Visual saliency estimation by nonlinearly integrating features using region covariances

Erkut Erdem

Department of Computer Engineering, Hacettepe University, Ankara, Turkey

Aykut Erdem

Department of Computer Engineering, Hacettepe University, Ankara, Turkey

To detect visually salient elements of complex natural scenes, computational bottom-up saliency models commonly examine several feature channels such as color and orientation in parallel. They compute a separate feature map for each channel and then linearly combine these maps to produce a master saliency map. However, only a few studies have investigated how different feature dimensions contribute to the overall visual saliency. We address this integration issue and propose to use covariance matrices of simple image features (known as *region covariance* descriptors in the computer vision community; Tuzel, Porikli, & Meer, 2006) as meta-features for saliency estimation. As low-dimensional representations of image patches, region covariances capture local image structures better than standard linear filters, but more importantly, they naturally provide nonlinear integration of different features by modeling their correlations. We also show that first-order statistics of features could be easily incorporated to the proposed approach to improve the performance. Our experimental evaluation on several benchmark data sets demonstrate that the proposed approach outperforms the state-of-art models on various tasks including prediction of human eye fixations, salient object detection, and image-retargeting.

Introduction

A natural scene typically contains many objects of various structures at different scales. This complexity poses a great challenge for our visual system since, with its limited capacity, it has to analyze a vast amount of visual information at any given time. To cope with this information overload, the human visual system has developed attentional mechanisms to select the most relevant (salient) parts of a scene that stand out relative to the other parts. What captures our attention depends on factors relevant to bottom-up or top-down selection processes, or a combination of those. While the bottom-up visual attention is mainly driven by intrinsic low-level properties of a scene, the top-down attention involves high-level visual tasks such as searching for a specific object.

Recent years have seen an increase in the number of computational approaches to visual saliency estimation. Starting from the seminal work by Itti, Koch, and Niebur (1998) most of the proposed saliency models consider a bottom-up strategy in which a saliency map is extracted in a purely data-driven manner by considering center-surround differences (e.g., Gao & Vasconcelos, 2007; Harel, Koch, & Perona, 2007; Seo & Milanfar, 2009). There are also some studies that carry out such computations in the frequency domain (Achanta, Hemami, Estrada, & Susstrunk, 2009; Hou & Zhang, 2007) or make use of natural image statistics (Bruce & Tsotsos, 2006; Zhang, Tong, Marks, Shan, & Cottrell, 2008). Another important line of models integrates low-level cues with some task-specific top-down knowledge such as face and object detectors (Cerf, Harel, Einhaeuser, & Koch, 2007; Goferman, Zelnik-Manor, & Tal, 2010; Judd, Ehinger, Durand, & Torralba, 2009), and global scene context (Torralba, Oliva, Castelhano, & Henderson, 2006) to improve their predictions. Lastly, some recent studies pose saliency estimation as a supervised learning problem (Judd et al., 2009; Liu, Jian Sun, & Shum, 2007; Zhao & Koch, 2011, 2012).

These computational models of visual saliency not only provide important insights into the underlying mechanisms of the human visual system but also have been shown to improve the performances of many computer vision applications such as scene classification (Siagian & Itti, 2007), object recognition (Gao, Han, & Vasconcelos, 2009; Rutishauser, Walther, Koch, & Perona, 2004), object tracking (Butko,

Citation: Erdem, E., & Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision, 13*(4):11, 1–20, http://www.journalofvision.org/content/13/4/11, doi:1167/13.4.11.

doi: 10.1167/13.4.11 Received August 17, 2012; published March 18, 2013 ISSN 1534-7362 © 2013 ARVO

Downloaded from jov.arvojournals.org on 07/28/2019

Lingyun, Cottrell, & Movellan, 2008), video compression (Wang, Lu, & Bovik, 2003), and image retargeting (Achanta & Susstrunk, 2009; Avidan & Shamir, 2007; Cheng, Zhang, Mitra, Huang, & Hu, 2011; Goferman et al., 2010; Wang, Tai, Sorkine, & Lee, 2008).

In most of the existing bottom-up models, we observe the following basic structure: (a) extract some basic visual features such as color and orientation, (b) investigate feature channels in parallel and extract a feature map for each dimension, and (c) integrate these maps to produce a master saliency map. Here, the most troublesome step is the last step, which is typically carried out by taking the weighted average (linear summation). In this regard, some recent saliency approaches try to overcome the feature integration problem by finding optimal values for the weights in the linear summation of feature maps in a supervised manner (Judd et al., 2009; Liu et al., 2007; Zhao & Koch, 2011). More recently, Zhao and Koch (2012) proposed combining feature maps in a nonlinear way by using AdaBoost learning method.

However, it is important to note that a limited number of studies have addressed how different feature dimensions contribute to the overall saliency (Callaghan, 1989, 1990; Eckstein, Thomas, Palmer, & Shimozaki, 2000; Rosenholtz, 1999, 2001; Rosenholtz, Nagy, & Bell, 2004). For instance, it has been argued that for certain tasks some visual features may become visually more salient than others. In detecting region boundaries, Callaghan (1989, 1990) showed that the human visual system favors color over shape and form. In a related discussion, Eckstein et al. (2000) suggested that the difficulty in searching for conjunctions (Treisman & Gelade, 1980) can be explained by examining the target and the distractors in a high-dimensional visual feature space and by looking into the feature dimensions along which the distractor and the target differ. Similarly, Rosenholtz (1999, 2001) and Rosenholtz et al. (2004) suggested that the covariance of the distractors in a higher dimensional feature space may provide an explanation for the difficulty of searching a target in motion or in different color.

Proposed approach

In this study, we aimed to perform the last two steps of the aforementioned general structure of bottom-up visual saliency estimation in a single shot. For that purpose, we proposed using covariance matrices of simple image features extracted from local image patches, known as *region covariance*s (Tuzel, Porikli, & Meer, 2006), as meta-features for saliency estimation. These second-order statistical descriptors capture local structure information in an effective manner by encoding pairwise correlations among features, but most notably, they provide a nonlinear solution to the aforementioned feature integration step (step c). We directly computed the saliency of a local image patch by means of the distances between its covariance descriptor and those of the surrounding patches. Nonlinear integration of different features makes our framework especially suitable for natural images containing texture elements or repeating patterns. Here, one reasonable concern is that it might fail in explaining search for conjunctions since our model considers the statistical relationships among different visual features and does not take into account the features in isolation. Another point that may be raised against the covariance-based saliency estimation based on covariances is that it does not take into account the differences in the means which could also indicate saliency. In this regard, we showed that first-order statistics can be easily incorporated to our saliency model to further improve the performance. We demonstrated through extensive experiments that the proposed approach achieves highly competitive results compared to many state-of-the-art models especially in predicting human eye fixations.

It is worth mentioning that incorporating higher-order statistics into saliency estimation has been previously investigated by a number of researchers. For example, Rosenholtz (1999, 2001) suggested computing the saliency of a region or a point as the Mahalanobis distance between its feature representation and the mean of the surround features by taking into account the covariance of the surround features. In a similar fashion, Torralba (2003) and Torralba et al. (2006) gave another bottom-up saliency definition by modeling the distribution of local features with a mixture of Gaussians or a multi-variate power-exponential distribution. Our approach differed from these methods in that we made use of covariance matrices of local features to represent image regions and considered the distance between their covariance matrices in visually comparing them. In this respect, it can be said that the proposed approach shares similarities with some methods proposed for texture segmentation and modelling (e.g., Bangalore & Ma, 1996; Portilla & Simoncelli, 2000; Puzicha, Hofmann, & Buhmann, 1997; Rosenholtz, 2000; Voorhees & Poggio, 1988). However, those works only consider variances of features but not the covariances as in the proposed model.

Previous work

Most of the previously proposed saliency models are grounded in theories of preattentive vision such as the *Feature Integration Theory* by Treisman and Gelade (1980) or the *Guided Search Model* of Wolfe, Cave, and

Downloaded from jov.arvojournals.org on 07/28/2019

Franzel (1989) and Wolfe (1994). They mostly differ in their computational properties.

The earliest computational approach to visual saliency is the biologically motivated model of Itti et al. (1998). It is basically an implementation of the approach of Koch and Ullman (1985) and employs a multi-scale center-surround mechanism that imitates the workings of the retinal receptive fields. Center-surround differences are calculated for a set of linear features, and then the resulting maps are combined together to estimate the master saliency map.

Harel et al. (2007) proposed a *graph-based* approach in which several feature maps are first extracted at multiple spatial scales, as in the model of Itti et al. (1998), and then represented as fully connected graphs. While the vertices of the graphs denote the grid positions, the edges represent the relationships between pairs of vertices, weighted in proportion to their dissimilarity in the corresponding feature space and their spatial distance. The resulting graphs are used to define Markov chains in which their equilibrium distribution is treated as in the activation and saliency maps.

Hou and Zhang (2007) suggested a simple way to perform saliency estimation in the frequency domain in which the saliency of an input image is computed as the inverse Fourier transform of the *spectral residual* of the image, which is defined as the difference between the log spectrum of the image and its smoothed version.

Torralba (2003) and Torralba et al. (2006) proposed a Bayesian *contextual guidance* model for visual search tasks which combines low-level salience and scene context. In estimating the probability of a target at each pixel location, the bottom-up saliency (which does not depend on the target) is modeled as $1/[p(F|G]$ where F denotes the local features and G denotes the global features of the image representing the gist of the scene.

The *Saliency using natural statistics* (SUN) model from Zhang et al. (2008) is another Bayesian framework that combines top-down and bottom-up information to guide visual search tasks. Unlike the approach by Torralba et al. (2006), the bottom-up saliency is based on the self-information of visual features. The authors implemented their approach with two different sets of features, one based on *difference of Gaussians* (DoG) filters and the other based on *independent component analysis* (ICA) features extracted from a training set of natural images.

Bruce and Tsotsos (2006, 2009) approached bottom-up saliency estimation from an information-theoretic perspective and modeled the problem based on the principle of sampling *information maximization* of a scene. Similar to the case in Zhang et al. (2008), their study also employed high-level features derived via ICA, but these features were learned from the input image itself. It is important to note that these models both consider a global definition of saliency in which the salient parts are estimated by considering the global rarity of the local visual features in the entire image.

Goferman et al. (2010) introduced a *context-aware* saliency model that aims to detect the important parts of the image representing the scene. While identifying these salient regions, they investigated a variety of factors including local color and contrast information, frequently occurring global features, and some visual organizational rules and high-level semantic information such as probability maps of face detectors.

Seo and Milanfar (2009) presented a *self-resemblance* measure based on nonparametric kernel density estimation in which local steering kernels (LSKs) are employed as features. LSKs are obtained by examining pixel value differences based on estimated gradients, and the saliency of a pixel is then measured as the likelihood of saliency of a feature matrix given its neighboring feature matrices. Our framework, to a certain extent, resembles the approach of Seo and Milanfar (2009) in the sense that both employ nonlinear features. Both region covariances and LSKs capture local image structures better than responses of standard linear filters such as Gabor, DoG, or ICA filters. Moreover, both approaches carry out feature integration in a nonlinear way. In particular, Seo and Milanfar (2009) used the matrix cosine similarity (MCS) between two LSK features; whereas, we utilize the geodesic distance between covariance descriptors.

