian smoothing, such that a correspondingly filtered function can be well approximated by a sparse number of samples. Adams *et al*. use samples on simplices of a high dimensional lattice structure to represent the result of the filtering operation. They then evaluate the filter by downsampling the input values onto the lattice, blur along each dimension of the lattice and reconstruct the resulting signal by interpolation.

By using a Gaussian weight $w_{ij}^{(p)}$ we are able to evaluate Eq. (1) in linear time, without crude approximations such as histograms or distance to mean color. Parameter σ_p was set to 0.25 in all experiments, which allows for a balance between local and global effects. Examples for the uniqueness measure are shown in Figure 3b.

4.3. Element distribution

Conceptually, we define the element distribution measure for a segment i using the spatial variance D_i of its color $\mathbf{c}_i$, *i.e*., we measure its occurrence elsewhere in the image. As motivated before, low variance indicates a spatially compact object which should be considered more salient than spatially widely distributed elements. Hence we compute



(a) Source image. (b) Uniqueness. (c) Distribution. (d) Saliency.

Figure 3: Uniqueness, spatial distribution, and the combined saliency map. The uniqueness prefers rare colors, whereas the distribution favors compact objects. Combined together those measures provide better perfomance.

$$D_i = \sum_{j=1}^{N} \|\mathbf{p}_j - \mu_i\|^2 \underbrace{w(\mathbf{c}_i, \mathbf{c}_j)}_{w_{ij}^{(c)}}, \quad (3)$$

where $w_{ij}^{(c)}$ describes the similarity of color $\mathbf{c}_i$ and color $\mathbf{c}_j$ of segments i and j, respectively, $\mathbf{p}_j$ is again the position of segment j, and $\mu_i = \sum_{j=1}^{N} w_{ij}^{(c)} \mathbf{p}_j$ defines the weighted mean position of color $\mathbf{c}_i$.

Again naive evaluation of Eq. (3) has quadratic runtime complexity. By choosing the color similarity to be Gaussian $w_{ij}^{(c)} = \frac{1}{Z_i} \exp(-\frac{1}{2\sigma_c^2}\|\mathbf{c}_i - \mathbf{c}_j\|^2)$, we can efficiently evaluate it in linear time:

$$\begin{aligned} D_i &= \sum_{j=1}^{N} \|\mathbf{p}_j - \mu_i\|^2 w_{ij}^{(c)} \\ &= \sum_{j=1}^{N} \mathbf{p}_j^2 w_{ij}^{(c)} - 2\mu_i \underbrace{\sum_{j=1}^{N} \mathbf{p}_j w_{ij}^{(c)}}_{\mu_i} + {\mu_i}^2 \underbrace{\sum_{j=1}^{N} w_{ij}^{(c)}}_{1} \\ &= \underbrace{\sum_{j=1}^{N} \mathbf{p}_j^2 w_{ij}^{(c)}}_{\text{blur } \mathbf{p}_j^2} - \underbrace{\mu_i}_{\text{blur } \mathbf{p}_j}{}^2. \end{aligned} \quad (4)$$

Here the position $\mathbf{p}_j$ and squared position $\mathbf{p}_j^2$ are blurred in the 3-dimensional color space. It can be efficiently evaluated by discretizing the color space and then evaluating a separable Gaussian blur along each of the L, a and b dimension. Since the Gaussian filter is additive, we can simply add position values associated to the same color. As in Eq. (2) we use the permutohedral lattice [4] as a linear approximation to the Gaussian filter in the Lab space.

The parameter σ_c controls the color sensitivity of the element distribution. We use $\sigma_c = 20$ in all our experiments.

See Figure 3 for a visual comparison of uniqueness and spatial distribution.

In summary, by simple evaluation of two Gaussian filters we can compute two non-trivial, but intuitively defined contrast measures on a per-element basis. By filtering color values in the image, we compute the uniqueness of an element, while filtering position values in the Lab color space gives us the element distribution. Next we will look at how to combine both measures, which have a different scaling and units associated to them, in order to compute a per-pixel saliency value.

