

Frequency-tuned Salient Region Detection

Radhakrishna Achanta[†], Sheila Hemami[‡], Francisco Estrada[†], and Sabine Süsstrunk[†]

[†]School of Computer and Communication Sciences (IC)

Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Switzerland.

[radhakrishna.achanta, francisco.estrada, sabine.susstrunk]@epfl.ch

[‡]School of Electrical and Computer Engineering

Cornell University, Ithaca, NY 14853, U.S.A.

hemami@ece.cornell.edu

Abstract

Detection of visually salient image regions is useful for applications like object segmentation, adaptive compression, and object recognition. In this paper, we introduce a method for salient region detection that outputs full resolution saliency maps with well-defined boundaries of salient objects. These boundaries are preserved by retaining substantially more frequency content from the original image than other existing techniques. Our method exploits features of color and luminance, is simple to implement, and is computationally efficient. We compare our algorithm to five state-of-the-art salient region detection methods with a frequency domain analysis, ground truth, and a salient object segmentation application. Our method outperforms the five algorithms both on the ground-truth evaluation and on the segmentation task by achieving both higher precision and better recall.

1. Introduction

Visual saliency is the perceptual quality that makes an object, person, or pixel stand out relative to its neighbors and thus capture our attention. Visual attention results both from fast, pre-attentive, bottom-up visual saliency of the retinal input, as well as from slower, top-down memory and volition based processing that is task-dependent [24].

The focus of this paper is the automatic detection of visually salient regions in images, which is useful in applications such as adaptive content delivery [22], adaptive region-of-interest based image compression [4], image segmentation [18, 9], object recognition [26], and content aware image resizing [2]. Our algorithm finds low-level, pre-attentive, bottom-up saliency. It is inspired by the biological concept of center-surround contrast, but is not based on any biological model.

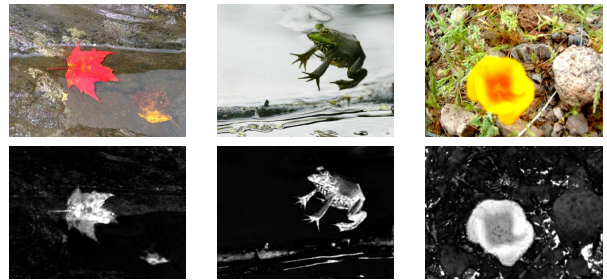


Figure 1. Original images and their saliency maps using our algorithm.

Current methods of saliency detection generate regions that have low resolution, poorly defined borders, or are expensive to compute. Additionally, some methods produce higher saliency values at object edges instead of generating maps that uniformly cover the whole object, which results from failing to exploit all the spatial frequency content of the original image. We analyze the spatial frequencies in the original image that are retained by five state-of-the-art techniques, and visually illustrate that these techniques primarily operate using extremely low-frequency content in the image. We introduce a frequency-tuned approach to estimate center-surround contrast using color and luminance features that offers three advantages over existing methods: uniformly highlighted salient regions with well-defined boundaries, full resolution, and computational efficiency. The saliency map generated can be more effectively used in many applications, and here we present results for object segmentation. We provide an objective comparison of the accuracy of the saliency maps against five state-of-the-art methods using a ground truth of a 1000 images. Our method outperforms all of these methods in terms of precision and recall.

2. General approaches to determining saliency

The term *saliency* was used by Tsotsos et al. [27] and Olshausen et al. [25] in their work on visual attention, and by Itti et al. [16] in their work on rapid scene analysis. Saliency has also been referred to as *visual attention* [27, 22], *unpredictability*, *rarity*, or *surprise* [17, 14]. Saliency estimation methods can broadly be classified as biologically based, purely computational, or a combination. In general, all methods employ a low-level approach by determining contrast of image regions relative to their surroundings, using one or more features of intensity, color, and orientation.

Itti et al. [16] base their method on the biologically plausible architecture proposed by Koch and Ullman [19]. They determine center-surround contrast using a Difference of Gaussians (DoG) approach. Frintrop et al. [7] present a method inspired by Itti's method, but they compute center-surround differences with square filters and use integral images to speed up the calculations.

Other methods are purely computational [22, 13, 12, 1] and are not based on biological vision principles. Ma and Zhang [22] and Achanta et al. [1] estimate saliency using center-surround feature distances. Hu et al. [13] estimate saliency by applying heuristic measures on initial saliency measures obtained by histogram thresholding of feature maps. Gao and Vasconcelos [8] maximize the mutual information between the feature distributions of center and surround regions in an image, while Hou and Zhang [12] rely on frequency domain processing.

The third category of methods are those that incorporate ideas that are partly based on biological models and partly on computational ones. For instance, Harel et al. [10] create feature maps using Itti's method but perform their normalization using a graph based approach. Other methods use a computational approach like maximization of information [3] that represents a biologically plausible model of saliency detection.

