Visual saliency in noisy images

Chelhwon Kim

Electrical Engineering Department, University of California, Santa Cruz, CA, USA

Peyman Milanfar

Electrical Engineering Department, University of California, Santa Cruz, CA, USA, Google, Inc., Mountain View, CA, USA

The human visual system possesses the remarkable ability to pick out salient objects in images. Even more impressive is its ability to do the very same in the presence of disturbances. In particular, the ability persists despite the presence of noise, poor weather, and other impediments to perfect vision. Meanwhile, noise can significantly degrade the accuracy of automated computational saliency detection algorithms. In this article, we set out to remedy this shortcoming. Existing computational saliency models generally assume that the given image is clean, and a fundamental and explicit treatment of saliency in noisy images is missing from the literature. Here we propose a novel and statistically sound method for estimating saliency based on a nonparametric regression framework and investigate the stability of saliency models for noisy images and analyze how state-of-the-art computational models respond to noisy visual stimuli. The proposed model of saliency at a pixel of interest is a data-dependent weighted average of dissimilarities between a center patch around that pixel and other patches. To further enhance the degree of accuracy in predicting the human fixations and of stability to noise, we incorporate a global and multiscale approach by extending the local analysis window to the entire input image, even further to multiple scaled copies of the image. Our method consistently outperforms six other state-of-the-art models (Bruce & Tsotsos, 2009; Garcia-Diaz et al., 2012; Goferman et al., 2010; Hou & Zhang, 2007; Itti, Koch, & Niebur, 1998; Seo & Milanfar, 2009; Seo & Milanfar, 2009) for both noise-free and noisy cases.

Introduction

Visual saliency is an important aspect of human vision, as it directs our attention to what we want to perceive. It also affects the processing of information in that it allocates limited perceptual resources to objects of interest and suppresses our awareness of areas worth ignoring in our visual field. In computer vision tasks, finding salient regions in the visual field is also essential because it allows computer vision systems to process a flood of visual information and allocate limited resources to relatively small but interesting regions or a few objects. In recent years, extensive research has focused on finding saliency in natural images and predicting where humans look in the image. As such, a wide diversity of computational saliency models have been introduced (Bruce & Tsotsos, 2009; Gao, Mahadevan, & Vasoncelos, 2008; Garcia-Diaz et al., 2012; Goferman et al., 2010; Hou & Zhang, 2007; Itti, Koch, & Niebur, 1998; Seo & Milanfar, 2009; Zhang et al., 2008) and are aimed at transforming a given image into a scalar-valued map (the saliency map) representing visual saliency in that image. This saliency map has been useful in many applications, such as object detection (Rosin, 2009; Rutishauser, Walther, Koch, & Perona, 2004; Seo & Milanfar, 2010; Zhicheng & Itti, 2011), image quality assessment (Ma & Zhang, 2008; Niassi, LeMeur, Lecallet, & Barba, 2007), and action detection (Seo & Milanfar, 2009) and more.

Most saliency models are biologically inspired and based on a bottom-up computational model. Itti et al. (1998) introduced a model based on the biologically plausible architecture proposed by Koch and Ullman (1985) and measure center-surround contrast using a difference of Gaussians approach. Bruce and Tsotsos (2009) proposed the Attention based on Information Maximization (AIM) model. They measured saliency at a pixel in the image by Shannon’s self-information of that location with respect to its surrounding context. To estimate the probability density of a visual feature in the high dimensional space, they employed a representation based on independent components, which are determined from natural scenes. The saliency model of Zhang et al. (2008) uses natural image statistics within a Bayesian framework from which bottom-up saliency...
emerges naturally as the self-information of visual features. Seo and Milanfar (2009) proposed the self-resemblance mechanism to measure saliency. At each pixel, they first extract visual features (local regression kernels) that are robust in extracting local geometry of the image. Then, matrix cosine similarity (Seo & Milanfar, 2009, 2010) is employed to measure the resemblance of each pixel to its surroundings. Hou and Zhang (2007) derived saliency by measuring the spectral residual of an image, which is the difference between the log spectrum of the image and its smoothed version. They posited that the statistical singularities in the spectrum may be responsible for anomalous regions in the image. Goferman et al. (2010) proposed context-aware saliency. Their saliency model aims to detect not only the dominant objects but also the parts of their surroundings that convey the context. This type of model is useful in applications in which the context of the dominant objects is just as essential as the objects themselves. Garcia-Diaz et al. (2012) proposed the Adaptive Whitening Saliency (AWS) model. The whitening process (decorrelation and variance normalization) is applied to the chromatic components of the image. Then, the multioriented and multiscale local energy representation of the image is obtained by applying a bank of log-Gabor filters parameterized by different scales and orientations. Visual saliency is measured by a simple vector norm computation in the obtained representation.