4.4. Saliency assignment

We start by normalizing both uniqueness U_i and distribution D_i to the range $[0..1]$. We assume that both measures are independent, and hence we combine these terms as follows to compute a saliency value S_i for each element:

$$S_i = U_i \cdot \exp(-k \cdot D_i), \tag{5}$$

In practice we found the distribution measure D_i to be of higher significance and discriminative power. Therefore, we use an exponential function in order to emphasize D_i. In all our experiments we use $k = 6$ as the scaling factor for the exponential. Figure 6 (middle) shows the performance of the uniqueness U_i, distribution D_i and their combination S_i, while Figure 3 shows a visual comparison.

As the final step, we need to assign a final saliency value to each image pixel, which can be interpreted as an up-sampling of the per-element saliency S_i. However, naive up-sampling by assigning S_i to every pixel contained in element i carries over all segmentation errors of the abstraction algorithm. Instead we adopt an idea proposed in the context of range image up-sampling [9] and apply it to our framework. We define the saliency $\tilde{S}_i$ of a pixel as a weighted linear combination of the saliency S_j of its surrounding image elements

$$\tilde{S}_i = \sum_{j=1}^{N} w_{ij} S_j. \tag{6}$$

By choosing a Gaussian weight $w_{ij} = \frac{1}{Z_i} \exp(-\frac{1}{2}(\alpha \|\mathbf{c}_i - \mathbf{c}_j\|^2 + \beta \|\mathbf{p}_i - \mathbf{p}_j\|^2)$, we ensure the up-sampling process is both local and color sensitive. Here α and β are parameters controlling the sensitivity to color and position. We found $\alpha = \frac{1}{30}$ and $\beta = \frac{1}{30}$ to work well in practice.

As for our contrast measures in Eq. (1) and (3), Eq. (6) describes a high-dimensional Gaussian filter and can hence be evaluated within the same filtering framework [4]. The saliency value of each element is embedded in a five-dimensional space using its position $\mathbf{p}_i$ and its color value $\mathbf{c}_i$ in RGB (as we found it to outperform CIELab for up-sampling). Since our abstract elements do not have a regular shape we create a point sample in RGBXY space at each pixel position $\tilde{\mathbf{p}}_i$ within a particular element and blur the RGBXY space along each of its dimensions. The per-pixel saliency values can then be retrieved with a lookup in that high-dimensional space using the pixel's position $\tilde{\mathbf{p}}_i$ and its color value $\tilde{\mathbf{c}}_i$ in the input image.

The resulting pixel-level saliency map can have an arbitrary scale. In a final step we rescale the saliency map to the range $[0..1]$ or to contain at least 10% saliency pixels.

In summary, our algorithm computes the saliency of an image by first abstracting it into small, perceptually homogeneous elements. It then applies a series of three Gaussian filtering steps in order to compute the uniqueness and spatial distribution of elements as well as to perform the final per-pixel saliency assignment.

5. Results

We provide an exhaustive comparison of our approach (SF) to several state-of-art methods on a database of 1000 images [23] with binary ground truth [2]. Saliency maps of previous works are provided by [7]. Figure 5 shows a visual comparison of the different methods.

5.1. Precision and Recall

Similar to [2, 7, 21], we evaluate the performance of our algorithm measuring its *precision* and *recall* rate. Precision corresponds to the percentage of salient pixels correctly assigned, while recall corresponds to the fraction of detected salient pixels in relation to the ground truth number of salient pixels.

High recall can be achieved at the expense of reducing the precision and vice-versa so it is important to evaluate both measures together. We perform two different experiments. In both cases we generate a binary saliency map based on some saliency threshold. In the first experiment we compare binary masks for every threshold in the range $[0..255]$. The resulting curves in Figure 4 show that our algorithm (SF) consistently produces results closer to ground truth at every threshold and for any given recall rate.

