Finding Saliency in Noisy Images

Chelhwon Kim and Peyman Milanfar

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

ABSTRACT

Recently, many computational saliency models have been introduced[2,5,7,13,23] to transform a given image into a scalar-valued map that represents visual saliency of the input image. These approaches, however, generally assume the given image is clean. Fortunately, most methods implicitly suppress the noise before calculating the saliency by blurring and downsampling the input image, and therefore tend to be apparently rather insensitive to noise.[11] However, a fundamental and explicit treatment of saliency in noisy images is missing from the literature. Indeed, as we will show, the price for this apparent insensitivity to noise is that the overall performance over a large range of noise strengths is diminished. Accordingly, the question is how to compute saliency in a reliable way when a noise-corrupted image is given. To address this problem, we propose a novel and statistically sound method for estimating saliency based on a non-parametric regression framework. The proposed estimate of the saliency at a pixel is a data-dependent weighted average of dissimilarities between a center patch and its surrounding patches. This aggregation of the dissimilarities is simple and more stable despite the presence of noise. For comparison's sake, we apply a state of the art denoising approach before attempting to calculate the saliency map, which obviously produces much more stable results for noisy images. Despite the advantage of preprocessing, we still found that our method consistently outperforms the other state-of-the-art[2,13] methods over a large range of noise strengths.

Keywords: Saliency, non-parametric regression, saliency for noisy images

1. INTRODUCTION

Visual saliency is important both because it directs our attention to what we want to perceive and also because it affects the processing of information. Visual saliency is also useful to allocate limited perceptual resources to objects of interest and decrease our awareness of areas worth ignoring in our visual field. From an engineering point of view, being able to identify small interesting regions or a few objects in a given image is a key to any machine vision system that could process a flood of visual information. In this paper, we propose a computational saliency model that transforms the given input image into a scalar-valued map representing visual saliency of the input image. This map is expected to be useful in applications such as object detection,[14,17,18,24] action detection,[13] and image quality assessment[10,12] and more. Recently, many computational saliency models have been introduced.[2,5,7,13,23] These approaches, however, generally assume the given image is clean. Fortunately, these saliency models implicitly suppress the noise by blurring and downsampling the input image, and therefore tend to be apparently rather insensitive to noise.[11] However, a fundamental and explicit treatment of saliency in noisy images is missing from the literature. Indeed, as we will show, the price for this apparent insensitivity to noise is that the overall performance over a large range of noise strengths is diminished. Therefore, the aim of this paper is two-fold. First, we propose a simple and statistically well-motivated computational saliency model which achieves a high degree of accuracy in predicting where humans look in the given image. Second, with explicit treatment of saliency in *noisy* images, the proposed model is more stable when a noise-corrupted image is given and improves on other state-of-the-art models[2,13] over a large range of noise strengths.

Most saliency models are biologically inspired and based on a bottom-up computational model. Itti et al.[7] introduced a saliency model based on the biologically plausible architecture proposed by Koch and Ullman[8] and measure center-surround contrast using a Difference of Gaussians (DoG) approach. Bruce and Tsotsos[2] measured saliency at a pixel in the image by the self-information of the location with respect to its surrounding

Further author information: (Send correspondence to Chelhwon Kim)
Chelhwon Kim: E-mail: chkim@soe.ucsc.edu, Telephone: 1 831 295 2406
Peyman Milanfar: E-mail: milanfar@soe.ucsc.edu, Telephone: 1 650 332 2311