Despite the wide variety of computational saliency models, they all assume the given image is clean and free of distortions. However, as in Figure 1, when we feed a noisy image instead of a clean image into existing saliency detection algorithms, many fail, frivolously declaring saliency in the noisy image. Especially, for the model by Bruce and Tsotsos (2009), Goferman et al. (2010), Seo and Milanfar (2009), and Zhang et al. (2008), it is apparent that any applications using these noisy saliency maps cannot perform well. In contrast, Garcia-Diaz et al. (2012) and Hou and Zhang (2007) provide more stable results because they implicitly suppress the noise during the process of computing saliency. In the model by Hou and Zhang, spectral filtering of the image will suppress the noise in the image. The AWS by Garcia-Diaz et al. (2012), incorporating multioriented and multiscale representation with their whitening process, also implicitly suppresses the noise. Although their results tend to be apparently somewhat insensitive to noise, a fundamental and explicit treatment of saliency in noisy images is missing from the literature (Le Meur, 2011). We shall provide this in this article. Furthermore, we will demonstrate that the price for this apparent insensitivity to noise is that the overall performance over a large range of noise strengths is diminished. In this article, we aim to achieve two goals simultaneously. First, we propose a simple and statistically well-motivated computational saliency model that achieves a high degree of accuracy in predicting where humans look. Second, we illustrate that the proposed model is stable when a noise-corrupted image is given and improves on other state-of-the-art models over a large range of noise strengths.

The proposed saliency model is based on a bottom-up computational model. As such, an underlying hypothesis is that human eye fixations are driven to conspicuous regions in the test image, which stand out from their surroundings. To measure this distinctiveness of region, we observe dissimilarities between a center patch of the region and other patches (Figure 2). Once we have measured these dissimilarities, the problem of interest is how to aggregate them to obtain an estimate of the underlying saliency of that region. We look at this problem from an estimation theory point of view and propose a novel and statistically sound saliency model. We assume that each observed dissimilarity has an underlying true value, which is measured with uncertainty. Given these noisy observations, we estimate the underlying saliency by solving a local data-dependent weighted least squares problem. As we will see in the next section, this results in an aggregation of the dissimilarities with weights depending on a kernel function to be specified. We define the kernel function so that it gives higher weight to similar patch pairs than dissimilar patch pairs. Giving higher weights to more similar patch pairs would seem counterintuitive at first.
But this process will ensure that only truly salient objects would be declared so, sparing us from too many false declarations of saliency. The proposed estimate of saliency at pixel $x_j$ is defined as:

$$\hat{s}(x_j) = \frac{\sum_{i=1}^{N} w_{ij} y_i}{\sum_{i=1}^{N} w_{ij}}$$

where $y_i$ and $w_{ij}$ are the observed dissimilarity (to be defined shortly in the next section) and the weight for the $ij$-th patch pair, respectively.

It is important to highlight the direct relation of our approach to two earlier approaches of Seo and Milanfar and Goferman et al. We make this comparison explicit here because these methods also involve aggregation of local dissimilarities. Although this was not made entirely clear in either Goferman et al. (2010) or Seo and Milanfar (2009), it is interesting to note that these methods employed arithmetic and harmonic averaging of local dissimilarities, respectively. In Seo and Milanfar (2009), they defined the estimate of saliency at pixel $x_j$ by

$$\hat{s}(x_j) = \exp\left(\frac{1}{N} \sum_{i=1}^{N} \frac{1}{y_i}\right)$$

where $y_i = \exp(-\rho_i/\tau)$ and $\rho_i$ is the cosine similarity between visual features extracted from the center patch around the pixel $x_j$ and its $i$-th nearby patch. This saliency model is (to within a constant) the harmonic mean of dissimilarities, $y_i$’s.