In the second experiment we use the image dependent adaptive threshold proposed by [2], defined as twice the mean saliency of the image:

$$T_a = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x, y), \tag{7}$$

where W and H are the width and the height of the saliency map S, respectively. In addition to precision and recall we compute their weighted harmonic mean measure or *F-measure*, which is defined as:

$$F_\beta = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}. \tag{8}$$

Similar to [2, 7] we set $\beta^2 = 0.3$.

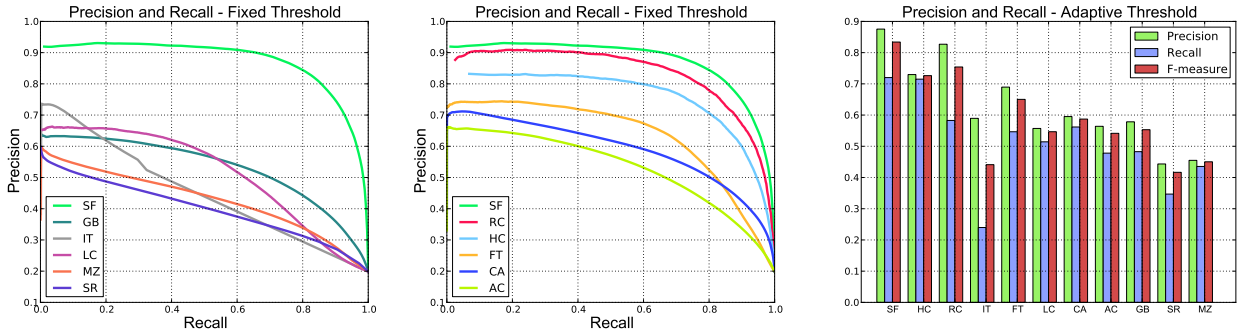
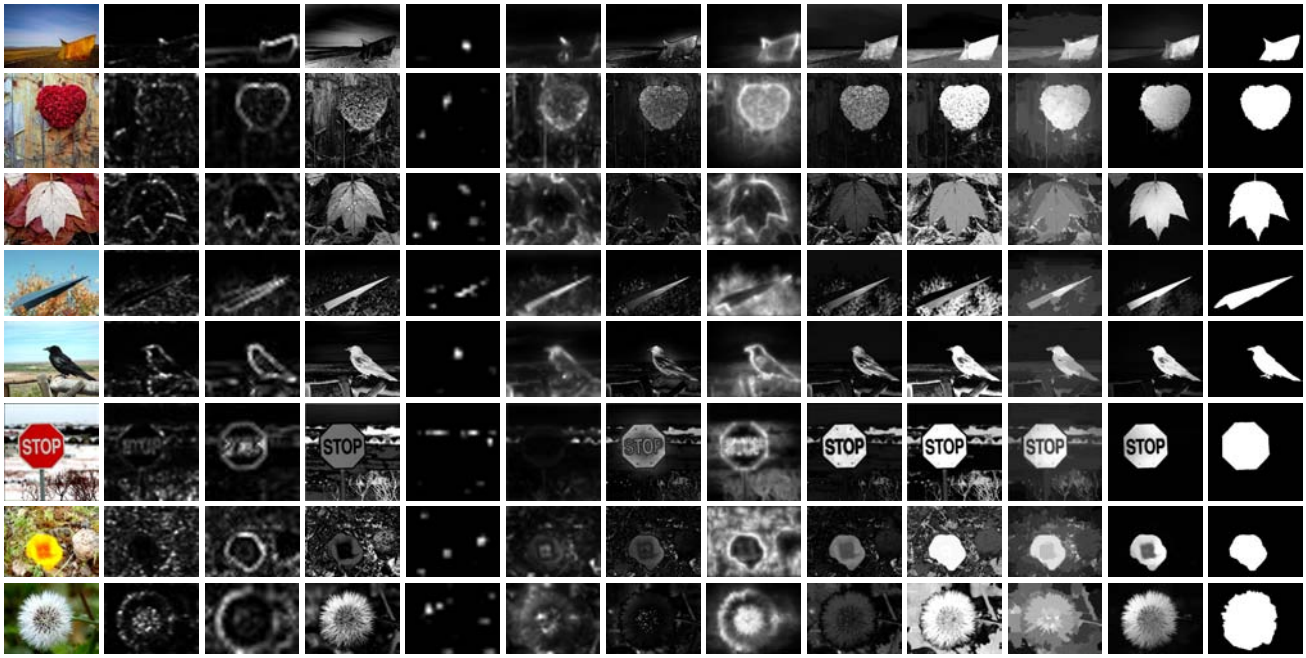