Some algorithms detect saliency over multiple scales [16, 1], while others operate on a single scale [22, 13]. Also, individual feature maps are created separately and then combined to obtain the final saliency map [15, 22, 13, 7], or a feature combined saliency map is directly obtained [22, 1].

2.1. Limitations of saliency maps

The saliency maps generated by most methods have low resolution [16, 22, 10, 7, 12]. Itti's method produces saliency maps that are just $1/256^{th}$ the original image size in pixels, while Hou and Zhang [12] output maps of size 64×64 pixels for any input image size. An exception is the algorithm presented by Achanta et al. [1] that outputs saliency maps of the same size as the input image. This is accomplished by changing the filter size to achieve a change in scale rather than the original image size.

Depending on the salient region detector, some maps additionally have ill-defined object boundaries [16, 10, 7], limiting their usefulness in certain applications. This arises from severe downsizing of the input image, which reduces the range of spatial frequencies in the original image considered in the creation of the saliency maps. Other methods highlight the salient object boundaries, but fail to uniformly map the entire salient region [22, 12] or better highlight smaller salient regions than larger ones [1]. These shortcomings result from the limited range of spatial frequencies retained from the original image in computing the final saliency map as well as the specific algorithmic properties.

3. Frequency Domain Analysis of Saliency Detectors

We examine the information content used in the creation of the saliency maps of five state-of-the-art methods from a frequency domain perspective. The five saliency detectors are Itti et al. [16], Ma and Zhang [22], Harel et al. [10], Hou and Zhang [12], and Achanta et al. [1], hereby referred to as IT, MZ, GB, SR, and AC, respectively. We refer to our proposed method as IG. The choice of these algorithms is motivated by the following reasons: citation in literature (the classic approach of IT is widely cited), recency (GB, SR, and AC are recent), and variety (IT is biologically motivated, MZ is purely computational, GB is a hybrid approach, SR estimates saliency in the frequency domain, and AC outputs full-resolution maps).

3.1. Spatial frequency content of saliency maps

To analyze the properties of the five saliency algorithms, we examine the spatial frequency content from the original image that is retained in computing the final saliency map. It will be shown in Sec. 4.3 that the range of spatial frequencies retained by our proposed algorithm is more appropriate than the algorithms used for comparison. For simplicity, the following analysis is given in one dimension and extensions to two dimensions are clarified when necessary.

In method IT, a Gaussian pyramid of 9 levels (level 0 is the original image) is built with successive Gaussian blurring and downsampling by 2 in each dimension. In the case of the luminance image, this results in a successive reduction of the spatial frequencies retained from the input image. Each smoothing operation approximately halves the normalized frequency spectrum of the image. At the end of 8 such smoothing operations, the frequencies retained from the spectrum of the original image at level 8 range within $[0, \pi/256]$. The technique computes differences of Gaussian-smoothed images from this pyramid, resizing them to size of level 4, which results in using frequency content from the original image in the range $[\pi/256, \pi/16]$. In this frequency range the *DC* (mean) component is removed

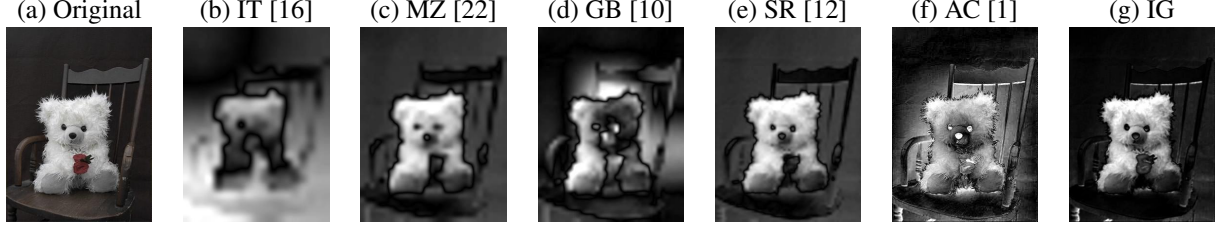


Figure 2. Original image filtered with band-pass filters with cut-off frequencies given in Table 3.1. (b)-(g) illustrate the spatial frequency information retained in the computation of each of the saliency maps.

along with approximately 99% ($(1 - \frac{1}{16^2}) \times 100$) of the high frequencies for a 2-D image. As such, the net information retained from the original image contains very few details and represents a very blurry version of the original image (see the band-pass filtered image of Fig. 2(b)).

In method MZ, a low-resolution image is created by averaging blocks of pixels and then downsampling the filtered image such that each block is represented by a single pixel having its average value. The averaging operation performs low-pass filtering. While the authors do not provide a block size for this operation, we obtained good with a block size of 10×10 pixels, and as such the frequencies retained from the original image are in the range $[0, \pi/10]$.