Goferman et al. (2010) formularized the saliency at pixel $x_j$ as

$$\hat{s}(x_j) = 1 - \exp\left(-\frac{1}{N} \sum_{i=1}^{N} y_i\right)$$

where $y_i$ is the dissimilarity measure between a center patch around the pixel $x_j$ and any other patch observed in the test image. This saliency model is the arithmetic mean of $y_i$’s. Besides the use of the exponential, the important difference as compared with our approach is that they use constant weights $w_{ij} = 1/N$ for the aggregation of dissimilarities, whereas we use data-dependent weights.

In summary, among those saliency models in which dissimilarities (either local or global) are combined by different aggregation techniques, our proposed method is simpler, better justified, and indeed a more effective arithmetic aggregation based on kernel regression.

Many saliency models have leveraged the multiscale approach (Gao et al., 2008; Garcia-Diaz et al., 2012;...
Goferman et al., 2010; Walther & Koch, 2006; Zhang et al., 2008; Zhicheng & Itti, 2011). In the proposed model, we also exploit the global and multiscale approach by extending the window to the whole image. By doing so, we enhance the degree of accuracy in predicting human fixations and further realize strong stability to noise as well.

The article is organized as follows. In the next section, we provide further technical details about the proposed saliency model and describe the global and multiscale approach to the saliency computational model. In the Performance Evaluation section, we demonstrate the efficacy of this saliency model in predicting human fixations with six other state-of-the-art models (Bruce & Tsotsos, 2009; Garcia-Diaz et al., 2012; Goferman et al., 2010; Hou & Zhang, 2007; Seo & Milanfar, 2009; Zhang et al., 2008) and investigate the stability of our method in the presence of noise. In the last section, we conclude the article.

Technical details

Nonparametric regression for saliency

In this section, we propose a measure of saliency at a pixel of interest from observations of dissimilarity between a center patch around the pixel and its nearby patches (see Figure 2). Let us denote by \( \rho \), the similarity between a patch centered at a pixel of interest and its \( i \)-th neighboring patch. Then, the dissimilarity is measured as a decreasing function of \( \rho \) as follows:

\[
y_i = e^{-\rho_i}, \quad i = 1, \ldots, N
\]

The similarity function \( \rho \) can be measured in a variety of ways (Rubner, Tomasi, & Guibas, 2000; Seo & Milanfar, 2009; Swain & Ballard, 1991), for instance, using the matrix cosine similarity between visual features computed in the two patches (Seo & Milanfar, 2009, 2010). For our experiments, we shall use the LARK features as defined in Takeda, Farsiu, and Milanfar (2007), which have been shown to be robust to the presence of noise and other distortions. Much detailed description of these features is given in Takeda et al. (2007) and Takeda, Milanfar, Proctor, and Elad (2009). We note that the effectiveness of LARK as a visual descriptor has led to its use for object and action detection and recognition, even in the presence of significant noise (Seo & Milanfar, 2009, 2010). From an estimation theory point of view, we assume that each observation \( y_i \) is in essence a measurement of the true saliency but measured with some error. This observation model can be posed as:

\[
y_i = s_i + \eta_i, \quad i = 1, \ldots, N
\]

where \( \eta_i \) is noise. Given these observations, we assume a locally constant model of saliency and estimate the expected saliency at pixel \( x_j \) by solving the weighted least squares problem

\[
\hat{s}(x_j) = \arg\min_{s_j} \sum_{i=1}^{N} [y_i - s(x_j)]^2 K(y_i, y_r)
\]

where \( y_r \) is a reference observation. We choose \( y_r \) where \( i = 1, \ldots, N \) ranges in a neighborhood of \( j \). As such, \( y_r \) is the most similar patch to the patch at \( j \). Depending on the difference between this reference observation \( y_r \) and each observation \( y_p \), the kernel function \( K(\cdot) \) gives higher or lower weight to each observation as follows:

\[
K(y_i, y_r) = e^{-\frac{(y_i - y_r)^2}{2}}
\]

Therefore, the weight function gives higher weight to similar patch pairs than dissimilar patch pairs. The rationale behind this way of weighting is to avoid easily declaring saliency; that is, the aggregation of dissimilarities for a truly salient region should be still high even if we put more weight on the most similar patch pairs. Put yet another way, we do not easily allow any region to be declared salient, and thus we reduce the likelihood of false alarms. We set the weight of the reference observation itself, \( w_r = \max_i w_i \). This setting avoids the excessive weighting of the reference observation in the average. The parameter \( h \) controls the decay of the weights and is determined empirically to get best performance.