Figure 4: Left, middle: precision and recall rates for all algorithms. Right: precision, recall, and F-measure for adaptive thresholds. In all experiments, our approach consistently produces results closest to ground truth. See the legend of Figure 5 for the references to all methods.



(a) SRC (b) SR [16](c) MZ [22](d) LC [29] (e) IT [18] (f) GB [15] (g) AC [1] (h) CA [12] (i) FT [2] (j) HC [7] (k) RC [7] (l) SF (m) GT

Figure 5: Visual comparison of previous approaches to our method (SF) and ground truth (GT). As also shown in the numerical evaluation, SF consistently produces saliency maps closest to ground truth. We compare to spectral residual saliency (SR [16]), fuzzy growing (MZ [22]), spatiotemporal cues (LC [29]), visual attention measure (IT [18]), graph-based saliency (GB [15]), salient region detection (AC [1]), context-aware saliency (CA [12]), frequency-tuned saliency (FT [2]) and global-contrast saliency (HC [7] and RC [7]).

Figure 6 shows that our algorithm performs consistently and robustly over a wide range of numbers of image elements. Furthermore, we evaluate precision and recall for each individual phase of our algorithm, showing the benefit of combining all steps.

5.2. Mean Absolute Error

Neither the precision nor recall measure consider the true negative saliency assignments, *i.e.*, the number of pixel correctly marked as non-salient. This favors methods that successfully assign saliency to salient pixels but fail to detect non-salient regions over methods that successfully detect non-salient pixels but make mistakes in determining the salient ones. Moreover, in some application scenarios [5] the quality of the weighted, continuous saliency maps may be of higher importance than the binary masks.

For a more balanced comparison that takes these effects into account we therefore also evaluate the *mean absolute error* (MAE) between the continuous saliency map S (prior to thresholding) and the binary ground truth GT. The mean

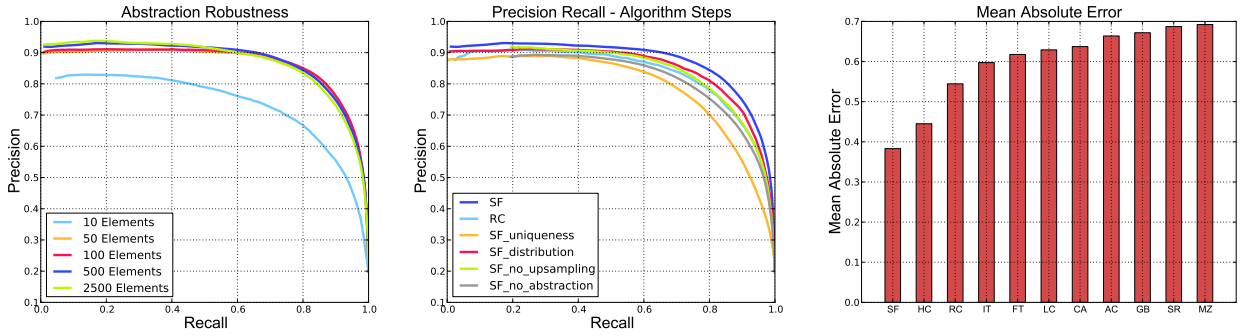
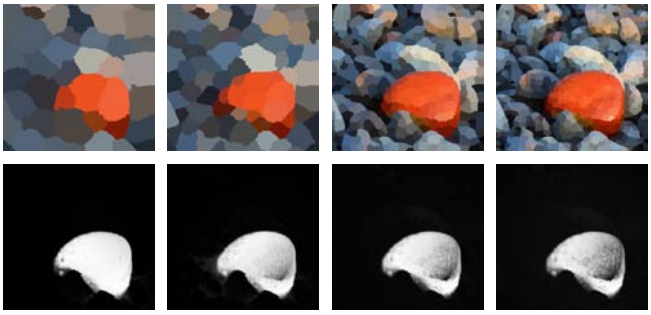