Figure 1 shows a highly textured image, taken from Bruce and Tsotsos, 2006, that contains birthday candles of various orientation and color, randomly distributed on a solid background. This example demonstrates the importance of using region covariances in saliency estimation. In the case of natural images, humans perceive a textured region as a whole and thus attend more to the texture discontinuities (changes in the textures) instead of the textured region. While our model accurately captured the white gap in the image as the most salient part, most of the state-of-the-art models produce high saliency scores on the highly textured regions since the linear features that they employ give strong responses in these areas. For this specific example, incorporating mean into our covariance-based model did not improve the prediction, although its result was superior to those of most of the other saliency models.

In the next section, Nonlinear feature integration using region covariances, we present a detailed description of the proposed framework. In the Experimental results section, we perform a comprehensive evaluation on some benchmark data sets. Finally, we conclude the paper with a brief discussion and possible directions for future work.

Downloaded from jov.arvojournals.org on 07/28/2019

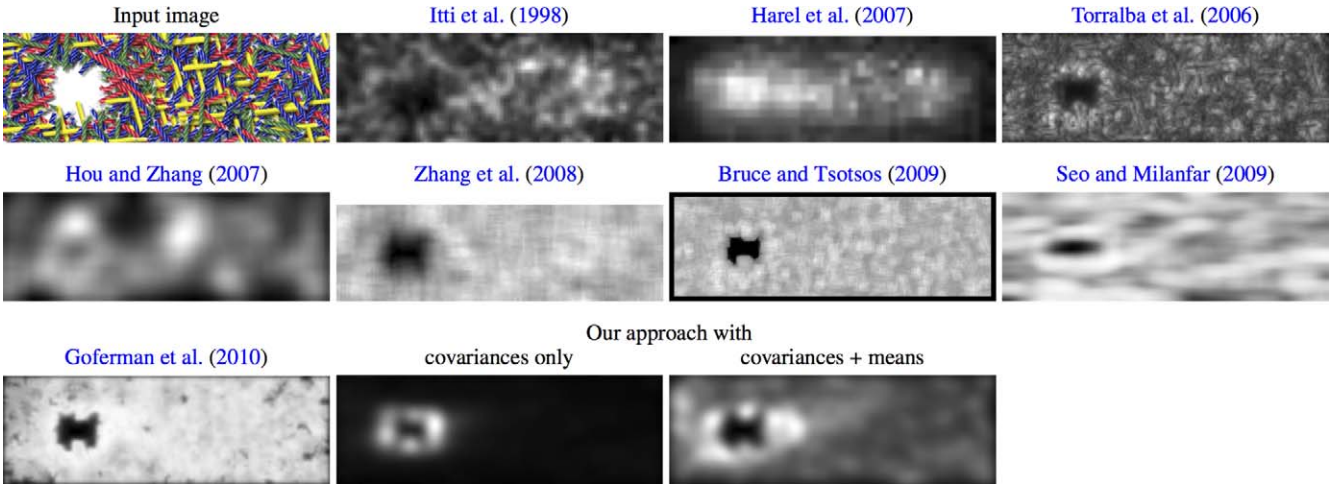
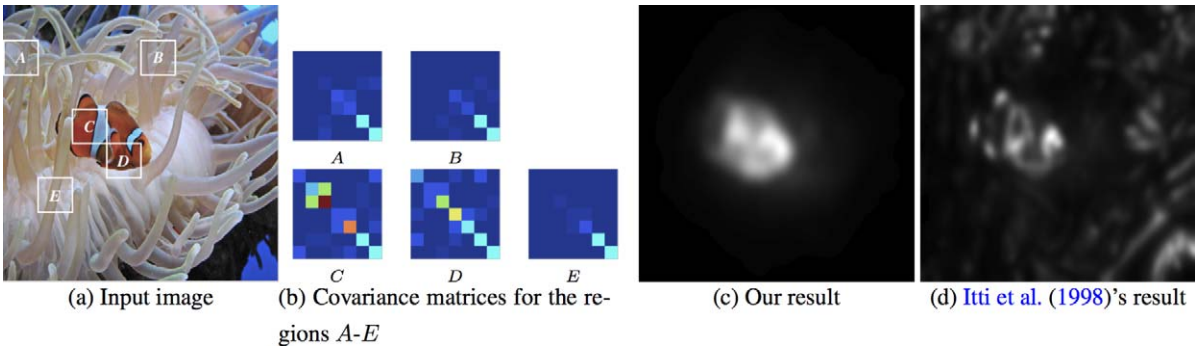


Figure 1. For a highly textured input image, most of the state-of-the-art saliency estimation algorithms respond strongly to the textured regions. In contrast, the proposed region covariance-based model detects the white gap as the salient part and gives a perceptually more meaningful result. This is mainly achieved by nonlinearly combining simple visual features through covariances. For this specific example, incorporating mean does not further improve the prediction.

Nonlinear feature integration using region covariances

Our model employs a local definition of saliency in which the saliency of a pixel is measured by how much it differs from its surroundings. This is carried out on a patch-by-patch basis in which each rectangular image region (local neighborhood of a pixel) is compared against its immediate context described by the nearby regions. We represented each patch by its region covariance descriptor (Tuzel et al., 2006), which naturally provided a nonlinear integration of features using second-order statistics. To illustrate the general idea, consider the image given in Figure 2a. Perceptually, the clownfish swimming around a coral reef stands outs as the most salient object in the cluttered background and immediately grabs the viewer's attention. The proposed model first decomposes the image into non-overlapping regions and estimates their covariances. Here, we employed some basic features, namely color, orientation, and spatial information (see the Implementation details section). For the regions highlighted in Figure 2a, the corresponding covariance matrices are shown in Figure 2b. They clearly demonstrate that the regions with similar characteristics have similar covariances; whereas, the visually dissimilar ones have dissimilar descriptions. As presented in Figure 2c, if our model is used, the fish pops out from the background. Compare this with the result of the seminal model of Itti et al. (1998) in Figure 2d.



(a) Input image (b) Covariance matrices for the regions *A*-*E* (c) Our result (d) Itti et al. (1998)'s result

Figure 2. The proposed saliency model. The input image is first decomposed into non-overlapping regions, and then the saliency of each region is measured by examining its surrounding regions. The salient regions are those that are highly dissimilar to their neighboring regions in terms of their covariance representations based on color, orientation, and spatial features. For the given input image, the covariance distances between the region *C* to the other regions in consideration are calculated as $\rho(C,A) = 6.09$, $\rho(C,B) = 7.33$, $\rho(C,D) = 3.37$, and $\rho(C,E) = 7.27$. In the saliency map computed by the proposed model, the fish pops out from the complex background, *cf.* Itti's saliency map (Itti, Koch, & Niebur, 1998).

Downloaded from jov.arvojournals.org on 07/28/2019

In the Region covariances section, we provide the technical details about the proposed model. First, we review the region covariance descriptor and then give the details of our saliency model. Finally, we provide the implementation details.

Region covariances

Covariance of features was first proposed as a compact region descriptor by Tuzel et al. (2006). Since then, it has been effectively utilized in various high-level computer vision problems such as texture discrimination (Tuzel et al., 2006), object detection (Tuzel et al., 2006; Tuzel, Porikli, & Meer, 2008), and object tracking (Porikli, Tuzel, & Meer, 2006). For the formal definition, let I denote an image, and F be the feature image extracted from I:

$$F(x,y) = \Phi(I,x,y) \tag{1}$$

where Φ denotes the d-dimensional function of features such as intensity, color, orientation, spatial attributes, etc.

Then, a region R inside F can be represented with a $d \times d$ covariance matrix $\mathbf{C}_R$ of the feature points:

$$\mathbf{C}_R = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{f}_i - \boldsymbol{\mu})(\mathbf{f}_i - \boldsymbol{\mu})^T \tag{2}$$

with $\{\mathbf{f}_i\}_{i=1\ldots n}$ denoting the d-dimensional feature points inside R, and μ being the mean of these points.

Tuzel et al. (2006) also proposed a fast way of computing covariance matrices of rectangular regions by using the first and the second-order integral image representations (Viola & Jones, 2001) with $O(d^2)$ computational complexity.

Note that covariance matrices do not lie on Euclidean space. Hence, to compute the distance between two covariances $\mathbf{C}_1$ and $\mathbf{C}_2$, Tuzel et al. (2006) suggested using the metric proposed by Föerstner and Moonen (1999):

$$\rho(\mathbf{C}_1,\mathbf{C}_2) = \sqrt{\sum_{i=1}^{n} ln^2\lambda_i(\mathbf{C}_1,\mathbf{C}_2)} \tag{3}$$

where $\{\lambda_i(\mathbf{C}_1,\mathbf{C}_2)\}_{i=1\ldots n}$ and $\{\mathbf{x}_i\}$ are the generalized eigenvalues and the generalized eigenvectors of $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively, satisfying

$$\lambda_i\mathbf{C}_1\mathbf{x}_i - \mathbf{C}_2\mathbf{x}_i = 0, \qquad i = 1\ldots d. \tag{4}$$

A covariance matrix provides a natural way of combining different visual features with its diagonal elements representing the feature variances and its nondiagonal elements representing the correlations among the features. Unlike the common practice in the existing computational saliency models that assume the responses of linear filters are independent of one another and combined linearly, incorporating second-order image statistics within a single descriptor encodes the local structure exceedingly well and provides robustness and high discriminative power. It is worth noting that Karklin and Lewicki (2009) speculated that nonlinearity in *complex cells* in the primary visual cortex (V1) can be explained by the higher-level visual neurons, which encode statistical variations describing local image regions through covariance matrices. Although we based our decision solely on computational grounds, the former argument provides an insight into the biological plausibility of using region covariances in saliency estimation.

Incorporating first-order statistics

In distinguishing between two different distributions of features, first-order statistics could also play an important role. To remedy this issue, Hong, Chang, Shan, Chen, and Gao (2009) proposed employing the notion of so-called Sigma Points (Julier & Uhlmann, 1996) in which covariance matrices are transformed directly on Euclidean vector space using the Cholesky decomposition. The idea is based on the property that every symmetric, positive definite matrices (covariance matrices) has a unique factorization whose elements can be used to construct a small set of points in Euclidean space. Once this is achieved, it becomes straightforward to incorporate the mean vector of the features, thus resulting in an enriched representation which encodes both first and second-order statistics.

Let $\mathbf{C}$ be a $d \times d$ covariance matrix, the corresponding set of Sigma Points $\mathrm{S} = \{\mathbf{s}_i\}$ can be computed as:

$$\mathbf{s}_i = \begin{cases} \alpha\sqrt{d}\mathbf{L}_i & \text{if } 1 \le i \le d \\ -\alpha\sqrt{d}\mathbf{L}_i & \text{if } d+1 \le i \le 2d \end{cases} \tag{5}$$

where $\mathbf{L}_i$ is the ith column of the lower triangular matrix $\mathbf{L}$ obtained with the Cholesky decomposition $\mathbf{C} = \mathbf{L}\mathbf{L}^T$. Using the set S given in Equation 5, a feature vector can be obtained by simply concatenating its elements. Moreover, first-order statistics can be easily incorporated to this representation scheme by adding the mean vector of the features $\boldsymbol{\mu}$. We denote this enriched feature vector as Ψ ($\mathbf{C}$):

$$\Psi(\mathbf{C}) = (\mu, \mathbf{s}_1, \ldots \mathbf{s}_d, , \mathbf{s}_{d+1}, \ldots, \mathbf{s}_{2d})^T \tag{6}$$

For all the experiments in this paper, we take $\alpha = \sqrt{2}$.