Minimizing Equation 6, the result is merely a weighted average of the measured dissimilarities, where the weights are computed based on distances between each observation and the reference observation,

\[
\hat{s}(x_j) = \frac{\sum_{i=1}^{N} K(y_i, y_r) y_i}{\sum_{i=1}^{N} K(y_i, y_r)} = \sum_{i=1}^{N} w_{yr} y_i
\]

Global and multiscale saliency

So far, the underlying idea is that the saliency of a pixel is measured by the distinctiveness of a center patch around that pixel relative to its neighbors. In this section, we extend our local analysis window (gray dashed rectangle in Figure 2) to the entire input image. By doing this, we aggregate all dissimilarities between the center patch and all patches observed from the entire image. This is a sensible and well-motivated extension because, in general, it is consistent with the way the human visual system inspects the global field of view at once to determine saliency. Furthermore, we
incorporate a multiscale approach by taking the patches (to be compared to the center patch) from the multiscale Gaussian pyramid constructed from the given image. Figure 3 illustrates the global and multiscale saliency computation. In our implementation, we follow the same general procedure as in Goferman et al. (2010). First, we denote by $R = \{r_1, r_2, \ldots, r_M\}$ the multiple scales applied to the input image (the horizontal axis in Figure 3), and then at each scale $r_m$, where $1 \leq m \leq M$, we compute the dissimilarity of the center patch relative to all patches observed in the images whose scales are $R_q = \{r_m, r_m/2, r_m/4\}$ (the vertical axis in Figure 3). Consequently, for $M$ scales, $M$ saliency maps are computed and resized to the original image size by bilinear interpolation. The resulting multiple saliency maps are then combined into one by simple averaging. Figure 4 demonstrates the difference between those saliency maps obtained at different scales. Although the fine scale result detects details such as textures and edges, the coarse scale result detects global features. Note that we fixed the size of the patch at each scale $r_m$ as the yellow rectangle shown in Figure 3.

To rewrite the saliency equation in a multiscale fashion, we denote again the dissimilarity measure defined in Equation 4 by $y_i = e^{-\rho(p_i, p)}$, where $p_i$ is the $i$-th patch observed across the multiscale pyramid and $p$ is the center patch at pixel $x$. Therefore, we can rewrite the saliency equation at each scale $r_m$ as follows:

![Figure 3. Global and multiscale saliency computation. At each scale $r_m \in R$ (column), we search all patches to be compared to the center patch (yellow rectangle) across multiple images whose scales are $R_q = \{r_m, r_m/2, r_m/4\}$.](image1)

![Figure 4. The saliency maps obtained at different scales. The multiscale approach not only gives high saliency values at object edges (from the fine scale result) but also detects global features (from the coarse scale result).](image2)
The saliency at pixel $x_j$ is taken as the mean of its saliency across all scales:

$$\hat{s}(x_j) = \frac{1}{M} \sum_{m=1}^{M} \hat{s}(x_j)^{(m)}$$  \hspace{1cm} (10)$$

In the next section, we first evaluate our saliency model for clean images against six existing saliency models (Bruce & Tsotsos, 2009; Garcia-Diaz et al., 2012; Goferman et al., 2010; Hou & Zhang, 2007; Seo & Milanfar, 2009; Zhang et al., 2008) and then investigate the stability of our saliency model for noisy images. We also see the effect of the global and multiscale approach on overall performance.

### Performance evaluation

#### Predicting human fixation data

In this section, we evaluate the proposed saliency model in predicting human eye fixations on Bruce and Tsotsos’s (2009) data set (available at http://www.cs.umanitoba.ca/~bruce/). This is a data set of 120 indoor and outdoor natural images and has been commonly used to validate many state-of-the-art saliency models (Bruce & Tsotsos, 2009; Garcia-Diaz et al., 2012; Seo & Milanfar, 2009; Zhang et al., 2008). The subjects were given no instructions except to observe the images, and the eye fixations were recorded during 4 s (Bruce & Tsotsos, 2009).