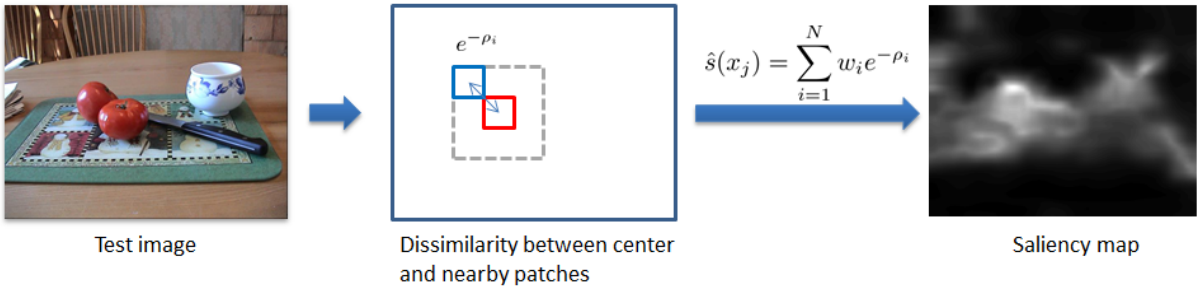

Figure 1. Overview of saliency detection: We observe dissimilarity of a center patch around $\mathbf{x}_j$ relative to its surrounding patches. The proposed saliency model is a weighted average of the observed dissimilarities.

context. Zhang et al.[23]'s saliency model is also based on the self-information and using natural image statistics within a Bayesian framework. Gao et al.[5] maximize the mutual information between distributions of a set of features from center and surround regions. Seo and Milanfar[13] uses the self-resemblance mechanism to compute saliency using a nonparametric kernel density estimation based on the matrix cosine similarity measure.

The proposed saliency model is also based on the bottom-up computation. 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. In order to measure this distinctiveness of region, we observe dissimilarities between a center patch of the region and nearby patches (Fig. 1). Once we have measured these dissimilarities in the region, 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 so as to propose a novel and statistically sound saliency model. We assume that each observed dissimilarity has an underlying true value, but with some noise. 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 a linear combination 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 counter-intuitive 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 $\mathbf{x}_j$ is defined as:

$$\hat{s}(\mathbf{x}_j) = \sum_{i=1}^{N} w_i y_i \tag{1}$$

where y_i and w_i are the observed dissimilarity (to be defined shortly in Section 2) and the weight of i-th patch pair, respectively. We will describe details later in the next section.

It is important to highlight the direct relation of our approach to two earlier approaches of Seo and Milanfar[13] and Goferman et al.[6] We make this comparison explicit here because these methods too involve aggregation of local dissimilarities. While this was not made entirely clear in either[6] or,[13] it is interesting to note that these methods employed arithmetic and harmonic averaging of local dissimilarities, respectively. In,[13] Seo and Milanfar defined the estimate of saliency at pixel $\mathbf{x}_j$ by

$$\hat{s}(\mathbf{x}_j) = \frac{\exp(1/\tau)}{N} \underbrace{\frac{N}{\sum_{i=1}^{N} 1/y_i}}_{\text{harmonic mean}} \tag{2}$$

where $y_i = \exp(\frac{-\rho_i}{\tau})$ and ρ_i is the cosine similarity between visual features extracted from the center patch around the pixel $\mathbf{x}_j$ and its nearby patch. This saliency model is (to within a constant) the harmonic mean of dissimilarities, y_i's.

Goferman et al.[6] proposed a context-aware saliency detection based on four basic principles of human visual

attention and formularized the saliency at pixel $\mathbf{x}_j$ as

$$\hat{s}(\mathbf{x}_j) = 1 - \exp(\quad \underbrace{-\frac{1}{N}\sum_{i=1}^{N} y_i}_{\text{arithmetic mean}} \quad) \tag{3}$$

where $y_i = d(p_j, q_i)$ and $d(\cdot)$ is the dissimilarity measure[6] between a center patch p_j around the pixel $\mathbf{x}_j$ and any other patch q_i observed in the test image. This saliency model is the arithmetic mean of y_i's, which is equivalent to ours except for the exponential and the subtraction from one. The important difference is that while the weights they have are constant, our approach is using more adaptive weights.

Consequently, among those saliency models in which the level of dissimilarities locally (or globally) observed are combined by different aggregation techniques, our proposed method is using a simpler, better justified, and indeed more effective *arithmetic* aggregation based on kernel regression. Moreover, while other saliency models described above use constant weights for their respective modes of aggregation, the proposed one is using varying weights based on non-parametric regression. Furthermore, our saliency model is not only mathematically simple but also stable when a noise-corrupted image is given, as will be discussed in Section 4.