Local saliency estimation

Given an input image I, our model reshapes the image to a square form and then decomposes it into non-overlapping regions of square blocks $\{R_i\}$, which

Downloaded from jov.arvojournals.org on 07/28/2019

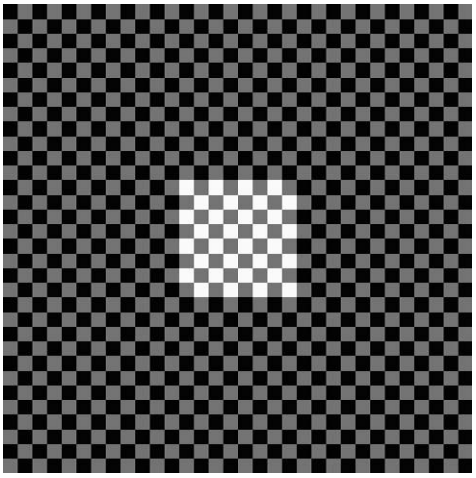


Figure 3. A synthetic image that highlights a case in which considering only covariance features could not provide an accurate saliency prediction whereas looking difference in the means could.

are of size $k \times k$ pixels. The saliency of a block is estimated by comparing it with its nearby context. If it locally displays distinct characteristics, it is regarded as salient. The region properties depend on the pixels within the region, and thus it can be argued that the region size k determines the scale at which the saliency prediction is performed. As compared to similar models that use local patch-based strategies (Borji & Itti, 2012; Duan, Wu, Miao, Qing, & Fu, 2011; Goferman et al., 2010; Seo & Milanfar, 2009), the main novelty of our model comes from using covariance descriptors of the regions to represent their visual characteristics.

In this study, we conducted experiments on two different versions of our model which respectively employed (a) covariance features only and (b) combined covariance and mean features.

Model 1: Saliency using covariance features

Let R_i denote the region under consideration whose immediate context is defined by the regions $\{R_j\}$ within a radius of r. The saliency of R_i is defined as the weighted average of the dissimilarities between R_i to the m most similar regions around it. More formally, the saliency of region R_i is given by:

$$S(R_i) = \frac{1}{m}\sum_{j=1}^{m} d(R_i,R_j) \tag{7}$$

where the m most similar regions to R_i is found according to the dissimilarity measure $d(R_i, R_j)$ defined as:

$$d(R_i,R_j) = \frac{\rho(\mathbf{C}_i,\mathbf{C}_j)}{1 + ||\mathbf{x}_i - \mathbf{x}_j||} \tag{8}$$

with $\mathbf{C}_i$ and $\mathbf{C}_j$ denoting the covariance matrices, and $\mathbf{x}_i$ and $\mathbf{x}_j$ being the image coordinates of the center of the regions R_i and R_j, respectively. In determining the distinctiveness of a region, weighting covariance distances by inverse spatial distance decreases the influence of visually similar nearby regions and somehow introduces a grouping-like effect (see Figure 2).

Note that the region size k specifies the resolution of the saliency map. Hence, in order to get a map at the resolution of the original image I, we resized the estimated saliency maps back to the original size. We refer to the interpolated map as $\hat{S}^k$ denoting the saliency at scale k.

Model 2: Saliency using covariance and mean features

In our first model, we employed covariance features to compute the saliency map of an image. Although covariance matrices can effectively encode local structure information by using the second-order statistical relations among features, first-order statistics (mean) can be also valuable in capturing saliency of an image region with respect to its surroundings. The importance of looking at the difference in the means is apparent in Figure 3. It depicts a checkerboard board image that contains a rectangular region at the center whose contrast is lower than the surrounding region and so draws our attention. This rectangular region receives a low saliency value from our first model because the covariances are the same for the center and the surrounding regions. In contrast, since the means are different, an analysis based on first-order statistics would make this region pop out from its surroundings.

To eliminate the shortcoming of the proposed Model 1 already mentioned, we incorporated the mean information into our covariance-based model and came up with a second model in which the saliency of region R_i is given by

$$S(R_i) = \frac{1}{m}\sum_{j=1}^{m} d'(R_i,R_j) \tag{9}$$

where the m most similar regions to R_i is found according to the dissimilarity measure $d'(R_i, R_j)$, which is defined as:

$$d'(R_i,R_j) = \frac{||\Psi(\mathbf{C}_i) - \Psi(\mathbf{C}_j)||}{1 + ||\mathbf{x}_i - \mathbf{x}_j||} \tag{10}$$

with $\Psi(\mathbf{C}_i)$ and $\Psi(\mathbf{C}_j)$ denoting the feature vectors with the incorporated first-order statistics (Equation 6). Again, the estimated saliency maps at scale k could be interpolated to obtain a map $\hat{S}^k$, which is of the same size as the input image.

Downloaded from jov.arvojournals.org on 07/28/2019

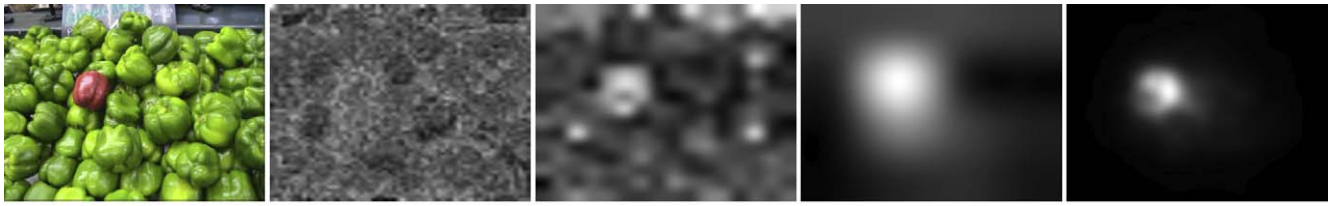


Figure 4. (a) Input image. (b–d) Predicted saliency maps obtained at different scales (from the finest to the coarsest). (e) Final saliency map according to the spatial coincidence assumption described in the text.

Incorporating center bias

The experiments on human eye fixations demonstrate that there is a tendency in humans to look towards the image center, which is called the *center bias*. This bias is mainly explained by several factors including (a) the *photographer bias* (tendency of photographers to place objects of interest in the center of photographs), (b) the *viewing strategy* (tendency of participants to focus on the center to obtain more information), or (c) the *motor bias* (center being the optimal location to initiate a visual search) (Judd et al., 2009; Tatler, 2007; Tseng, Carmi, Cameron, Munoz, & Itti, 2009; Zhang et al., 2008). However, only a limited number of studies additionally consider the center bias in their models (Harel et al., 2007; Judd et al., 2009; Zhao & Koch, 2011). It has been shown that adding the center bias to the saliency models improves the quality of their predictions. Thus, we included a center bias into our second model by defining the saliency of region R_i as follows:

$$S'(R_i) = \left(1 - \frac{||\mathbf{x}_i - \mathbf{x}_c||}{Z}\right) \cdot S(R_i) \tag{11}$$

where $\mathbf{x}_c$ is the coordinates of the image center and Z is a normalization factor equal to $max_{i' \in I}||\mathbf{x}_{i'} - \mathbf{x}_c||$. This additional weight reflects the proximity to the region R_i to the image center and thus signifies the *center bias*.

Scale-space extension

The objects that can be treated as salient in an image can and do appear over a wide range of scales. This suggests that saliency detection should be carried out simultaneously at all possible scales. For that purpose, most multiscale saliency models extract multiple saliency maps, each at a different scale, and then employ a fusion strategy to combine these maps to come up with one final saliency map. The single-scale saliency models described in the previous section can be easily extended to operate on multiple scales by following a similar idea.

Let $K = \{k\}$ denote the set of region sizes representing the scales at which the saliency predictions is carried out. The master saliency map is given by the product of individual saliency maps extracted at different scales, convolved with a Gaussian, as follows:

$$S(x) = G_\sigma(x) * \prod_{k \in K} \hat{S}^k(x) \tag{12}$$

where $\hat{S}_k(x)$ denotes the saliency score of pixel x at scale k, and σ is the standard deviation of the Gaussian filter.

The above definition considers a *spatial coincidence assumption* that an image part should treated as salient if it is salient at all scales. For a sample image, Figure 4 presents saliency maps extracted at three different scales. As can be seen, as we moved to coarser scales, the model tended to capture the location of the visually most prominent region in the image. Figure 4e shows the combined saliency map obtained with the suggested multiscale approach using covariance features. In the master map, the red bell pepper in the image stands out among the surrounding green peppers.

Implementation details

In our implementation, we used very simple visual features, namely color, orientation, and spatial information. Based on these features, an image pixel is represented with a seven-dimensional feature vector:

$$F(x,y) = \left[L(x,y)\ a(x,y)\ b(x,y)\ \left|\frac{\partial I(x,y)}{\partial x}\right|\ \left|\frac{\partial I(x,y)}{\partial y}\right|\ x\ y\right]^T \tag{13}$$

where L, a, and b denote the color of the pixel in $L^*a^*b^*$ color space, $|\partial I/\partial x|$, $|\partial I/\partial y|$, are the edge orientation information, and (x,y) denotes the pixel location. Hence, the covariance descriptor of a region is computed as a 7×7 matrix.

In our model, there are three parameters related to the notion of scale: (a) the set of region sizes K, (b) the neighborhood radius r, and (c) the smoothing param-

Downloaded from jov.arvojournals.org on 07/28/2019

eter σ. The number of most similar neighbors m is another parameter that needs be decided. In the experiments, we first rescaled the input image of $w \times h$ pixels to 512×512 pixels, and fixed the parameter set as $K = \{8, 16, 32, 64, 128\}$, $r = 3$, $\sigma = 0.02*w$, and $m = 1/10$ of the number of the surrounding regions defined by r.

The saliency computation for a single image took a few seconds with a Matlab implementation (MATLAB, Mathworks, Natick, MA) on a Apple Macbook with a 2.53 GHz Intel Core2 Duo processor and 4 GB RAM. It should be noted that the run-time performance of the model can be improved by including some MEX C++ subroutines and/or parallelizing the code.