The six state-of-the-art models used for comparison are the Saliency Using Natural statistics model by Zhang et al. (2008), the AIM model by Bruce and Tsotsos (2009), the Spectral Residual model by Hou and Zhang (2007), the Context Aware model by Goferman et al. (2010), the Self-resemblance model by Seo and Milanfar (2009), and the AWS model by Garcia-Diaz et al. (2012). For each model, we used the default parameters suggested by the respective authors.

In our implementation, the similarity function $\rho$ in Equation 4 is computed using the matrix cosine similarity between the LARK features as in the model of Seo and Milanfar (2009). We sample patches of $7 \times 7$ with 50% overlap from multiple scale images. We use three scales ($M = 3$), $R = \{1.0, 0.6, 0.4\}$, and the smallest scale allowed in $R_q$ is 20% of the original size, as in Goferman et al. (2010).

Figure 5 demonstrates a qualitative comparison of the proposed model with the fixation density map and the saliency maps produced by the six competing models. All saliency maps were normalized to range between zero and one to make a comparison with equalized contrast. There is qualitative similarity between Seo and Milanfar (2009) and our model, except that ours is seen to have fewer spurious salient regions. We also note that our global and multiscale approach also contributes to the advantage of our model. More precisely, as similarly argued in Goferman et al. (2010), background pixels are likely to have similar patches in the entire image at multiple scales, whereas the salient pixels have similar patches in the nearby region and at a few scales. Therefore, incorporating global and multiscale approach not only emphasizes the contrast between salient and nonsalient regions but also suppresses frequently occurring features in the background.

Figure 6 shows three different test images, each of which has one or two salient objects. The saliency models with the global and multiscale considerations such as Goferman et al.’s model and the proposed model produce more reliable results than others. It seems that Goferman et al.’s model suppress saliencies on the frequently occurring features more efficiently than ours. However, we note that Goferman et al. simulated the visual contextual effect by identifying the attended areas where the saliency value exceeds a certain threshold and weighting each pixel outside the attended areas according to its Euclidean distance to the closest attended pixel. This apparently suppresses more saliencies on the features outside the attended areas such as the bricks in the first test image. By contrast, the output of Zhang et al. (2008), Bruce and Tsotsos (2009), and Seo and Milanfar (2009) found high saliency in the uninteresting background. In particular, the saliency model by Seo and Milanfar is seen to be most sensitive to the frequently occurring features. Although it seems that Hou and Zhang’s (2007) model is also robust to the frequently occurring features as seen in the first and third image, it still declares high saliency values in the pile of green peppers.

For quantitative performance analysis, we use area under the receiver-operating characteristic curve (AUC) and Spearman’s rank correlation coefficient (SCC). The AUC metric determines how well fixated and nonfixated locations can be discriminated by the saliency map using a simple threshold (Tatler, Baddeley, & Gilchrist, 2010). If the values of the saliency map exceed the threshold, then we declare them as fixated. By sweeping the threshold between the minimum and maximum values in the saliency map, the true positive rate (declaring fixated locations as fixated) and the false-positive rate (declaring nonfixated locations as fixated) are calculated, and the receiver-operating characteristic curve is constructed by plotting the true-positive rate as a function of the false-positive rate across all possible thresholds. The SCC metric measures the degree of similarity between two ranked saliency and fixation density maps (see Figure 5 for...
examples of the fixation density map). If they are not well matched, the correlation coefficient is zero.

Zhang et al. (2008) pointed out two problems in using the AUC metric: First, simply using a Gaussian blob centered in the middle of the image as the saliency map produces excellent results because most human eye fixation data have a center bias as photographers tend to place objects of interest in the center (Parkhurst &
Table 2. AUC for the various standard deviations (std) from the original fixation data.