In method GB, the initial steps for creating feature maps are similar to IT, with the difference that fewer levels of the pyramid are used to find center-surround differences. The spatial frequencies retained are within the range $[\pi/128, \pi/8]$. Approximately 98% ($(1 - \frac{1}{8^2}) \times 100$) of the high frequencies are discarded for a 2D image. As illustrated in Fig. 2(d), there is slightly more high frequency content than in 2(b).

In method SR, the input image is resized to 64×64 pixels (via low-pass filtering and downsampling) based on the argument that the spatial resolution of pre-attentive vision is very limited. The resulting frequency content of the resized image therefore varies according to the original size of the image. For example, with input images of size 320×320 pixels (which is the approximate average dimension of the images of our test database), the retained frequencies are limited to the range $[0, \pi/5]$. As seen in Fig. 2(e), higher frequencies are smoothed out.

In method AC, a difference-of-means filter is used to estimate center-surround contrast. The lowest frequencies retained depend on the size of the largest surround filter (which is half of the image's smaller dimension) and the highest frequencies depend on the size of the smallest center filter (which is one pixel). As such, method AC effectively retains the entire range of frequencies $(0, \pi]$ with a notch at DC. All the high frequencies from the original image are retained in the saliency map but not all low frequencies (see Fig. 2(f)).

Method	Freq. range	Res.	Complexity
IT	$[\pi/256, \pi/16]$	$S/256$	$O(k_{IT}N)$
MZ	$[0, \pi/10]$	$S/100$	$O(k_{MZ}N)$
GB	$[\pi/128, \pi/8]$	$S/64$	$O(k_{GB}N^4K)$
SR	$[0, \pi/5]$	64×64	$O(k_{SR}N)$
AC	$(0, \pi]$	S	$O(k_{AC}N)$
IG	$(0, \pi/2.75]$	S	$O(k_{IG}N)$

Table 1. A comparison of 1-D frequency ranges, saliency map resolution, and computational efficiency. S is the input image size in pixels. Although the complexity of all methods except GB is proportional to N , the operations per pixel in these methods vary ($k_{MZ} < k_{SR} < k_{IG} < k_{AC} < k_{IT} < k_{GB}$). GB has an overall complexity of $O(k_{GB}N^4K)$, depending on the number of iterations K .

3.2. Other properties of methods MZ, SR, and AC

In MZ, the saliency value at each pixel position (i, j) is given by:

$$S(x, y) = \sum_{(m,n) \in N} d[\mathbf{p}(x, y), \mathbf{q}(m, n)] \quad (1)$$

where N is a small neighborhood of a pixel (in the resized image obtained by 10×10 box filtered and downsampled image) at position (x, y) and d is a Euclidean distance between Luv pixel vectors \mathbf{p} and \mathbf{q} . In our experiments, we choose N to be a 3×3 neighborhood. The method is fast but has the drawback that the saliency values at either side of an edge of a salient object are high, i.e the saliency maps show the salient object to be bigger than it is, which gets more pronounced if block sizes are bigger than 10×10 . In addition, for large salient objects, the salient regions are not likely to be uniformly highlighted (see Fig.3(c)).

In SR, the *spectral residual* R is found by subtracting a smoothed version of the FFT (Fast Fourier Transform) log-magnitude spectrum from the original log-magnitude spectrum. The saliency map is the inverse transform of the spectral residual. The FFT is smoothed using a separable 3×3 mean filter. Examining this operation in one dimension, this is equivalent to forming the residue $R(k)$ as:

$$R(k) = \ln|X(k)| - g_n * \ln|X(k)| \quad (2)$$

with $g_n = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$, and $*$ denoting convolution. A simple