Experimental results

In this section, we demonstrate the effectiveness of the proposed models with a series of experiments involving (a) prediction of human eye fixations, (b) salient object detection, (c) image retargeting (or content-aware image resizing) (Rubinstein, Gutierrez, Sorkine, & Shamir, 2010), and (d) some psychophysical patterns. Results of these experiments are available in high resolution at http://web.cs.hacettepe.edu.tr/~erkut/projects/CovSal. We compared the proposed approaches both qualitatively and quantitatively with the state-of-the-art saliency detection methods, including the model from Itti et al. (1998) (Itti), Graph-based visual saliency (GBVS) (Harel et al., 2007), the model from Torralba et al. (2006) (Torralba) (excluding the task prior), Spectral residual (Hou & Zhang, 2007), SUN (Zhang et al., 2008), Attention based on information maximization (AIM) (Bruce & Tsotsos, 2009), Saliency detection by self-resemblance (Seo & Milanfar, 2009), and Context aware based saliency detection (Goferman et al., 2010). The source codes of these models were, respectively, downloaded from http://www.klab.caltech.edu/~harel/share/gbvs.php, http://people.csail.mit.edu/tjudd/SaliencyBenchmark/Code/torralbaSaliency.m, http://www.klab.caltech.edu/~xhou/projects/spectralResidual/spectralresidual.html, http://cseweb.ucsd.edu/~l6zhang/code/imagesaliency.zip, http://www-sop.inria.fr/members/Neil.Bruce/, http://users.soe.ucsc.edu/~milanfar/research/rokaf/.html/SaliencyDetection.html, and http://webee.technion.ac.il/labs/cgm/Computer-Graphics-Multimedia/Software/Saliency/Saliency.html, and the results of these models were obtained by the default parameters provided by the authors, which matched with the values reported in the related papers. The only exceptions were the parameters for the GBVS and Itti models. For the GBVS model, the implementation we used employs DKL color space (instead of the "color double-opponent" model used in the original paper) because it has been reported to provide better results. Furthermore, for the Itti model, the center scales $c = \{2,3\}$ and the center-surround scale differences $\delta \in \{2, 3\}$ were employed instead of $c = \{2,3,4\}$ and $\delta \in \{2, 3\}$.

Predicting human eye fixations

The most common way of evaluating the performances of bottom-up saliency models is to measure how well the areas identified as attractive in computed saliency maps coincide with the actual human eye fixations. We tested the proposed method on this task using three publicly available data sets (Bruce & Tsotsos, 2006; Judd et al., 2009; Judd, Durand, & Torralba, 2012).

Data sets

The first data set was from Bruce and Tsotsos (2006), referred to as the Toronto data set, contains 120 natural color images of size 681×511, each depicting an outdoor or an indoor urban scene. The eye movement data were collected from 20 subjects who free-viewed each image for 4 s. The participants were given no particular instructions except to observe the images.

The second data set was introduced by Judd et al. (2009), referred to as the MIT1003 data set, and it was the largest of all, with a total of 1003 natural color images (779 landscape images and 228 portrait images), which were randomly crawled from Flickr creative commons and LabelMe. The images in this data set are mostly of 1024×768 pixels (the longest dimension was 1024 pixels and the other one ranged from 405 to 1024 pixels). The data set contains eye fixation data from 15 viewers who performed a 3-s-long free-viewing task on each image.

The last data set was from Judd et al. (2012), referred to as the MIT300 data set, and it had eye fixation data collected from 39 subjects for a total of 300 natural images (223 landscape images and 77 portrait images). The dimensions of the images were similar to those in the MIT1003 data set, i.e., while the longest dimension of each image was 1024 pixels and the other dimension varied from 457 to 1024 pixels, with mostly 768 pixels. Similarly, the participants also free-viewed each image for 3 s.

Evaluation scores

We used several complementary metrics for a comprehensive quantitative analysis. These are (a) the area under the receiver operator characteristics (ROC) (Green & Swets, 1966) curve (AUC), (b) the Normalized Scanpath Saliency (NSS) (Parkhurst, Law, & Niebur, 2002; Peters, Iyer, Itti, & Koch, 2005), (c) the

Downloaded from jov.arvojournals.org on 07/28/2019

Earth Movers Distance (EMD) (Pele & Werman, 2009; Rubner, Tomasi, & Guibas, 2000), and (d) the similarity score suggested by Judd et al. (2012). In the next subsection, we report the mean values of these metrics averaged over all the images in the data sets.

The ROC score is the most commonly used metric in the literature. In a ROC analysis, the saliency map is thresholded and treated as a binary classifier on every pixel in a given image such that those pixels with saliency values greater than the applied threshold level is classified as fixated; whereas, the rest are considered as nonfixated (Tatler, Baddeley, & Gilchrist, 2005). The fixation data from the subjects are used as ground truth. Then, the consistency between a saliency model and the set of human fixations is given by AUC, obtained by varying the threshold level. An AUC of 1 indicates a perfect prediction, while the chance performance is around an area of 0.5.

The NSS is defined as the average value of the responses at the human fixation points in the predicted saliency map that has been normalized to have zero mean and unit standard deviation. While an NSS ≤ 0 indicates chance level, an NSS ≥ 1 indicates that the responses at fixated points are significantly higher than those at the nonfixated points in the saliency map.

The EMD is a measure of dissimilarity between two probability distributions and is defined as the minimum cost required to transform one distribution into the other. An EMD of zero indicates that two distributions are the same. A larger EMD score suggests that two distributions are significantly different. In our analysis, we used a fast implementation of the EMD provided by Pele and Werman (2009).

The similarity score suggested by Judd et al. (2012) is a measure of similarity between two saliency maps, which is defined as $S = \sum_{i,j} min(P_{i,j}, Q_{i,j})$ with (i, j) denoting the pixel location and P and Q being the saliency maps that have been normalized to have $\sum_{i,j} P_{i,j} = \sum_{i,j} Q_{i,j} = 1$. While a perfect match has a similarity value of 1, comparing two completely different saliency maps results in a similarity score of zero.

A good saliency model should have an AUC value close to 1, a large NSS score, a low EMD value, and a similarity value close to 1. An analysis based on AUC or NSS depends solely on the exact locations of fixation, while the EMD and the similarity score in Judd et al. (2012) compared maps in a more global manner. In this respect, it can be argued that these metrics reveal different characteristics of the saliency models. Despite the widespread use of AUC score as a performance measure for visual saliency, it suffers from the drawback that it only depends on the ordering of the fixations (Zhao & Koch, 2011). That is, as long as the hit rates are high, the AUC is always high regardless of the false alarm rate. Additionally, it does not consider the spatial deviation of the computed saliency map from the actual fixation density map.

Another important point in performance analysis is taking into consideration the center bias. As pointed out by Zhang et al. (2008), a saliency map formed alone by a Gaussian blob centered in the middle of the image also yields very good results in a ROC analysis. Instead of trying to remove the center bias (such as using the unshuffled version of AUC metric suggested by Zhang et al., 2008), we decided to include a center bias into all the models tested in this study. In the following experiments, we follow (Judd et al., 2012) and linearly combine the saliency map of each model with a center map as follows:

$$\text{newSMap} = w * \text{centerSMap} + (1 - w)\text{SMap} \qquad (14)$$

where $w \in [0, 1]$ is the weight of the center map.

Performance

Figures 5 and 6 present results of the proposed approach and the state-of-the-art saliency models on some sample images from the Toronto and the MIT1003 data sets, respectively. In these images, human participants tended to primarily fixate on only one salient area containing a single object surrounded by a highly textured background. Our saliency models gave perceptually more accurate results as compared with other models. This was primarily achieved by performing nonlinear integration of visual features via region covariances, which caused nonsalient regions (textured backgrounds) to be suppressed very effectively. Most of the saliency models, which employ linear integration of features, respond to the true salient regions as well as these textured regions and have high saliency scores at these regions of repeating distractors.

We provide quantitative analysis of the saliency models on the Toronto, MIT1003, and MIT300 data sets in Tables 1 through 3. For each data set, we also show the results of two baseline models referred to as Chance and Center, which stand for the random and the centered Gaussian models, respectively. Tables 1 and 2 present evaluation scores of the saliency models on the Toronto and MIT1003 data sets, respectively (the EMD metric is left out for the MIT1003 data set due to time constraints). The proposed approach, which uses covariances and means with implicit center bias, outperformed the other saliency models in terms of all evaluation metrics. For a fair comparison, we also incorporated an explicit bias towards the center by taking a linear combination of the predicted the saliency map of each model with the center model (denoted by "with CB" in the tables). As can be seen, including a center bias to the models boosted the performances of all saliency models (results were

Downloaded from jov.arvojournals.org on 07/28/2019

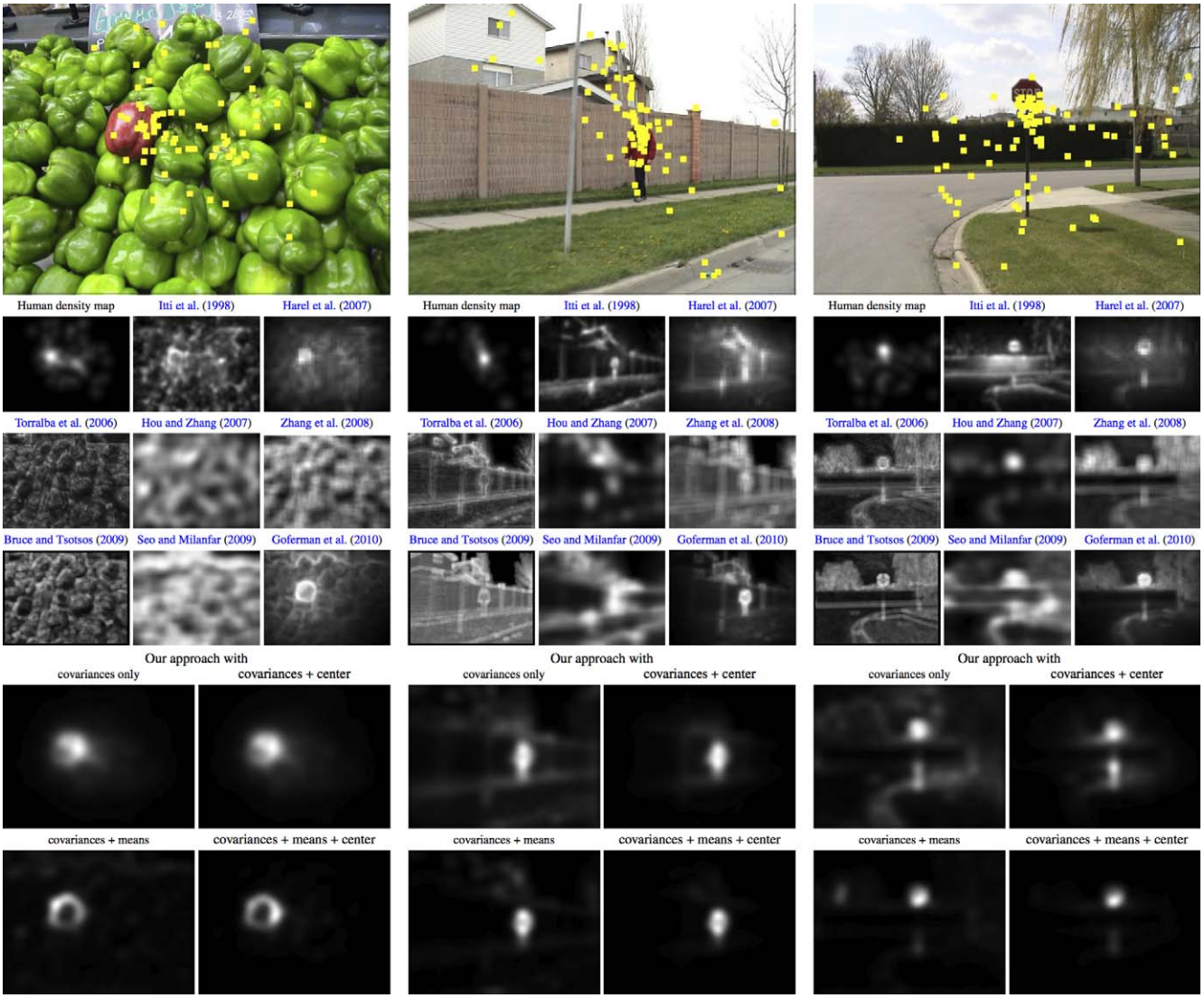


Figure 5. Comparison to the state-of-the-art methods. The top row shows three sample images from the Toronto data set with the superimposed eye fixations from all subjects (drawn with yellow dots). Our saliency model is much less sensitive to background texture as opposed to other models, and correctly predicts the fixations.

obtained with the center weights being optimized separately for each model on each data set). The performances of our models with implicit center bias and the GBVS model did not change much because they inherently had a center bias. Table 3 illustrates the model performances on the MIT300 data set. Similarly, the proposed approach (both versions) gave very competitive results on this data set compared with the state-of-the-art models. As another baseline, we also report the scores of the SVM-based model from Judd et al. (2009), the best performing model in Judd et al. (2012), which learns optimal feature weights through a training process. It should be noted that our results were quite close even if we used very simple features, without any learning.