<table>
<thead>
<tr>
<th>Model</th>
<th>std(0)</th>
<th>std(5)</th>
<th>std(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method</td>
<td>0.713</td>
<td>0.713</td>
<td>0.709</td>
</tr>
<tr>
<td>Garcia-Diaz et al. (2012)</td>
<td>0.714</td>
<td>0.713</td>
<td>0.710</td>
</tr>
<tr>
<td>Seo and Milanfar (2009)</td>
<td>0.696</td>
<td>0.695</td>
<td>0.693</td>
</tr>
<tr>
<td>Goferman et al. (2010)</td>
<td>0.686</td>
<td>0.686</td>
<td>0.681</td>
</tr>
<tr>
<td>Hou and Zhang (2007)</td>
<td>0.672</td>
<td>0.670</td>
<td>0.668</td>
</tr>
<tr>
<td>Bruce and Tsotsos (2009)</td>
<td>0.672</td>
<td>0.672</td>
<td>0.672</td>
</tr>
<tr>
<td>Zhang et al. (2008)</td>
<td>0.639</td>
<td>0.638</td>
<td>0.638</td>
</tr>
</tbody>
</table>

In the next section, we will see that the proposed model is more stable than others when the input images are corrupted by noise and thus produces better performance overall across a large range of noise strengths.

**Stability of saliency models for noisy images**

In this section, we investigate the stability of saliency models for noisy images. The same original test images from Bruce and Tsotsos’s (2009) data set are used, and the noise added to the test images is white Gaussian with variance $\sigma^2$, which equals 0.01, 0.05, 0.1, or 0.2 (the intensity value for each pixel of the image ranges from 0 to 1). The saliency maps computed from the noisy images are compared to the human fixations through the same procedure. One may be concerned that the human fixations used in this evaluation were recorded from noise-free images and not the corrupted images. However, we focus on investigating the sensitivity of computational models of visual attention subjected to visual degradations rather than evaluating the performance in predicting human fixation data in noisy images. Therefore, we use the same human fixations to see if the computational models achieve the same performance as in the noise-free case. Also, to the best of our knowledge, there is no available public fixation database on noisy images. So, we resorted instead to analyzing how state-of-the-art computational models respond to noisy visual stimuli.

Examples of noisy test images and their saliency maps are depicted in Figure 7. We observe that the proposed model shows more stable responses in flat regions in the background such as sky, road, and wall than in the regions containing high-frequency textured areas such as leaves of a tree. This phenomenon can be explained as follows: Background pixels in such flat regions are likely to have more similar patches in the entire image, whereas salient pixels have similar patches in the nearby region. Furthermore, because we give higher weights to similar patch pairs than dissimilar ones when we aggregate the dissimilarities of those
Figure 7. Saliency maps produced by the proposed method on increasingly noisy images. From left to right, a clean image, and noisy images with noise variance $\sigma^2 = \{0.01, 0.05, 0.1, 0.2\}$, respectively.
patches, we tend to average more dissimilarities for the background pixels and suppress more noise than on the salient pixels.

For quantitative performance analysis, we plot the AUC values for each method against the noise strength. As one may expect, the performance in predicting human fixations generally decreases as the noise strength increases (see the curves in Figure 8). However, our saliency model outperforms the six other state-of-the-art models over a wide range of noise strengths. Only Garcia-Diaz et al.’s (2012) model shows similar performance for the noise-free case, but the proposed model shows better performance for the noisy case.

As we alluded to earlier, most saliency models implicitly suppress the noise by blurring and down-sampling the input image. Hou and Zhang (2007) and Seo and Milanfar (2009) down-sampled the input image to 64 × 64. Bruce and Tsotsos (2009), Zhang et al. (2008), and Garcia-Diaz et al. (2012) also used an input image down-sampled by a factor of 2. In Goferman et al.’s (2010) model, the input image was down-sampled to 250 pixels. However, as illustrated in Figure 8, the price for this implicit treatment is that the overall performance over a large range of noise strengths is diminished, except in Hou and Zhang’s (2007) model. Because Hou and Zhang removed redundancies in the frequency domain after the input image was down-sampled, they suppressed more noise and showed stable results. However, we note that their method does not achieve a high degree of accuracy overall in predicting human fixations. In contrast, our regression-based saliency model achieves a high degree of accuracy for noise-free and noisy cases simultaneously and improves on competing models over a large range of noise strengths.