Figure 6: Left: a comparison of precision and recall curves for different numbers of image elements shows that our method performs robustly over a wide range of image elements (see also Figure 7). A significant drop is only visible for an extremely low number of 10 elements. Middle: evaluation of each individual phase of our algorithm. Both contrast measures alone achieve a performance close to the state-of-the-art RC [7]. However, the combination of all steps in our algorithm is crucial for optimal performance. Right: Mean absolute error of the different saliency methods to ground truth.



(a) 50. (b) 100. (c) 500. (d) 1000.

Figure 7: Visual comparison of resulting saliency maps for different numbers of image elements.

absolute error is then defined as

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - GT(x,y)|, \quad (9)$$

where W and H are again the width and the height of the respective saliency map and ground truth image.

Figure 6 shows that our method also outperforms the other approaches in terms of the MAE measure, which provides a better estimate of the dissimilarity between the saliency map and ground truth. Results have been averaged over all images in [23], and all results have been generated with the same parameter settings. It is also interesting to observe that the HC method has a lower MAE than RC, which is in contrast to the precision and recall results.

5.3. Performance

In Table 1 we compare the average running time of our approach to the currently best performing methods on the benchmark images. Timings have been taken on an Intel Core i7-920 2.6 GHz with 3GB RAM. Our running times

Method	CA [12]	FT [2]	HC [7]	RC [7]	SF
Time(s)	51.2	0.012	0.011	0.144	0.153
Code	Matlab	C++	C++	C++	C++

Table 1: Comparison of running times.

are similar to that of RC (both methods involve segmentation), with our method spending most of the processing time on abstraction (about 40%) and the final saliency upsampling (50%). Only 10% account for the actual per-element contrast and saliency computation. The CA method is slower because it requires an exhaustive nearest-neighbor search among patches.

5.4. Limitations

Saliency estimation based on color contrast may not always be feasible, *e.g.*, in the case of lighting variations, or when fore- and background colors are very similar. In such cases, the thresholding procedures used for all the above evaluations can result in noisy segmentations (see Figure 8).

One option to significantly reduce this effect is to perform a single min-cut segmentation [6] as a post process, using our saliency maps as a prior for the min-cut data term, and color differences between neighboring pixels for the smoothness term. The graph structure facilitates smoothness of salient objects and significantly improves the performance of our algorithm, when binary saliency maps are required for challenging images. As Figure 8 shows, even in cases where thresholded saliency masks are not of the desired quality, the original continuous saliency maps are of sufficient coherence so that a straight forward min-cut segmentation produces high quality masks.