Detecting salient objects

Salient object detection refers to the task of identifying foreground objects that attract more attention in a given image. This definition of saliency is directly related to figure-ground grouping, and there are some computer vision studies that tackle this problem with a segmentation-type binary labeling formulation (e.g., Cheng et al., 2011; Rahtu, Kannala, Salo, & Heikkilä, 2010). Although the saliency models analyzed in this study are not designed to capture exact (salient) object boundaries, such binary maps can be obtained by thresholding the predicted saliency maps.

Downloaded from jov.arvojournals.org on 07/28/2019

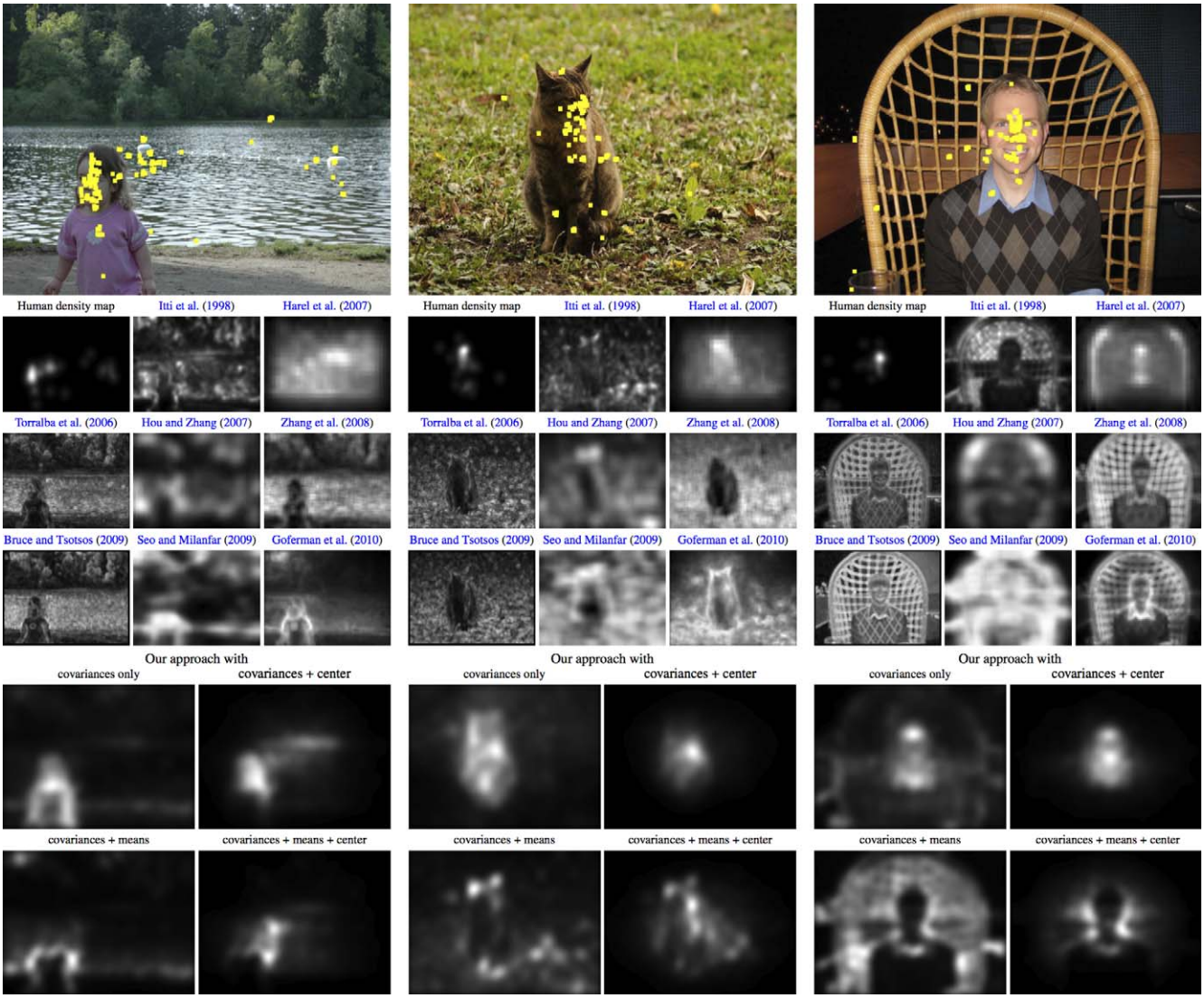


Figure 6. Comparison to the state-of-the-art methods. The top row shows three sample images from the MIT1003 data set with the superimposed eye fixations from all subjects (drawn with yellow dots). Our saliency models again provide results better than most of the state-of-the-art saliency models. The predictions of our approach with covariance features are specifically very close to the ground truth.

Data set

We used recently published ImgSal data set (Li, Levine, An, Xu, & He, 2012), which contains a total of 235 natural color images of size 480 × 640 pixels collected from either Google or from other data sets available on the web. The images in this data set were divided into six categories based on difficulty levels for saliency detection. These categories involve images with *large salient regions* (50 images), *intermediate salient regions* (80 images), *small salient regions* (60 images), *cluttered backgrounds* (15 images), *repeating distractors* (15 images), and *large and small salient regions* (15 images).

The ImgSal data set contains eye fixation records as well as binary maps of the salient objects. Here we only concentrated on region ground truth obtained from 19 subjects who were asked to sit in front of a computer screen and to label the most salient objects in the images presented to them. The value of a pixel in a ground truth map was set to 1 if the majority of the subjects agreed that it belonged to a salient region, otherwise it was set to 0.

Evaluation scores

In Li et al. (2012), the AUC score and the maximal value of the Dice Similarity Coefficient (DSC) curve were used in quantitative evaluation. The predicted saliency maps were thresholded, and the thresholded binary maps were compared against the binary ground truth images provided in the data set using these scores.

Downloaded from jov.arvojournals.org on 07/28/2019

	AUC		NSS		EMD		Similarity	
	Without CB	With CB	Without CB	With CB	Without CB	With CB	Without CB	With CB
Itti et al. (1998)	0.771	0.825	1.137	1.264	2.906	2.002	0.397	0.521
Harel et al. (2007)	0.829	0.835	1.533	1.533	2.014	1.886	0.519	0.556
Torralba et al. (2006)	0.710	0.832	0.805	1.185	3.467	1.868	0.330	0.528
Hou & Zhang (2007)	0.736	0.835	0.964	1.271	3.791	1.959	0.360	0.550
Zhang et al. (2008)	0.718	0.832	0.884	1.194	3.954	1.968	0.347	0.541
Bruce & Tsotsos (2009)	0.728	0.835	0.896	1.165	3.127	1.809	0.351	0.535
Seo & Milanfar (2009)	0.766	0.845	1.100	1.320	3.222	1.759	0.415	0.579
Goferman et al. (2010)	0.784	0.841	1.272	1.370	3.520	1.819	0.431	0.574
Our approach with								
Covariances only	0.767	0.834	1.184	1.342	3.142	1.931	0.408	0.546
Covariances + means	0.765	0.834	1.198	1.396	3.398	1.896	0.402	0.548
Covariances + center	0.840	0.840	1.753	1.753	1.901	1.901	0.561	0.561
Covariances + means + center	**0.851**	**0.851**	**1.891**	**1.898**	**1.728**	**1.728**	**0.581**	**0.581**
Center	–	0.803	–	0.969	–	2.401	–	0.478
Chance	0.505	0.803	−0.001	0.969	5.159	2.339	0.187	0.479

Table 1. Performance comparisons of the saliency models on the Toronto data set. Chance and Center are the baselines, which respectively stand for the random and the centered Gaussian models. CB denotes center bias. The best performing model is shown in bold type.

The DSC is a measure of set agreement defined by $DSC = 2TP/[(TP + FP) + (TP + FN)]$ where TP is the true positive, FP is the false positive, and FN is the false negative counts. A DSC value of 1 indicates a perfect agreement whereas a DSC value of 0 means no overlap, so a good salient object model should give a DSC value close to 1.

Performance

Detecting salient objects on the ImgSal data set poses some great challenges such as variation in scale, cluttered backgrounds, repeating distractors, etc. The images contain one or more objects that are distinguishable from the background by their visual characteristics but with different difficulty levels. In Figure 7, we present some qualitative examples. The illustrated object maps were obtained by setting the threshold as the average intensity of the saliency map plus one standard deviation. Our saliency model detected the salient objects accurately under these difficult scenarios.

We provide quantitative analysis of our model and the state-of-the-art saliency models on the ImgSal data set in Table 4. The proposed models outperformed the other saliency models in three out of six categories, and it was the second best or third best model in other

	AUC		NSS		Similarity	
	Without CB	With CB	Without CB	With CB	Without CB	With CB
Itti et al. (1998)	0.741	0.827	0.921	1.170	0.273	0.402
Harel et al. (2007)	0.791	0.829	1.150	1.182	0.319	0.415
Torralba et al. (2006)	0.700	0.832	0.771	1.156	0.244	0.412
Hou & Zhang (2007)	0.713	0.833	0.855	1.200	0.264	0.421
Zhang et al. (2008)	0.703	0.834	0.829	1.177	0.261	0.418
Bruce & Tsotsos (2009)	0.709	0.835	0.813	1.148	0.254	0.415
Seo & Milanfar (2009)	0.712	0.836	0.826	1.171	0.263	0.424
Goferman et al. (2010)	0.758	0.840	1.053	1.241	0.297	0.431
Our approach with						
Covariances only	0.715	0.826	0.862	1.169	0.261	0.410
Covariances + means	0.740	0.832	0.940	1.240	0.287	0.417
Covariances + center	0.833	0.833	1.468	1.486	0.417	0.418
Covariances + means + center	**0.843**	**0.843**	**1.488**	**1.543**	**0.428**	**0.432**
Center	–	0.810	–	1.004	–	0.379
Chance	0.500	0.810	−0.000	1.004	0.131	0.383

Table 2. Performance comparisons of the saliency models on the MIT1003 data set. The best performing model is shown in bold type.