The paper is organized as follows. Section 2 provides further technical details about the proposed saliency model. Section 3 demonstrates the performance of this saliency model in predicting human fixations with other two state-of-the-art models[2,13] . In section 4, we investigate the stability of our method in the presence of noise. Section 5 will conclude the paper.

2. NON-PARAMETRIC REGRESSION FRAMEWORK FOR SALIENCY

In this section, we seek to estimate the saliency at a pixel of interest from observations of dissimilarity between a center patch around the pixel and its nearby patches (See Fig. 1). Let us denote by ρ_i 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 ρ as follows:

$$y_i = e^{-\rho_i}, \quad i = 1, ..., N \tag{4}$$

The similarity function ρ can be measured in a variety of ways,[9,13,15,19] for instance using the matrix cosine similarity between visual features computed in the two patches.[13] For our experiments, we shall use the LARK features as defined in,[21] which have been shown to be robust to the presence of noise and other distortions. From an estimation theory point of view, we assume that each observation y_i is in essence a measurement of the true saliency s_i, but measured with some error. This observation model can be posed as:

$$y_i = s_i + \eta_i, \quad i = 1, ..., N \tag{5}$$

where η_i is some noise. Given these observations, we assume a locally constant model of saliency and estimate the expected saliency at pixel $\mathbf{x}_j$ by solving the weighted least squares problem

$$\hat{s}(\mathbf{x}_j) = \arg\min_{s(\mathbf{x}_j)} \sum_{i=1}^{N} [y_i - s(\mathbf{x}_j)]^2 K(y_i, y_r) \tag{6}$$

where $s(\mathbf{x}_j)$ is an underlying saliency and y_r is a reference observation. Depending on the difference between this reference observation y_r and each observation y_i, the kernel function $K(\cdot)$ gives higher or lower weight to each observation as follows:

$$K(y_i, y_r) = e^{-\frac{(y_i - y_r)^2}{h}} \tag{7}$$

The reference observation can be chosen as the minimum of the N given observations, which corresponds to the most similar patch pair. 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 frivolously 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. As we illustrate later in the experimental results, this approach is quite useful when

the input image is corrupted by noise. Namely, we do not easily allow any region to be declared salient and thus we reduce the likelihood of false alarms. While we define the way of weighting each noisy observation according to the kernel function, we set the weight of the reference observation itself as the maximum of the rest of weights for other observations. 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 as a function of the variance of noise η defined in the observation model Equation (5).

Solving 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,

$$\hat{s}(\mathbf{x}_j) = \frac{\sum_{i=1}^{N} K(y_i, y_r)\ y_i}{\sum_{i=1}^{N} K(y_i, y_r)} = \sum_{i=1}^{N} w_i y_i \tag{8}$$

In Section 3, we first evaluate our saliency model for clean images against two existing saliency models,[2,13] and then investigate the stability of our saliency model for noisy images in Section 4.

3. PERFORMANCE EVALUATION

In this section, we evaluate the proposed saliency model in predicting human eye fixations on Bruce & Tsotsos's dataset*. For quantitative performance analysis, we use area under the receiver operating characteristic curve (AUC) and the linear correlation coefficient (CC). The AUC metric determines how well fixated and non-fixated locations can be discriminated by the saliency map using a simple threshold.[20] If the values of 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 non-fixated locations as fixated) are calculated and the ROC curve is constructed by plotting the true positive rate as a function of the false positive rate across all possible thresholds. The CC metric measures the degree of similarity between the saliency map and a fixation density map produced based on fixation points (See Fig. 2). If they are independent, the correlation coefficient is zero.