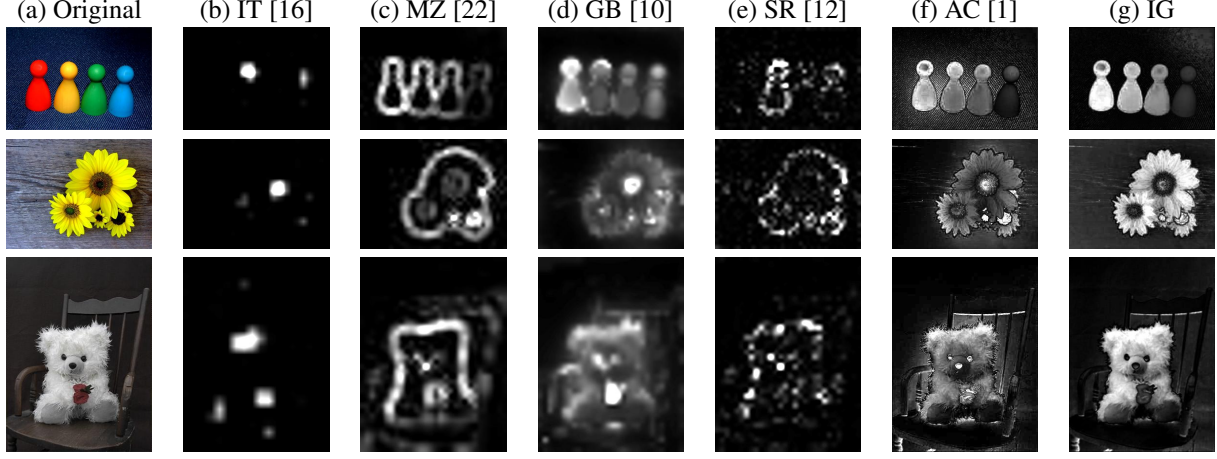


Figure 3. Visual comparison of saliency maps. (a) original image, (b) saliency maps using the method presented by, Itti [16], (c) Ma and Zhang [22], (d) Harel et al. [10], (e) Hou and Zhang [12], (f) Achanta et al. [1], and (g) our method. Our method generates sharper and uniformly highlighted salient regions as compared to other methods.

manipulation of this equation demonstrates that the (1-D) spectral residue $R(k)$ can be written as:

$$R(k) = \frac{1}{3} \ln \left[\frac{|X(k)|^2}{|X(k-1)||X(k+1)|} \right] \quad (3)$$

When this is performed in two dimensions on the 2-D FFT coefficients, only 3 low-frequency AC coefficients are divided by the DC (mean) value (if the FFT coefficients are circularly extended for the filtering, then the 3 highest-frequency FFT AC coefficients are also divided by the mean). In comparison, contrast measures typically normalize all FFT AC coefficients by the mean value [11].

In method AC, center-surround contrast is computed as a Euclidean distance between average Lab vectors of the center and surround regions. The saliency value at pixel position (x, y) is given as:

$$S(x, y) = \frac{1}{3} [F_{\frac{W}{2}}(x, y) + F_{\frac{W}{4}}(x, y) + F_{\frac{W}{8}}(x, y)]$$

$$F_t(x, y) = d(\mathbf{c}(x, y), \mathbf{s}_t(x, y)) \quad (4)$$

where feature map value $F_t(x, y)$ is found as a Euclidean distance d between the Lab pixel vector $\mathbf{c}(x, y)$ (center) and the average Lab pixel vector $\mathbf{s}_t(x, y)$ in window t (surround). The square surround region is varied as $t = \{\frac{W}{2}, \frac{W}{4}, \frac{W}{8}\}$, assuming W to be the smaller of the two dimensions of the image.

Objects that are smaller than a filter size are detected completely, while objects larger than a filter size are only partially detected (closer to edges). Smaller objects that are well detected by the smallest filter are detected by all three filters, while larger objects are only detected by the larger filters. Since the final saliency map is an average of the three feature maps (corresponding to detections of the three filters), small objects will almost always be better

highlighted. This explains why the toy bear's eyes (or the centers of the flowers) are more salient than the rest of the bear (or flowers) in Fig. 3(f).

4. Frequency-tuned Saliency Detection

In sections 2.1 and 3, shortcomings of existing saliency detection methods were mentioned. Motivated by these deficiencies, we propose a new algorithm in this section.

4.1. Requirements for a saliency map

We set the following requirements for a saliency detector:

- Emphasize the largest salient objects.
- Uniformly highlight whole salient regions.
- Establish well-defined boundaries of salient objects.
- Disregard high frequencies arising from texture, noise and blocking artifacts.
- Efficiently output full resolution saliency maps.

Let ω_{lc} be the low frequency cut-off value and ω_{hc} be the high frequency cut-off value. To highlight large salient objects, we need to consider very low frequencies from the original image, i.e. ω_{lc} has to be low (first criterion). This also helps highlight salient objects uniformly (second criterion). In order to have well defined boundaries, we need to retain high frequencies from the original image, i.e. ω_{hc} has to be high (third criterion). However, to avoid noise, coding artifacts, and texture patterns, the highest frequencies need to be disregarded (fourth criterion). Since we are interested in a saliency map containing a wide range of frequencies, combining the outputs of several band pass filters with contiguous $[\omega_{lc}, \omega_{hc}]$ pass bands is appropriate.