Downloaded from jov.arvojournals.org on 07/28/2019

	AUC		EMD		Similarity	
	Without CB	With CB	Without CB	With CB	Without CB	With CB
Itti et al. (1998)	0.750	0.806	4.560	3.394	0.405	0.493
Harel et al. (2007)	0.801	0.813	3.574	3.315	0.472	0.501
Torralba et al. (2006)	0.684	0.806	4.715	**3.036**	0.343	0.488
Hou & Zhang (2007)	0.682	0.804	5.368	3.200	0.319	0.487
Zhang et al. (2008)	0.672	0.799	5.088	3.296	0.340	0.473
Bruce & Tsotsos (2009)	0.751	**0.820**	4.236	3.085	0.390	0.507
Goferman et al. (2010)	0.742	0.815	4.900	3.219	0.390	**0.509**
Our approach with						
Covariances + center	0.800	0.800	3.422	3.422	0.487	0.487
Covariances + means + center	**0.806**	0.811	**3.109**	3.109	**0.502**	0.503
Center	–	0.783	–	3.719	–	0.451
Chance	0.503	0.783	6.352	3.506	0.327	0.482
Judd et al. (2009)	0.811	0.813	3.130	3.130	0.506	0.511

Table 3. Performance comparisons of the saliency models on the MIT300 data set. The best performing model is shown in bold type.

categories. Our models achieved the best performance especially when the images contained small salient regions, cluttered backgrounds, and repeating distractors (Categories 3–5) in which nonlinear integration of visual features were required in order to respond to discontinuity in textures. We suspect that the reason why our models performed poorly on the images involving both large and small salient objects (Category 6) was due to the spatial coincidence assumption that was considered in our multiscale saliency definition.

Image retargeting by seam-carving

Image retargeting or content aware image resizing has emerged as an interesting computer vision problem, which deals with automatically resizing an image to arbitrary aspect ratios while trying to preserve important content and internal structure and to prevent visual artifacts (Rubinstein et al., 2010). To achieve these objectives, most retargeting methods assume that an importance map is available that highlights the most prominent objects or the structures in the image so that unimportant regions can be discarded during the resizing process. In this regard, image retargeting has proved to be a good application area for saliency estimation (Achanta & Susstrunk, 2009; Cheng et al., 2011; Goferman et al., 2010; Wang et al., 2008). However, the literature lacks a quantitative analysis of the performance of saliency models on retargeting tasks. To our knowledge, our analysis is the first comprehensive study that compares different saliency models according to objective measures. For that purpose, we used the ReTargetMe benchmark data set

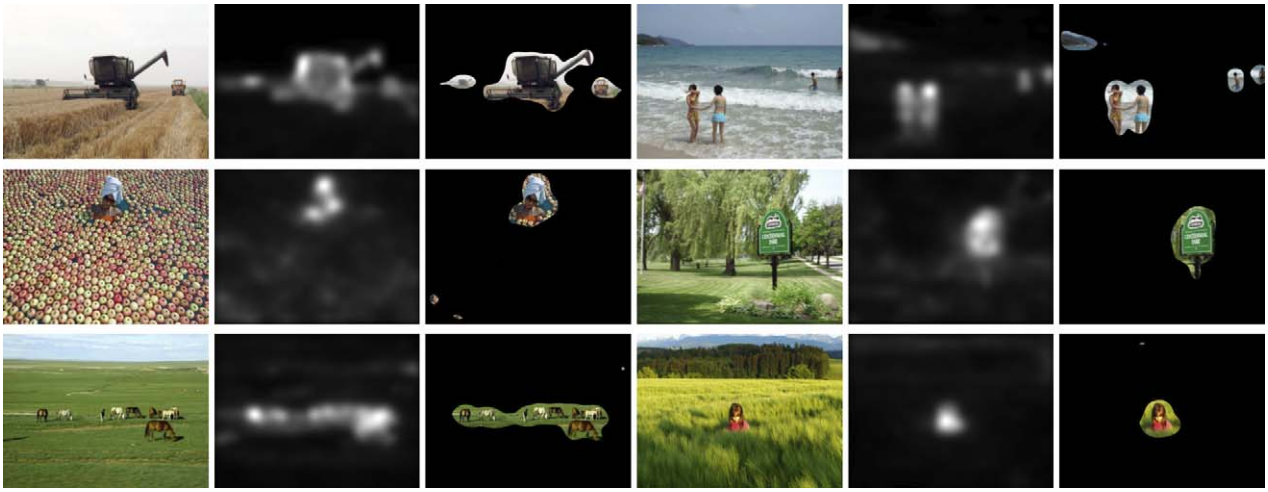


Figure 7. Sample salient object detection results in the ImgSal data set. Here, only covariance features are used in saliency estimation, and the object maps are then obtained from the saliency maps by thresholding them according to the average saliency score plus one standard deviation. As can be seen, the salient objects are captured quite well by the proposed approach.

Downloaded from jov.arvojournals.org on 07/28/2019

	Large salient regions		Intermediate salient regions		Small salient regions		Cluttered backgrounds		Repeating distractors		Large and small salient regions	
	AUC	DSC	AUC	DSC	AUC	DSC	AUC	DSC	AUC	DSC	AUC	DSC
Itti et al. (1998)	0.897	0.610	0.897	0.473	0.937	0.401	0.824	0.335	0.891	0.439	0.936	0.639
Harel et al. (2007)	**0.945**	**0.694**	0.925	0.529	0.951	0.463	0.916	0.499	0.934	0.557	**0.952**	0.688
Torralba et al. (2006)	0.790	0.469	0.825	0.377	0.929	0.372	0.700	0.239	0.750	0.306	0.870	0.515
Hou & Zhang (2007)	0.833	0.524	0.861	0.448	0.939	0.411	0.769	0.308	0.809	0.369	0.918	0.584
Zhang et al. (2008)	0.760	0.461	0.813	0.391	0.895	0.366	0.676	0.270	0.755	0.325	0.850	0.504
Bruce & Tsotsos (2009)	0.798	0.480	0.825	0.383	0.914	0.357	0.759	0.288	0.788	0.350	0.855	0.494
Seo & Milanfar (2009)	0.842	0.563	0.896	0.474	0.948	0.430	0.776	0.284	0.878	0.451	0.916	0.611
Goferman et al. (2010)	0.905	0.636	**0.950**	**0.610**	0.970	0.553	0.919	0.509	0.914	0.581	0.947	0.723
Our approach with												
Covariances only	0.920	0.666	0.928	0.548	0.957	0.470	**0.933**	**0.554**	0.947	0.664	0.946	0.645
Covariances + means	0.866	0.614	0.924	0.584	**0.972**	**0.586**	0.818	0.425	0.948	0.635	0.938	**0.728**
Covariances + center	0.919	0.681	0.909	0.517	0.919	0.329	0.905	0.500	**0.961**	0.654	0.893	0.574
Covariances + means + center	0.865	0.673	0.912	0.580	0.954	0.508	0.879	0.441	0.960	**0.698**	0.888	0.664

Table 4. Performance comparisons of the saliency models on the ImgSal data set. The best performing model is shown in bold type.

(Rubinstein et al., 2010), which also provides an online user survey at http://people.csail.mit.edu/mrub/retargetme.

Data set

The ReTargetMe data set contains 80 images with 92 different resizing scenarios such as reducing or increasing the width or the height of an images by 25% or 50%. The images in this data set are categorized into nine groups based on some visual attributes identified by the authors as important to retargeting objectives. These attributes are *lines/clear edges* (50 images), *faces/people* (26 images), *recurring texture* (10 images), *evident foreground objects* (39 images), *geometric structures* (33 images), *symmetry* (16 images), *textual elements* (five images), *outdoor/nature* (29 images), and *indoor* (nine images). Note that an image can belong to more than one set since it may contain different attributes.

We based our analysis on a widely known image retargeting approach called Seam Carving (Avidan & Shamir, 2007) whose source code is available at http://people.csail.mit.edu/mrub/code/seam_carving-1.0.zip. This method identifies a number of the so-called seams which are paths of least importance in an image, and then removes or inserts them to automatically shrink or enlarge the size of the given image. In particular, we used predicted saliency maps in combination with the forward energy criterion in (Avidan & Shamir, 2007) to guide the image retargeting process.

Evaluation scores

In the literature, little work has been published to quantitatively measure retargeting quality due to the subjective nature of the problem definition. To close this gap, Rubinstein et al. (2010) examined the degree of agreements between the human judgments and several computational image distance metrics through an online user study involving a total of 210 participants at the time of publication. The authors argued that the EMD and SIFTflow scores are the only two objective metrics that were highly consistent with human rankings. They better related to the deformations caused by the retargeting methods, resulting in a more reliable comparison of the image content after resizing.

The EMD is a global metric that measures the distance between the color histograms of two images. SIFTflow is an image registration algorithm that aims at aligning a query image to its neighbors in a large image set. The alignment is formulated as an energy minimization in which the correspondences between similar structures across the images are determined through densely sampled SIFT features. The cost of the estimated alignment gives a matching score between two images. Comparing two images with visually similar contents results in smaller EMD and SIFTflow alignment scores.

Performance

In Figure 8, we present comparative results of three of our models and the GBVS saliency model on two sample images from the ReTargetMe data set. These are the four best models according to overall rankings. As the baseline method, we also include the result of the original Seam Carving method by Avidan and Shamir (2007), which uses only the edge information while enlarging or shrinking the images. Particularly in these examples, our model with covariance features and

Downloaded from jov.arvojournals.org on 07/28/2019

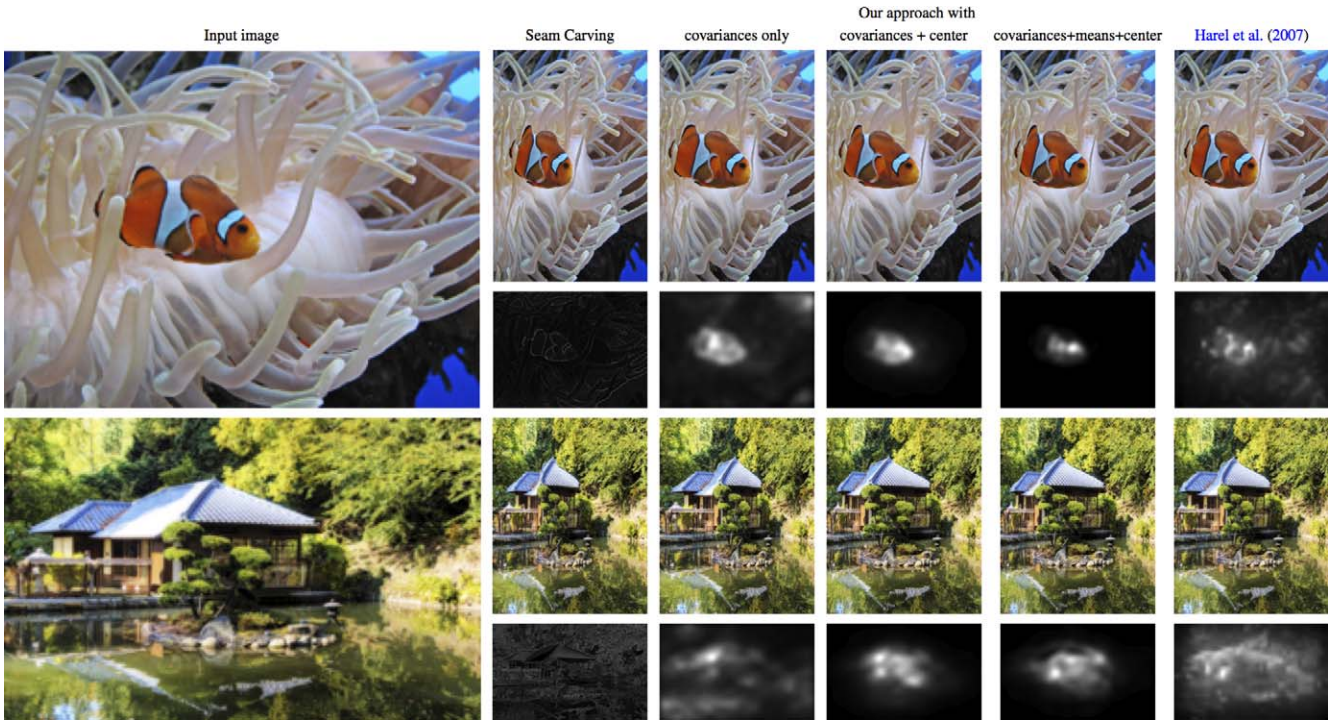


Figure 8. Comparison of retargeting results. The leftmost column shows two sample images from the ReTargetMe dataset. The remaining columns present retargeting results with the corresponding saliency maps. The proposed approach with covariance features successfully preserves the regions corresponding to the prominent objects such the clownfish and the chimney of the house.

implicit center bias preserved the structures of the prominent objects quite well since these objects were predicted more accurately in the related saliency maps.