Zhang *et al.*[23] 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 and human photographers tend to place objects of interest in the center of photograph.[16,20] Secondly, most of saliency models[2,13] have image border effects due to the invalid filter responses at the borders of images and this also produces an artificial improvement in AUC metric. To avoid these problems, they set the non-fixated locations of a test image as the fixated locations of a randomly chosen image from the test set. We follow the same procedure: For each test image, we first compute a histogram of saliency at the fixated locations of the test image and a histogram of saliency at the fixated locations, but of a randomly chosen image from the test set. Then, we compute all possible true positive and false positive rates by varying the threshold on these two histograms respectively. Finally, we compute the AUC. All AUC's computed for the various images in the database are averaged to derive and overall AUC. Because the test images for the non-fixations setting are randomly chosen, we repeat this procedure 100 times and compute the mean and the standard error of the results. The mean of AUC's of the proposed saliency model and the other state-of-the-arts models are shown in Table 1†. As Table 1 shows, our saliency model is competitive with the other state-of-the-art models in terms of both AUC and CC metrics. Fig. 2 demonstrates a qualitative comparison of the proposed model with the human eye fixations, the fixation density map and the saliency maps produced by other models. Our method is seen to have fewer spurious salient regions than others because we avoid frivolously declaring saliency by giving higher weight to similar patch pairs when we combine the observed dissimilarities.

From the quantitative performance analysis in Table 1, it seems that our saliency model does not significantly outperform Seo and Milarfar[13]'s method, though it is significantly simpler to describe and implement. However, as we will see in the next section, our regression based model is more stable when the input images are corrupted by noise, and thus produces better performance as the noise strength is increased.

*Available at `http://www-sop.inria.fr/members/Neil.Bruce/`

†The evaluation code is available at `https://sites.google.com/site/saliencyevaluation/evaluation-measures`

Table 1. Performance in predicting human fixations

Model	AUC (SE)	CC
Burce & Tsotsos[2]	0.672 (0.0006)	0.257
Seo & Milanfar[13]	0.696 (0.0007)	0.327
Proposed method	0.697(0.0007)	0.338

Figure 2. Examples of saliency map. From left to right in each row: An original image with human fixations shown as yellow crosses; the fixation density map produced based on the fixation points; the saliency map by Bruce & Tsotsos; the saliency map by Seo & Milanfar; and the saliency map by the proposed saliency model

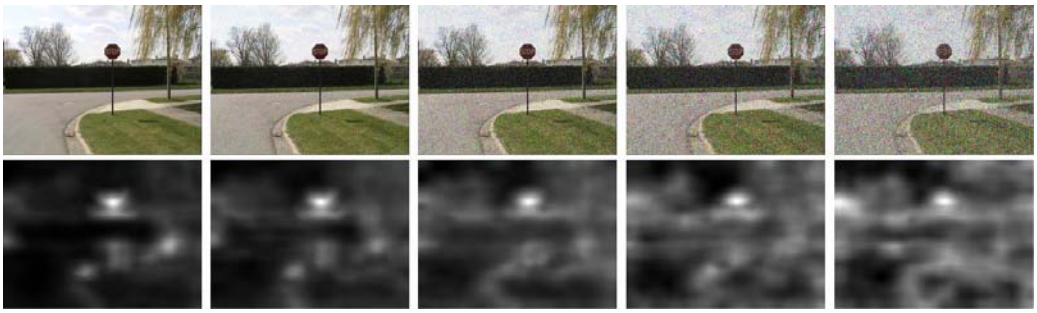

Figure 3. Examples of noisy images and saliency maps of the proposed method. From left to right, a clean image, and noisy images with noise variance $\sigma^2 = \{0.01, 0.05, 0.1, 0.2\}$, respectively

4. 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 database are used and the noise added to the test images is a white Gaussian noise with different variance σ^2 which equals to 0.01, 0.05, 0.1 or 0.2[‡]. Examples of noisy test images and their saliency maps are depicted on Fig. 3. As we may expect, the performance in predicting human fixations decreases in noisy images (See solid curves in Fig. 4).

However, we note that our saliency model produces better performance than Seo and Milanfar's model when the noise strength is increased because of our explicit treatment of noisy images. More specifically, when we compute the weights using the kernel function Equation (7), we set the smoothing parameter h as $1.6\sigma_\eta^2$, where σ_η^2 is the variance of η defined in the observation model, $y_i = s_i + \eta_i$. Because σ_η^2 is not known, we learned it from the clean and noisy images as follows: First, we assume that the variance of noise in the test image, σ^2 is known and add that amount of noise to the training images[§]. Next, according to the observation model, we subtract the dissimilarities observed from the clean training images from the ones observed from the noisy training images, which is $\eta_i = y_i - s_i$. Last, the estimate of σ_η^2 is the sample variance of $(y_i - s_i)$'s that are computed using all the training images. We precomputed σ_η^2 for each different noise variance σ^2. The constant 1.6 is determined empirically and fixed for all experiments.