4.2. Combining DoG band pass filters

We choose the DoG filter (Eq. 5) for band pass filtering. The DoG filter is widely used in edge detection since it closely and efficiently approximates the Laplacian of Gaussian (LoG) filter, cited as the most satisfactory operator for detecting intensity changes when the standard deviations of the Gaussians are in the ratio 1:1.6 [23]. The DoG has also been used for interest point detection [21] and saliency detection [16, 10]. The DoG filter is given by:

$$\begin{aligned} DoG(x, y) &= \frac{1}{2\pi} \left[\frac{1}{\sigma_1^2} e^{-\frac{(x^2+y^2)}{2\sigma_1^2}} - \frac{1}{\sigma_2^2} e^{-\frac{(x^2+y^2)}{2\sigma_2^2}} \right] \\ &= G(x, y, \sigma_1) - G(x, y, \sigma_2) \end{aligned} \quad (5)$$

where σ_1 and σ_2 are the standard deviations of the Gaussian ($\sigma_1 > \sigma_2$).

A DoG filter is a simple band-pass filter whose passband width is controlled by the ratio $\sigma_1 : \sigma_2$. Let us consider combining several narrow band-pass DoG filters. If we define $\sigma_1 = \rho\sigma$ and $\sigma_2 = \sigma$ so that $\rho = \sigma_1/\sigma_2$, we find that a summation over DoG with standard deviations in the ratio ρ results in:

$$\begin{aligned} \sum_{n=0}^{N-1} G(x, y, \rho^{n+1}\sigma) - G(x, y, \rho^n\sigma) \\ = G(x, y, \sigma\rho^N) - G(x, y, \sigma) \end{aligned} \quad (6)$$

for an integer $N \geq 0$, which is simply the difference of two Gaussians (since all the terms except the first and last add up to zero) whose standard deviations can have any ratio $K = \rho^N$. That is, we can obtain the combined result of applying several band pass filters by choosing a DoG with a large K . If we assume that σ_1 and σ_2 are varied in such a way as to keep ρ constant at 1.6 (as needed for an ideal edge detector), then we essentially add up the output of several edge detectors (or selective band pass filters) at several image scales. This gives an intuitive understanding of why the salient regions will be fully covered and not just highlighted on edges or in the center of the regions.

4.3. Parameter selection

Based on the arguments in the previous section, a strategic selection of σ_1 and σ_2 will provide an appropriate band-pass filter to retain the desired spatial frequencies from the original image when computing the saliency map. With sufficiently long filters and a sufficiently large difference in σ_1 and σ_2 , the passband of the resulting band-pass filter given in Eq. 5 can be easily approximated from the two constituent Gaussians. With $\sigma_1 > \sigma_2$, ω_{lc} is determined by σ_1 and ω_{hc} is determined by σ_2 . However, use of filters of a practical length, providing a correspondingly simple implementation, renders this approximation inaccurate.

The two σ and therefore frequency parameters are therefore selected as follows. To implement a large ratio in standard deviations, we drive σ_1 to infinity. This results in a notch in frequency at DC while retaining all other frequencies. To remove high frequency noise and textures, we use a small Gaussian kernel keeping in mind the need for computational simplicity. For small kernels, the binomial filter approximates the Gaussian very well in the discrete case [6]. We use $\frac{1}{16}[1, 4, 6, 4, 1]$ giving $\omega_{hc} = \pi/2.75$. We therefore retain more than twice as much high-frequency content from the original image as GB and at least 40% more than SR.

4.4. Computing saliency

Our method of finding the saliency map S for an image I of width W and height H pixels can thus be formulated as:

$$S(x, y) = |I_\mu - I_{\omega_{hc}}(x, y)| \quad (7)$$

where I_μ is the arithmetic mean pixel value of the image and $I_{\omega_{hc}}$ is the Gaussian blurred version of the original image to eliminate fine texture details as well as noise and coding artifacts. The norm of the difference is used since we are interested only in the magnitude of the differences. This is computationally quite efficient (fourth criterion). Also, as we operate on the original image without any downsampling, we obtain a full resolution saliency map (last criterion).

To extend Eq. 7 to use features of color and luminance, we rewrite it as:

$$S(x, y) = \|\mathbf{I}_\mu - \mathbf{I}_{\omega_{hc}}(x, y)\| \quad (8)$$

where \mathbf{I}_μ is the mean image feature vector, $\mathbf{I}_{\omega_{hc}}(x, y)$ is the corresponding image pixel vector value in the Gaussian blurred version (using a 5×5 separable binomial kernel) of the original image, and $\|\cdot\|$ is the L_2 norm. Using the Lab color space, each pixel location is an $[L, a, b]^T$ vector, and the L_2 norm is the Euclidean distance. Our method, summarized in Eq. 8 allows us to fulfill all of the requirements for salient region detection listed earlier in this section.

5. Comparisons

The true usefulness of a saliency map is determined by the application. In this paper we consider the use of saliency maps in salient object segmentation. To segment a salient object, we need to binarize the saliency map such that ones (white pixels) correspond to salient object pixels while zeros (black pixels) correspond to the background¹.