6. Conclusions

We presented Saliency Filters, a method for saliency computation based on an image abstraction into struc-

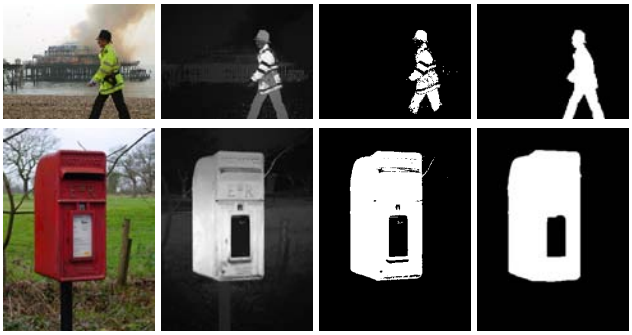


Figure 8: Limitations and min-cut segmentation. From left to right: Input image, saliency map computed with our method, the noisy result of simple thresholding, and min-cut segmentation applied to the saliency map.

turally representative elements and contrast-based saliency measures, which can be consistently formulated as high-dimensional Gaussian filters. Our filter-based formulation allows for efficient computation and produces per-pixel saliency maps, with the currently best performance in a ground truth comparison to various state-of-the-art works.

For future work we believe that investigating more sophisticated techniques for image abstraction, including robust color or structure distance measures, will be beneficial. Moreover, our proposed filter-based formulation is sufficiently general to serve as an extendable framework, *e.g.*, to incorporate higher-level features such as face detectors.

References

[1] R. Achanta, F. J. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. In *ICVS*, pages 66–75, 2008. 6

[2] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 2, 4, 5, 6, 7

[3] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. SLIC Superpixels. Technical report, 2010. 3

[4] A. Adams, J. Baek, and M. A. Davis. Fast high-dimensional filtering using the permutohedral lattice. *Comput. Graph. Forum*, 29(2):753–762, 2010. 3, 4, 5

[5] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3):10, 2007. 1, 6

[6] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE TPAMI*, 26(9):1124–1137, 2004. 7

[7] 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, 2011. 1, 2, 3, 4, 5, 6, 7

[8] A. Criminisi, T. Sharp, C. Rother, and P. Pérez. Geodesic image and video editing. *ACM Trans. Graph.*, 29(5):134, 2010. 3

[9] J. Dolson, J. Baek, C. Plagemann, and S. Thrun. Upsampling range data in dynamic environments. In *CVPR*, pages 1141–1148, 2010. 5

[10] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu. Visual saliency detection by spatially weighted dissimilarity. In *CVPR*, pages 473–480, 2011. 2

[11] W. Einhäuser and P. König. Does luminance-contrast contribute to a saliency map for overt visual attention? *Eur J Neurosci*, 17(5):1089–1097, Mar. 2003. 1

[12] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, pages 2376–2383, 2010. 2, 3, 6, 7

[13] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR*, 2008. 2

[14] J. Han, K. N. Ngan, M. Li, and H. Zhang. Unsupervised extraction of visual attention objects in color images. *IEEE Trans. Circuits Syst. Video Techn.*, 16(1):141–145, 2006. 1

[15] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 2, 6

[16] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007. 2, 6

[17] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *NIPS*, 2005. 1, 2

[18] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998. 2, 6

[19] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001. 3

[20] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985. 1, 2

[21] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *CVPR*, 2007. 1, 2, 3, 5

[22] Y.-F. Ma and H. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM Multimedia*, pages 374–381, 2003. 2, 3, 6

[23] 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 *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001. 5, 7

[24] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Res*, 42(1):107–123, Jan. 2002. 1

[25] P. Reinagel and A. M. Zador. Natural scene statistics at the centre of gaze. In *Network: Computation in Neural Systems*, pages 341–350, 1999. 1

[26] Z. Ren, Y. Hu, L.-T. Chia, and D. Rajan. Improved saliency detection based on superpixel clustering and saliency propagation. In *ACM Multimedia*, pages 1099–1102, 2010. 2

[27] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *CVPR (2)*, pages 37–44, 2004. 1

[28] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. A. Rowley. Image saliency: From intrinsic to extrinsic context. In *CVPR*, pages 417–424, 2011. 2

[29] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *ACM Multimedia*, pages 815–824, 2006. 6