Table 5 provides the quantitative analysis of the models. In terms of the EMD scores, the proposed model with covariance features and implicit center bias outperformed the other saliency models and the baseline Seam Carving method in eight out of nine categories, and it was the second best model in the remaining category. In terms of the SIFTflow score, the best performing model was mainly the original Seam Carving method, followed by our saliency models and the GBVS model. Our models preserved the prominent areas well and during resizing did not distort them much.

Psychophysical patterns

We tested our saliency models on some psychophysical patterns that are commonly used to explore the mechanisms underlying preattentive visual search (Treisman & Gelade, 1980; Wolfe, 1994). However, we should note that the main goal of this study was to devise a novel saliency model that can accurately detect eye fixations. We do not claim that the proposed framework completely explains all of the psychophysical phenomena but aim to shed some light on the plausibility of our model on some pop-out examples.

Figure 9 shows the set of synthetic patterns considered in the experiments. In each case, there is a target object that has a unique (basic) feature surrounded by an array of distracting objects, thus the target will pop out effortlessly. These patterns include *color pop-out*, *orientation pop-out*, *orientation and color pop-out*, *local pop-out*, *curvature pop-out*, *texture pop-out*, and *conjunction search* examples. A good saliency model should correctly identify the target objects in these images.

As shown in Figure 9, in general, the information-theoretic models proposed in Bruce and Tsotsos (2009) and Zhang et al. (2008) performed poorly as compared to the approaches in Harel et al. (2007) and Itti et al. (1998). This was mostly because SUN and AIM models both employ global image information instead of local surround data. The approach of Goferman et al. (2010) overall performed well on the first six cases; however, it could not distinguish the local pop-out phenomenon in the seventh example. Our model, on the other hand, successfully reproduced the pop-out phenomena in all the patterns considered here, except possibly the last one, which involved a conjunction search. In this example, while human subjects can immediately notice the small, rotated, and red 5's, it takes more effort to spot the 2 on the bottom right. As we considered feature statistics in a combined manner (as opposed to traditional way of treating them separately), our

Downloaded from jov.arvojournals.org on 07/28/2019

	Lines/clear edges		Faces/ people		Recurring texture		Evident foreground objects		Geometric structures		Symmetry		Textual elements		Outdoor / Nature		Indoor	
	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow
Itti et al. (1998)	7.62	8.85	6.48	8.45	6.80	8.50	6.13	8.41	8.08	9.29	8.06	8.06	6.50	4.67	6.62	7.41	8.10	9.70
Harel et al. (2007)	3.95	5.71	4.10	**5.03**	4.20	8.40	4.37	5.17	4.39	6.03	3.59	**4.59**	**2.83**	8.33	**3.41**	5.78	4.50	6.00
Torralba et al. (2006)	7.71	8.60	7.86	8.17	7.20	9.00	8.48	8.57	7.63	8.84	9.65	8.94	9.83	6.67	7.97	7.68	6.40	8.40
Hou & Zhang (2007)	9.71	8.04	8.59	9.17	9.80	5.80	8.61	8.52	9.34	7.84	10.24	6.82	10.00	9.33	9.59	8.32	8.60	9.80
Zhang et al. (2008)	9.05	7.87	10.38	8.07	9.10	7.60	9.59	8.87	9.26	7.95	9.47	8.41	11.00	9.50	9.14	8.14	8.10	7.90
Bruce & Tsotsos (2009)	5.76	10.45	5.86	10.10	5.90	10.80	6.85	9.91	5.61	10.42	6.00	9.94	6.33	8.67	6.32	9.59	6.00	9.40
Seo & Milanfar (2009)	8.42	8.25	8.17	9.00	7.60	7.90	7.91	8.61	7.95	8.21	8.71	9.65	7.00	11.17	7.59	9.41	7.00	8.20
Goferman et al. (2010)	8.78	6.80	9.07	5.69	9.10	6.70	8.61	5.76	8.87	7.03	7.12	7.88	4.67	7.33	7.86	5.97	7.60	8.10
Our approach with																		
Covariances only	5.85	5.56	5.86	5.69	6.50	**4.40**	6.63	6.50	5.29	5.05	5.24	5.82	7.00	6.50	6.59	**5.68**	5.60	4.90
Covariances + means	9.04	**5.09**	8.31	6.00	9.90	5.10	7.89	5.39	9.68	**4.61**	8.24	4.82	8.67	7.00	8.59	5.95	10.50	**4.30**
Covariances + center	**2.84**	5.67	**3.10**	5.52	**2.90**	4.70	**3.26**	5.54	**2.68**	5.45	**2.59**	5.53	**2.83**	5.50	3.73	6.03	**2.70**	4.60
Covariances + means + center	4.75	5.75	5.28	5.14	4.70	5.80	4.87	**4.93**	4.76	5.97	4.18	5.53	5.17	**4.33**	5.30	6.32	7.20	5.90
Seam carving (Avidan & Shamir, 2007)	7.53	4.35	7.93	4.97	7.30	6.30	7.80	4.80	7.45	4.32	7.94	5.00	9.17	2.00	8.27	4.73	8.70	3.80

Table 5. The average rankings of the saliency models and the baseline Seam Carving method on the ReTargetMe data set. The best performing model is shown in bold type.

approach was able to detect only the red 5 and failed to spot the other targets.

Evidence from psychophysical studies suggests that in some situations, there exist asymmetries in visual search. A well-known example of this is the presence–absence asymmetry. It refers to the observation that while the existence of a certain feature in a target object makes it pop-out easily among distractors lacking that feature, the reverse, the absence of a feature of distractors in a target might not lead to a clear pop-out. Figure 10 shows a typical example in which our approach successfully reproduced the asymmetric behavior that detecting the plus symbol among the minus symbols is much easier than the reverse distribution.

Discussion and concluding remarks

In this study, we presented a novel bottom-up saliency model that employs region covariances as features. Our experimental evaluation showed that the proposed approach is highly competitive with the state-of-the-art algorithms on several tasks, including prediction of human eye fixations, salient object detection, and image retargeting. Our framework differed from traditional bottom-up approaches in that it carried out feature map extraction and feature integration steps in a single shot. This was made possible by the use of region covariances, which was the key to the success of our framework. Modeling the statistical dependencies among different features, region covariances efficiently encode local structure information. More importantly, they provide a natural mechanism to nonlinearly integrate different features. This allowed our approach to produce especially accurate predictions for natural images containing texture elements or repeating patterns. We also showed that first-order statistics can be incorporated into our framework as well in a fairly straightforward way.

From the computational perspective, as we stated in the Introduction section, our model and the self-resemblance model by Seo and Milanfar (2009) can be considered somewhat similar in the sense that both models use high-level features (region covariances and LSKs) that nonlinearly combine some basic-level linear features. Seo and Milanfar (2009) computed the MCS between two LSK features, which is equivalent to a weighted sum of vector cosine similarities between each pair of column vectors in the feature matrices, with the weights indicating the relative importance of each feature. As compared to LSKs, which are obtained by the radiometric differences based on image gradients, region covariances are much richer descriptors since they allow encoding correlations among every pair of

Downloaded from jov.arvojournals.org on 07/28/2019

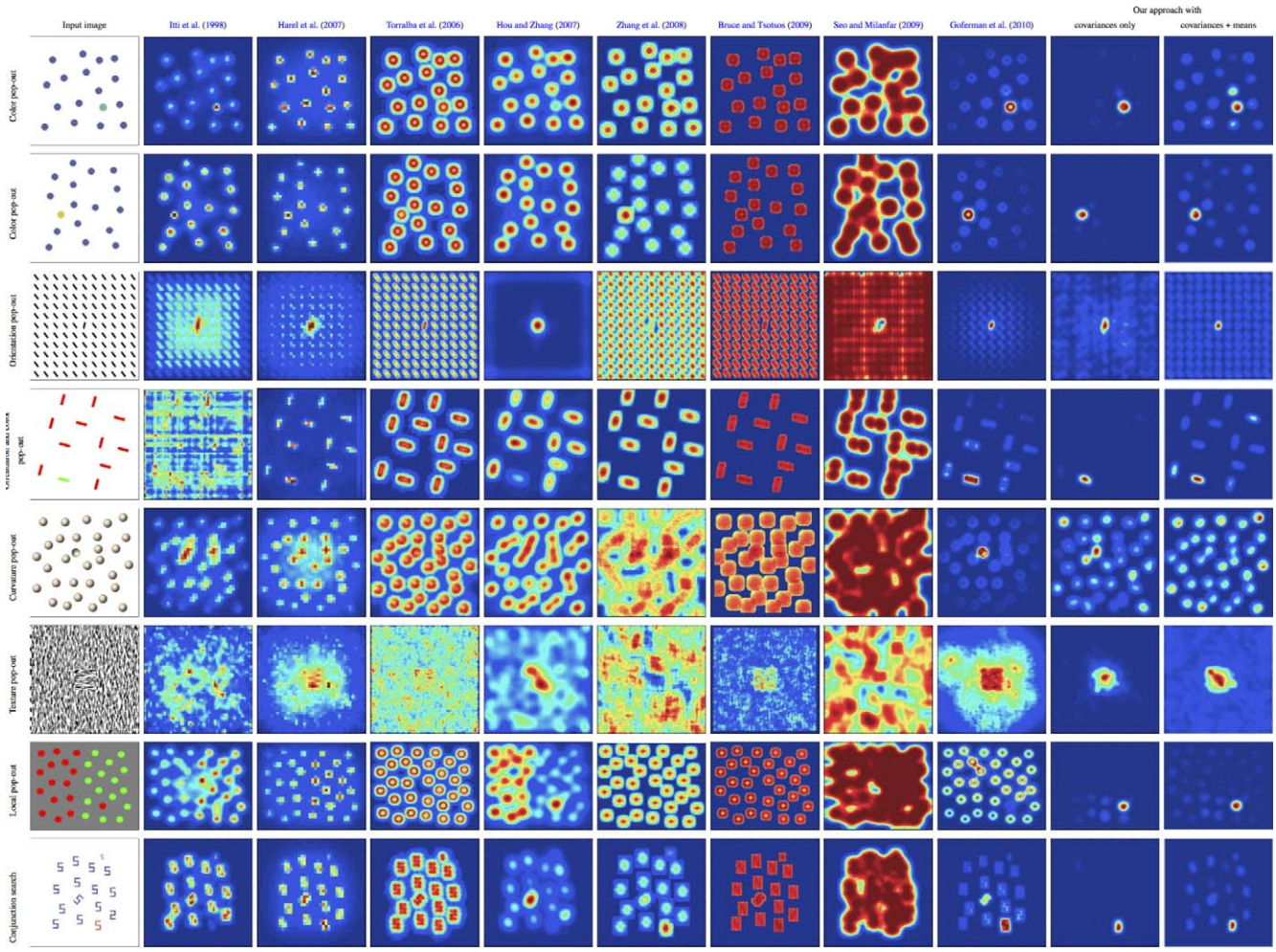


Figure 9. Example psychophysical patterns showing various types of pop-out stimuli and the saliency maps of different bottom-up saliency models. For these examples, our saliency models produce meaningful predictions but fails to correctly predict easy and difficult searches in the final conjunction search example.

different feature channels. The experimental results clearly demonstrate that our approach is significantly better than that of Seo and Milanfar (2009).