We present comparisons with our method against the five methods mentioned above. In the first experiment, we use

¹For the four methods that give lower resolution saliency maps, bicubic interpolation is used to resize them to the original image size.

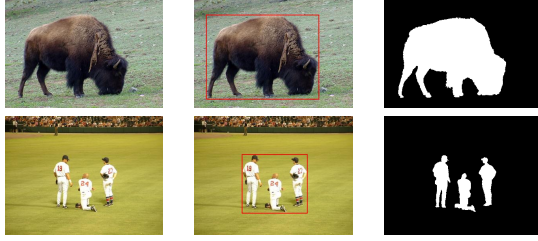


Figure 4. Ground truth examples. Left to Right, original image, ground truth rectangles from [28], and our ground truth, which is both more accurate and treats multiple objects separately.

a fixed threshold to binarize the saliency maps. In the second experiment, we perform image-adaptive binarization of saliency maps.

In order to obtain an objective comparison of segmentation results, we use a ground truth image database. We derived the database from the publicly available database used by Liu et al. [20]. This database provides bounding boxes drawn around salient regions by nine users. However, a bounding box-based ground truth is far from accurate, as also stated by Wang and Li [28]. Thus, we created an accurate object-contour based ground truth database² of 1000 images (examples in Fig. 4).

5.1. Segmentation by fixed thresholding

For a given saliency map, with saliency values in the range $[0, 255]$, the simplest way to obtain a binary mask for the salient object is to threshold the saliency map at a threshold T_f within $[0, 255]$. To compare the quality of the different saliency maps, we vary this threshold from 0 to 255, and compute the precision and recall at each value of the threshold. The resulting precision versus recall curve is shown in Fig. 5. This curve provides a reliable comparison of how well various saliency maps highlight salient regions in images. It is interesting to note that Itti’s method shows high accuracy for a very low recall (when $T_f > 240$), and then the accuracy drops steeply. This is because the salient pixels from this method fall well within salient regions and have near uniform values, but do not cover the entire salient object. Methods GB and AC have similar performance despite the fact that the latter generates full resolution maps as output. At maximum recall, all methods have the same low precision value. This happens at threshold zero, where all pixels from the saliency maps of each method are retained as positives, leading to an equal value for true and false positives for all methods.

5.2. Segmentation by adaptive thresholding

Maps generated by saliency detectors can be employed in salient object segmentation using more sophisticated

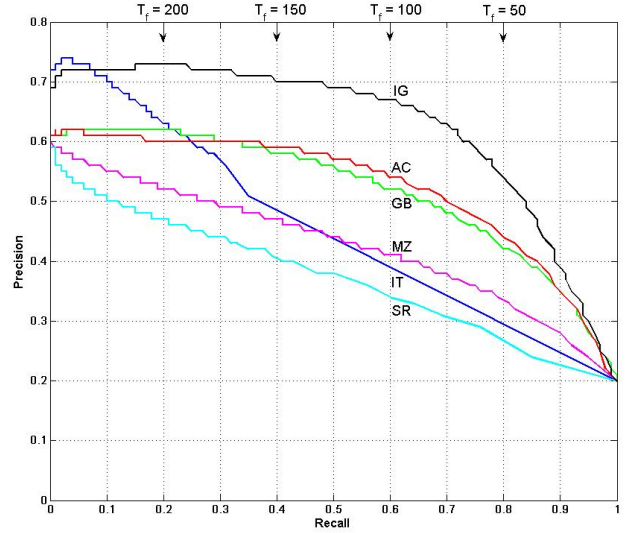


Figure 5. Precision-recall curve for naïve thresholding of saliency maps. Our method *IG* is compared against the five methods of *IT* [16], *MZ* [22], *GB* [10], *SR* [12], and *AC* [1] on 1000 images.

methods than simple thresholding. Saliency maps produced by Itti’s approach have been used in unsupervised object segmentation. Han et al. [9] use a Markov random field to integrate the seed values from Itti’s saliency map along with low-level features of color, texture, and edges to grow the salient object regions. Ko and Nam [18] utilize a Support Vector Machine trained on image segment features to select the salient regions of interest using Itti’s maps, which are then clustered to extract the salient objects. Ma and Zhang [22] use fuzzy growing on their saliency maps to confine salient regions within a rectangular region.

We use a simpler method for segmenting salient objects, which is a modified version of that presented in [1]. Their technique makes use of the intensity and color properties of the pixels along with their saliency values to segment the object. Considering the full resolution saliency map, their technique over-segments the input image using k-means clustering and retains only those segments whose average saliency is greater than a constant threshold. The binary maps representing the salient object are thus obtained by assigning ones to pixels of chosen segments and zeroes to the rest of the pixels.