As we alluded to earlier, all saliency models implicitly suppress the noise by blurring and downsampling the input image. Seo and Milanfar downsampled the input image to 64×64. Bruce and Tsotsos used input image down-sampled by a factor of two. We also downsample the input image to 64×64. However, as illustrated in Fig. 4, the price for this implicit treatment is that the overall performance over a large range of noise strengths is diminished. Although our regression based model gives better performance compared to other models when the noise strength is increased, the accuracy for the noisy case is still much lower than the one for the clean case. As such, we can apply a denoising approach before attempting to calculate the saliency map, which is an obvious alternative but surprisingly has not been tried. We use a state of the art denoising method, the block-matching 3-D filter[3] (BM3D)[¶]. In Fig. 4, a significant improvement is obtained for all models with this preprocessing step, and we note that even in this case, our saliency model still consistently outperforms others in a wide range of noise strengths.

5. CONCLUSION

In this paper, we have proposed a simple and statistically well-motivated saliency model based on non-parametric regression. The proposed method is practically appealing and more effective because of its simple mathematical form which is a convex combination of the dissimilarities observed in the given image. Due to its data-dependent weights and the setting of the smoothing parameter in the kernel function based on the noise variance, the proposed saliency model not only achieves a high degree of accuracy, but also improves on other state-of-the-art models for noisy images. Furthermore, even with the advantage of the denoising preprocessing, we found that our method still consistently outperforms others over a large range of noise strengths.

ACKNOWLEDGMENTS

This work was supported by AFOSR Grant FA9550-07-1-0365 and NSF Grant CCF-1016018

REFERENCES

[1] Achanta R., Hemami S., Estrada F., and Süsstrunk S., "Frequency-tuned Salient Region Detection," IEEE International Conference on Computer Vision and Pattern Recognition, (2009).

[2] Bruce N. D. B. and Tsotsos J. K., "Saliency, attention, and visual search: An information theoretic approach," Journal of Vision, Vol. 9(3), pp. 1-24, (2009).

[‡]The intensity value for each pixel in image ranges from 0 to 1.

[§]We use a different image set in MSRA Salient Object Database: `http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm`.

[¶]The software is available at `http://www.cs.tut.fi/~foi/GCF-BM3D/`.

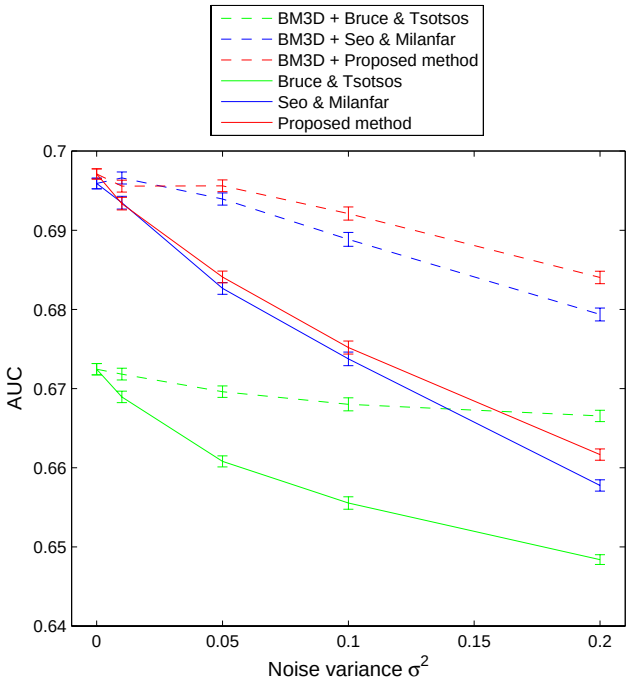

Figure 4. Solid curves: The performance in predicting the human fixations decreases as the amount of noise increases. The proposed method outperforms others over a wide range of noise strengths; Dashed curves: With the denoising preprocessing, the performances of saliency models are significantly improved in terms of the stability to noise. However, we note that the proposed method still consistently outperforms others across different noise strengths.