In our implementation, we used very simple visual features such as color and orientation, but it is possible to incorporate more complex semantic features. For example, one can use gist of the scene, faces, pedestrians, or text (Cerf, Frady, & Koch, 2009; Judd et al., 2009; Torralba et al., 2006) as such additional features. It might be interesting to seek that direction to incorporate task-oriented top-down influences into our model such as looking for faces or people. Furthermore, we also plan to explore estimation of spatio-temporal saliency in dynamic scenes as another future work. A shortcoming of the proposed model is the strong spatial coincidence assumption considered in the integration of saliency maps extracted at different scales. With this assumption, it might be hard to detect multiple salient objects that appear at different scales. Thus, it would be interesting to extend the proposed

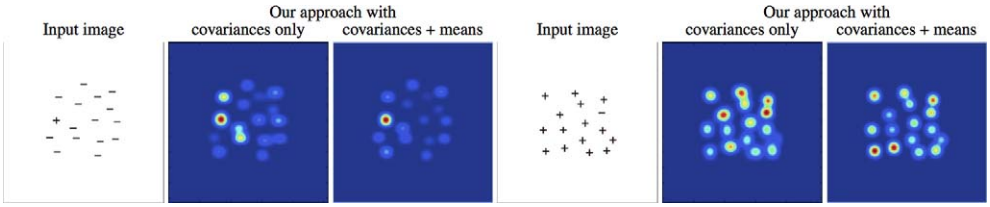


Figure 10. Search asymmetry example. The plus symbol (target) is clearly captured in the saliency map for the first pattern, while the minus symbol is not easily distinguishable among the plus symbols.

Downloaded from jov.arvojournals.org on 07/28/2019

framework to work simultaneously across multiple scales.

Keywords: visual attention, computational saliency model, feature integration, region covariances

Acknowledgments

This research was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK), Career Development Award 112E146. The authors thank anonymous reviewers for their helpful comments.

Author contribution: Both authors contributed equally to this work.
Corresponding author: Erkut Erdem.
Email: erkut@cs.hacettepe.edu.tr.
Address: Department of Computer Engineering, Hacettepe University, Ankara, Turkey.

References

Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1597–1604).

Achanta, R., & Susstrunk, S. (2009). Saliency detection for content-aware image resizing. In *IEEE International Conference on Image Processing (ICIP)* (pp. 1005–1008).

Avidan, S., & Shamir, A. (2007). Seam carving for content-aware image resizing. *ACM Transactions on Graphics, 26*(3).

Bangalore, S. M., & Ma, W.-Y. (1996). Texture features and learning similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(8), 837–842.

Borji, A., & Itti, L. (2012). Exploiting local and global patch rarities for saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 478–485).

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In Y. Weiss, B. Scholkopf, & J. Platt, *Advance in Neural Information Processing Systems (NIPS)*, pp. 155–162. Cambridge, MA: MIT Press.

Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3):5, 1–24, http://www.journalofvision.org/content/9/3/5, doi:10.1167/9.3.5. [PubMed] [Article]

Butko, N. J., Lingyun, Z., Cottrell, G. W., & Movellan, J. R. (2008). Visual saliency model for robot cameras. In *IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2398–2403).

Callaghan, T. C. (1989). Interference and domination in texture segregation: Hue, geometric form, and line orientation. *Perception and Pyschophysics, 46*(4), 299–311.

Callaghan, T. C. (1990). Interference and dominance in texture segregation. In D. Brogan (Ed.), *Visual search* (pp. 81–87). New York: Taylor & Francis.

Cerf, M., Frady, E., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12):10, 1–15, http://www.journalofvision.org/content/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article]

Cerf, M., Harel, J., Einhaeuser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. In J. C. Platt, D. Koller, Y. Singer, & S. Rowels, *Advance in Neural Information Processing Systems (NIPS)*, pp. 241–248. Cambridge, MA: MIT Press.

Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X., & Hu, S.-M. (2011). Global contrast based salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 409–416).

Duan, L., Wu, C., Miao, J., Qing, L., & Fu, Y. (2011). Visual saliency detection by spatially weighted dissimilarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 473–480).

Eckstein, M. P., Thomas, J. P., Palmer, J., & Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception and Pyschophysics, 62*(3), 425–451.

Föerstner, W., & Moonen, B. (1999). *A metric for covariance matrices (Tech. Rep.)*. Department of Geodesy and Geoinformatics, Stuttgart University, Germany.

Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(6), 989–1005.

Gao, D., & Vasconcelos, N. (2007). Bottom-up saliency is a discriminant process. In *IEEE International Conference on Computer Vision (ICCV)* (pp. 1–6).

Downloaded from jov.arvojournals.org on 07/28/2019

Goferman, S., Zelnik-Manor, L., & Tal, A. (2010). Context-aware saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2376–2383).

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In B. Scholkopf, J. Platt, & T. Hoffman (Eds.), *Advance in Neural Information Processing Systems (NIPS)*, pp. 545–552. Cambridge, MA: MIT Press.

Hong, X., Chang, H., Shan, S., Chen, X., & Gao, W. (2009). Sigma set: A small second order statistical region descriptor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1802–1809).

Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Judd, T., Durand, F., & Torralba, A. (2012. January). *A benchmark of computational models of saliency to predict human fixations* (Tech. Rep. No. MIT-CSAIL-TR-2012-001). Cambridge, MA: MIT Computer Science and Artificial Intelligence Laboratory.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)* (pp. 2106–2113).

Julier, S., & Uhlmann, J. K. (1996). *A general method for approximating nonlinear transformations of probability distributions* (Tech. Rep.). Robotics Research Group, Department of Engineering Science, University of Oxford.

Karklin, Y., & Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature, 457*, 83–86.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*(4), 219–227.

Li, J., Levine, M. D., An, X., Xu, X., & He, H. (2012). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(4), 996–1010.

Liu, T., Jian Sun, X. T., Nan-Ning Zheng, & Shum, H.-Y. (2007). Learning to detect a salient object. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Pele, O., & Werman, M. (2009). Fast and robust earth movers distances. In *IEEE International Conference on Computer Vision (ICCV)* (pp. 460–467).

Peters, R., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416.

Porikli, F., Tuzel, O., & Meer, P. (2006). Covariance tracking using model update based on Lie algebra. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 728–735).

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–71.

Puzicha, J., Hofmann, T., & Buhmann, J. M. (1997). Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 267–272).

Rahtu, E., Kannala, J., Salo, M., & Heikkilä, J. (2010). Segmenting salient objects from images and videos. In Kostas Daniilidis, Petros Maragos, & Nikos Paragios (Eds.), *European Conference of Computer Vision (ECCV)*, Series: Lecture Notes in Computer Science, pp. 366–379. Berlin, Heidelberg: Springer.

Rosenholtz, R. (2000). Significantly different textures: A computational model of pre-attentive texture segmentation. In David Vernon (Ed.), *European Conference of Computer Vision*, Series: Lecture Notes in Computer Science, pp. 197–211. Berlin, Heidelberg: Springer.

Rosenholtz, R. (2001). Search asymmetries? What search asymmetries? *Perception & Psychophysics*, *63*(3), 476–489.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research, 39*, 3157–3163.

Rosenholtz, R., Nagy, A. L., & Bell, N. R. (2004). The effect of background color on asymmetries in color search. *Journal of Vision, 4*(3):9, 224–240, http://www.journalofvision.org/content/4/3/9, doi:10.1167/4.3.9. [PubMed] [Article]

Rubinstein, M., Gutierrez, D., Sorkine, O., & Shamir, A. (2010). A comparative study of image retargeting. *ACM Transactions on Graphics, 29*(5):160, 1–10.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The

Downloaded from jov.arvojournals.org on 07/28/2019

earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision, 40*, 99–121.

Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004). Is bottom-up attention useful for object recognition? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 37–44).

Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision, 9*(12):15, 1–27, http://www.journalofvision.org/content/9/12/15, doi:10.1167/9.12.15. [PubMed] [Article]

Siagian, C., & Itti, L. (2007). Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(2), 300–312.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14):4, 1–17, http://www.journalofvision.org/content/7/14/4, doi:10.1167/7.14.4. [PubMed] [Article]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A*, *20*(7), 1407–1418.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97–136.

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7):4, 1–16, http://www.journalofvision.org/content/9/7/4, doi:10.1167/9.7.4. [PubMed] [Article]

Tuzel, O., Porikli, F., & Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In Aleš Lonards, Horst Bischof, & Axel Pinz (Eds.), *European Conference of Computer Vision (ECCV)*, Series: Lecture Notes in Computer Science (3952), pp. 589–600. Berlin, Heidelberg: Springer.

Tuzel, O., Porikli, F., & Meer, P. (2008). Pedestrian detection via classification on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(10), 1713–1727.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 511–518).

Voorhees, H., & Poggio, T. (1988). Computing texture boundaries from images. *Nature, 333*, 364–367.

Wang, Y.-S., Tai, C.-L., Sorkine, O., & Lee, T.-Y. (2008). Optimized scale-and-stretch for image resizing. *ACM Transaction on Graphics, 27*(5):118, 1–8.

Wang, Z., Lu, L., & Bovik, A. C. (2003). Foveation scalable video coding with automatic fixation selection. *IEEE Transactions on Image Processing, 12*, 243–254.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception & Performance*, *15*(3), 419–433.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7):32, 1–20, http://www.journalofvision.org/content/8/7/32, doi:10.1167/8.7.32. [PubMed] [Article]

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision, 11*(3):9, 1–15, http://www.journalofvision.org/content/11/3/9, doi:10.1167/11.3.9. [PubMed] [Article]

Zhao, Q., & Koch, C. (2012). Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost. *Journal of Vision, 12*(6):22, 1–15, http://www.journalofvision.org/content/12/6/22, doi:10.1167/12.6.22. [PubMed] [Article]

Downloaded from jov.arvojournals.org on 07/28/2019