We make two improvements to this method. First, we replace the hill-climbing based k-means segmentation algorithm by the mean-shift segmentation algorithm [5], which provides better segmentation boundaries. We perform mean-shift segmentation in *Lab* color space. We use fixed parameters of 7, 10, 20 for σ_{color} , σ_{spatial} , and \minRegion , respectively, for all the images (see [5]).

We also introduce an adaptive threshold that is image saliency dependent, instead of using a constant threshold for each image. This is similar to the adaptive threshold pro-

²http://ivrg.epfl.ch/supplementary_material/RK_CVPR09/index.html

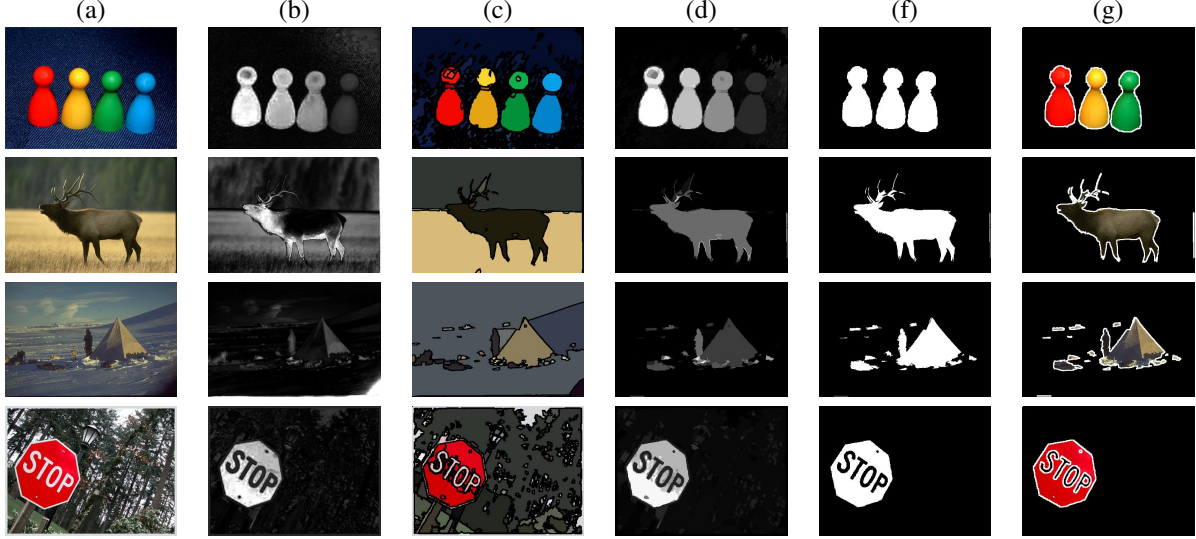


Figure 6. (a) is the original image. The average saliency per segment (d) is computed using the saliency map (b) and the mean-shift segmented image (c). Those segments that have a saliency value greater than the adaptive threshold computed in Eq. 9 are assigned ones (white) and the rest zeroes (black) in (f). The salient objects resulting from binary map (f) are shown in (g).

posed by Hou and Zhang [12] to detect proto-objects. The adaptive threshold (T_a) value is determined as two times the mean saliency of a given image:

$$T_a = \frac{2}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} S(x, y) \quad (9)$$

where W and H are the width and height of the saliency map in pixels, respectively, and $S(x, y)$ is the saliency value of the pixel at position (x, y) . A few results of salient object segmentation using our improvements are shown in Fig. 6.

Using this modified approach, we obtain binarized maps of salient object from each of the saliency algorithms. Average values of precision, recall, and F-Measure (Eq. 10) are obtained over the same ground-truth database used in the previous experiment.

$$F_\beta = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (10)$$

We use $\beta^2 = 0.3$ in our work to weigh precision more than recall. The comparison is shown in Fig. 7. Itti's method (IT) shows a high precision but very poor recall, indicating that it is better suited for gaze-tracking experiments, but perhaps not well suited for salient object segmentation. Among all the methods, our method (IG) shows the highest precision, recall, and F_β values.

Our method clearly outperforms alternate, state-of-the-art algorithms. However, like all saliency detection methods, it can fail if the object of interest is not distinct from the background in terms of visual contrast (see Fig 6(b), first row). Also, to fulfill the first criterion of saliency in Sec. 4.1 our method has a preference for larger salient objects

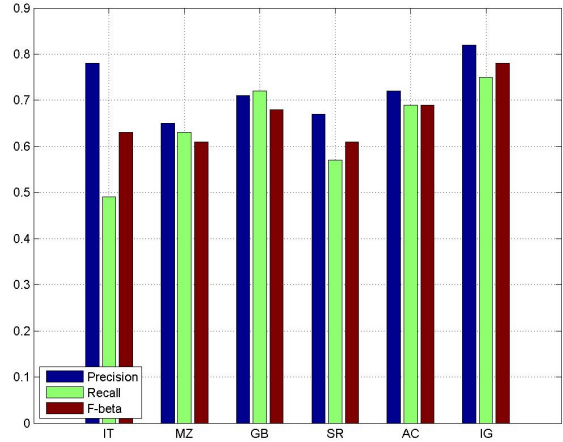


Figure 7. Precision-Recall bars for binarization of saliency maps using the method of Achanta et al [1]. Our method IG shows high precision, recall and F_β values on the 1000 image database.