[3] Dabov K., Foi A., Katkovnic V., and Egiazarian K., "Image denoising by sparse 3D transform-domain collaborative filtering," IEEE Transactions of Image Processing, Vol. 16(8), pp. 2080-2095, (2007).

[4] Elazary L. and Itti L., "Interesting objects are visually salient," Journal of Vision, Vol. 8(3), pp. 1-15, (2008).

[5] Gao D., Mahadevan V., and Vasoncelos N., "On the plausibility of the discriminant center-surround hypothesis for visual saliency," Journal of Vision, Vol. 8(7):13, pp. 1-18, (2008).

[6] Goferman S., Zelnik-Manor L., and Tal. A., "Context-aware saliency detection," IEEE International Conference on Computer Vision and Pattern Recognition, (2010).

[7] Itti L., Koch C., and Niebur E., "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, pp. 1254-1259, (1998).

[8] Koch C. and Ullman S., "Shifts in selective visual attention: Towards the underlying neural circuitry," Human Neurobiology, Vol. 4(4), pp. 219-227, (1985).

[9] Kullback S., [Information Theory and Statistics], Dover:New York, NY.

[10] Ma Q. and Zhang L., "Saliency-based image quality assessment criterion," Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, Vol. 5226, pp. 1124-1133, (2008).

[11] Meur O. Le, "Robustness and repeatability of saliency models subjected to visual degradations," IEEE International Conference on Image Processing, (2011).

[12] Niassi A., LeMeur O., Lecallet P., and Barba D., "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," IEEE International Conference on Image Processing, pp.II-169-II-172, (2007).

[13] Seo H. and Milanfar P., "Static and Space-time Visual Saliency Detection by Self-Resemblance," Journal of Vision, Vol. 9(12), pp. 1-27, (2009).

[14] Seo H. and Milanfar P., "Training-free, Generic Object Detection using Locally Adaptive Regression Kernels," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 32(9), pp. 1688-1704 , (2010).

[15] Swain M. J. and Ballard D. H., "Color indexing," International Journal of Computer Vision, Vol. 7(1), pp. 11-32, (1991).

[16] Parkhurst D. and Niebur E., "Scene content selected by active vision," Spatial Vision, Vol. 16(2), pp. 125-154, (2003).

[17] Rosin P. L., "A simple method for detecting salient regions," Pattern Recognition, Vol. 42(11), pp. 2363-2371, (2009).

[18] Rutishauser U., Walther D., Koch C., and Perona P., "Is bottom-up attention useful for object recognition?," IEEE Conference on Computer Vision and Pattern Recognition, Vol 2, pp. II-37-II-44, (2004).

[19] Rubner Y., Tomasi C., and Guibas L. J., "The earth movers distance as a metric for image retrieval," International Journal of Computer Vision,Vol. 40, pp. 99-121, (2000).

[20] Tatler B. W., Baddeley R. J., and Gilchrist I. D., "Visual correlates of fixation selection: Effects of scale and time," Vision Research, Vol. 45(5), pp. 643-659, (2010).

[21] Takeda H., Farsiu S., and Milanfar P., "Kernel Regression for Image Processing and Reconstruction," IEEE Trans. on Image Processing, Vol. 16(2), pp. 349-366, (2007).

[22] Walther D. and Koch C., "Modeling attention to salient proto-objects," Neural Networks, Vol. 19(9), pp. 1395-1407, (2006).

[23] Zhang L., Tong M. H., and Marks T. K., "SUN: A Bayesian framework for saliency using natural statistics," Journal of Vision, Vol. 8(7), pp. 1-20, (2008).

[24] Zhicheng Li and Itti L., "Saliency and Gist Features for Target Detection in Satellite Images," IEEE Transactions on Image Processing, Vol. 20(7), pp. 2017-2029, (2011).