We investigated how state-of-the-art computational models respond to noisy visual stimuli. Based on the Helmholtz principle (Desolneux, Moisan, & Morel,
the human visual system does not perceive structure in a uniform random image. Only when some relatively large deviation from randomness occurs is a structure perceived. According to this principle, the bottom-up approaches should result in roughly similar saliency maps to those produced using clean images because the random features in the input image are largely suppressed. That is to say, a good computational saliency model should behave similarly in the presence of noise and return stable results. We made several noisy synthetic images by adding different amounts of white Gaussian noise to a 128 × 128 gray image containing a 19 × 19 black square in the center. The saliency maps computed from these noisy synthetic images are normalized to range from zero to one. We note that the input images were not down-sampled or blurred before calculating the saliency map, and thus the implicit noise suppression was not included in this experiment. Figure 9 shows results produced by the six other state-of-the-art models and the proposed model. Only Garcia-Diaz et al.’s (2012) model and the proposed model remained robust to the noise in the saliency map. We also observed that the saliency maps from Seo and Milanfar’s (2009) model and Hou and Zhang’s (2007) model in the second row (noise variance, 0.05) are severely degraded compared with the ones in Figure 1. In addition, Hou and Zhang’s model detects only details (notice the white in the boundary of the square), and Zhang et al.’s (2008) model does not respond to the black square of a given size. We believe that each model has different inherent sensitivity to noise and different responses to image features at a given scale. Therefore, different degrees of blurring and down-sampling will no doubt affect the result of each model differently. To investigate the inherent sensitivity of each model to noise, we performed the same evaluation on the saliency maps but with the same degree of resizing and blurring applied to input images. To do this, we down-sampled all the images to the same size of 250 pixels. Figure 10 shows the performance in predicting the human fixations. We observed that the proposed model still outperforms other models and achieves a high degree of accuracy for both noise-free and noisy cases.

Finally, for the sake of completeness, we show the effect of the global and multiscale approach on our saliency computational model. To this end, we first evaluated the proposed model without the global and multiscale approach. In other words, we considered only the patches in the 7 × 7 local window to measure the dissimilarities for each pixel (we denote this by “Local + Single-scale” in Figure 11). Then, we extended the local analysis window to the entire image and evaluated it again (“Global + Single-scale”). Last, we further observed those patches from multiple scale images (“Global + Multi-scale”). As seen in Figure 11,

Figure 9. Examples of saliency on noisy synthetic images.
we can get better performance with this global approach. The multiple scale approach also improves the performance, but the amount of improvement is not as significant as that obtained by the global approach.

**Conclusion and future work**

In this article, we have proposed a simple and statistically well-motivated saliency model based on nonparametric regression, which is a data-dependent weighted combination of dissimilarities observed in the given image. The proposed method is practically appealing and effective because of its simple mathematical form. To enhance its performance, we incorporate the global and multiscale approach by extending the local analysis window to the entire input image, even further to multiple scaled copies of the image. Experiments on challenging sets of human fixations data demonstrate that the proposed saliency model not only achieves a high degree of accuracy in the standard noise-free scenario but also improves on other state-of-the-art models for noisy images. Because of its robustness to noise, we expect the proposed model to be quite effective in other computer vision applications subject to severe degradation by noise.

We investigated how different computational saliency models predict human fixations on images corrupted by white Gaussian noise. For future work, it would be interesting to investigate how they do on other type of distortions such as blur, low resolution, snow, rain, or air turbulence, which occur often in real-world applications. In addition, it would be interesting to preprocess degraded data before attempting to calculate the saliency map. As observed in our earlier work (Kim & Milanfar, 2012), the performance of saliency models can be improved by applying a denoising approach first. Unfortunately, this is not consistent with the way the human visual system operates; thus, an algorithm based on filtering first would not seem to be well-motivated by biology. In any case, the described model offers a simple yet effective approach to address this problem.
event, imperfect de-noising might further distort the data, and thus this is at best suboptimal.

**Keywords:** saliency, nonparametric regression, saliency for noisy images, global approach, multiscale approach

**Acknowledgements**

Source code for the proposed algorithm is available at the following link: http://users.soe.ucsc.edu/~chkim/SaliencyDetection.html. This work was supported by Air Force Office of Scientific Research Grant FA9550-07-1-0365 and National Science Foundation Grant CCF-1016018.

Commercial relationships: none.
Corresponding author: Chelhwon Kim.
Email: chkim@soe.ucsc.edu.
Address: Electrical Engineering Department, University of California, Santa Cruz, CA, USA.

**References**