(which comes from our choice of filters). In addition to this, segmentation of the salient object itself can fail despite having a good saliency map if the mean-shift pre-segmentation is flawed.

6. Conclusions

We performed a frequency-domain analysis on five state-of-the-art saliency methods, and compared the spatial frequency content retained from the original image, which is then used in the computation of the saliency maps. This analysis illustrated that the deficiencies of these techniques arise from the use of an inappropriate range of spatial fre-

quencies. Based on this analysis, we presented a frequency-tuned approach of computing saliency in images using low level features of color and luminance, which is easy to implement, fast, and provides full resolution saliency maps. The resulting saliency maps are better suited to salient object segmentation, demonstrating both higher precision and better recall than the five state-of-the-art techniques.

7. Acknowledgements

This work is in part supported by the National Competence Center in Research on Mobile Information and Communication Systems (NCCR-MICS), a center supported by the Swiss National Science Foundation under grant number 5005-67322, the European Commission under contract FP6-027026 (K-Space), and the PHAROS project funded by the European Commission under the 6th Framework Programme (IST Contract No. 045035).

References

- [1] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk. Salient region detection and segmentation. *International Conference on Computer Vision Systems*, 2008.
- [2] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Transactions on Graphics*, 26(3), 2007.
- [3] N. Bruce and J. Tsotsos. Attention based on information maximization. *Journal of Vision*, 7(9):950–950, 2007.
- [4] C. Christopoulos, A. Skodras, A. Koike, and T. Ebrahimi. The JPEG2000 still image coding system: An overview. *IEEE Transactions on Consumer Electronics*, 46(4):1103–1127, 2000.
- [5] C. Christoudias, B. Georgescu, and P. Meer. Synergism in low level vision. *IEEE Conference on Pattern Recognition*, 2002.
- [6] J. L. Crowley, O. Riff, and J. H. Piater. Fast computation of characteristic scale using a half octave pyramid. *International Conference on Scale-Space theories in Computer Vision*, 2003.
- [7] S. Frintrop, M. Klodt, and E. Rome. A real-time visual attention system using integral images. *International Conference on Computer Vision Systems*, 2007.
- [8] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. *IEEE Conference on Computer Vision*, 2007.
- [9] J. Han, K. Ngan, M. Li, and H. Zhang. Unsupervised extraction of visual attention objects in color images. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(1):141–145, 2006.
- [10] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in Neural Information Processing Systems*, 19:545–552, 2007.
- [11] S. S. Hemami and T. N. Pappas. Perceptual metrics for image quality evaluation. Tutorial presented at IS&T/SPIE Human Vision and Electronic Imaging, 2007.
- [12] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [13] Y. Hu, X. Xie, W.-Y. Ma, L.-T. Chia, and D. Rajan. Salient region detection using weighted feature maps based on the human visual attention model. *Pacific Rim Conference on Multimedia*, 2004.
- [14] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, 19:547–554, 2005.
- [15] L. Itti and C. Koch. Comparison of feature combination strategies for saliency-based visual attention systems. *SPIE Human Vision and Electronic Imaging IV*, 3644(1):473–482, 1999.
- [16] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
- [17] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. *European Conference on Computer Vision*, 2004.
- [18] B. C. Ko and J.-Y. Nam. Object-of-interest image segmentation based on human attention and semantic region clustering. *Journal of Optical Society of America A*, 23(10):2462–2470, 2006.
- [19] C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227, 1985.
- [20] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [21] D. G. Lowe. Distinctive image features from scale-invariant feature points. *International Journal of Computer Vision*, 60:91–110, 2004.
- [22] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM International Conference on Multimedia*, 2003.
- [23] D. Marr. *Vision: a computational investigation into the human representation and processing of visual information*. W. H. Freeman, San Francisco, 1982.
- [24] E. Niebur and C. Koch. *The Attentive Brain*, chapter Computational architectures for attention, pages 163–186. Cambridge MA:MIT Press, October 1995.
- [25] B. Olshausen, C. Anderson, and D. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13:4700–4719, 1993.
- [26] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 2004.
- [27] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nufo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.
- [28] Z. Wang and B. Li. A two-stage approach to saliency detection in images. *IEEE Conference on Acoustics, Speech and Signal Processing*, 2008.