Evaluating Visual Saliency Algorithms: Past, Present and Future

Puneet Sharma
Department of Engineering & Safety (IIS), University of Tromsø(UiT), Tromsø, Norway
E-mail: er.puneetsharma@gmail.com

Abstract. *A salient image region is defined as an image part that is clearly different from its surround. This difference is measured in terms of a number of attributes, namely, contrast, brightness and orientation. By measuring these attributes, visual saliency algorithms aim to predict the regions in an image that would attract our attention under free viewing conditions. As the number of saliency models has increased significantly in the past two decades, one is faced with the challenge of finding a metric that can be used to objectively quantify the performance of different saliency algorithms. To address this issue in this article, first, the state of the art of saliency models is revisited. Second, the major challenges associated with the evaluation of saliency models are discussed. Third, ten frequently used evaluation metrics are examined and their results are discussed for ten latest state-of-the-art saliency models. For the analysis, a comprehensive open source fixations database has been quantitatively examined.* *© 2015 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2015.59.5.050501]

INTRODUCTION

Our visual system is selective, i.e., we concentrate on certain aspects of a scene while neglecting other things. This is evident from studies on change blindness,[1–3] which show that large changes can be made in a visual scene that can remain unnoticed. The reason why our visual system is selective is because our brains do not process all the visual information in a scene. In fact, while the optic nerve receives information at the rate of approximately 3×10^6 bits/s, the brain processes less than 10^4 bits/s of this information.[4] In other words, the brain uses a tiny fraction ($<1\%$) of the collected information to build a representation of the scene, a representation that is good enough to perform a number of complex activities in the environment such as walking, aiming at objects and detecting objects. Based on this, we can ask what mechanisms are responsible for building this representation of the scene.

In the literature, two main attention mechanisms are discussed: top-down and bottom-up.[5–11] Top-down is voluntary, goal-driven and slow, i.e., typically in the range between 100 ms and several seconds.[9] It is assumed that the top-down attention is closely linked with cognitive aspects such as memory, thought and reasoning. For example, by employing top-down mechanisms, we can attend to a person sitting next to us in a busy restaurant and neglect other people and visual information in the background. In contrast, bottom-up attention (also known as visual saliency) is associated with attributes of a scene that draw our attention to a particular location. These attributes include motion, contrast, orientation, brightness and color.[12] Bottom-up mechanisms are involuntary, and faster than top-down.[9] For instance, flickering lights, a yellow target among green objects, and a horizontal target among vertical objects are some stimuli that would automatically capture our attention in the environment. Studies[13,14] show that in search tasks, such as looking for a target object among distractors, both bottom-up and top-down mechanisms work together to guide our attention. While bottom-up attention is based on elementary attributes of a scene, top-down is quite complex and strongly influenced by task demands.[15] For example, the studies (Refs. 16, 17) suggest that for tasks such as picking up and placing objects by hand in the environment, attention is mainly driven by top-down mechanisms.

In the past two decades, modeling of visual saliency has generated a lot of interest in the research community. In addition to contributing towards the understanding of human vision, it has also paved the way for a number of computer vision applications. These applications include target detection,[18] image and video compression,[19–22] image segmentation,[23] context aware image resizing,[24] robot localization,[25,26] image retrieval,[27] image and video quality assessment,[28,29] dynamic lighting,[30] advertisement,[31] artistic image rendering[32] and human–robot interaction.[33,34]

To evaluate the performance of visual saliency algorithms, the two-dimensional saliency maps are compared with the image regions that attract observers' attention.[35–37] This is done by displaying to the observers a set of images and using an eye tracker to record their eye fixations. Further, it is thought that a higher number of fixations correspond to salient image regions. The recorded fixations are thus compared with the associated visual saliency maps in a pairwise manner.[38–40] Unfortunately, studies[32,41,42] have shown that while viewing images, observers tend to fixate on the center of the image more than the peripheral regions. This effect is known as center bias and is well documented in vision studies.[41,43] The presence of center bias in fixations makes it difficult to evaluate the correspondence between the fixated regions and the salient image regions. This can be explained by the fact that in a comprehensive eye tracking study by Judd et al.,[32] it was shown that a dummy

Received Oct. 23, 2013; accepted for publication June 1, 2015; published online Aug. 4, 2015. Associate Editor: Susan Farnand.
1062-3701/2015/59(5)/050501/17/$25.00

classifier defined by a Gaussian blob at the center of the image was better at predicting the eye fixations than any of the visual saliency models.[35,44,45] In light of these results, one can ask what factors influence the performance of saliency algorithms, and which metric is a good candidate for comparing the different saliency models. This article is an attempt to address these issues.

The rest of the article is organized as follows. In the second section, the literature on visual saliency algorithms is reviewed. Next, in the third section, the different metrics used for judging the performance of saliency models are discussed. Finally, in the fourth section, the results obtained for the evaluation metrics are examined.

STATE OF THE ART IN MODELING VISUAL ATTENTION

In this section, the computer models for predicting eye fixations in still images are discussed.

In the study by Borji et al.,[46] the authors state that the visual attention models in the literature can be divided into eight classes: Bayesian, cognitive, decision theoretic, graphical, information theoretic, pattern classification, spectral analysis, and others. The classification made by Borji et al.[46] has been updated with the latest saliency models. The different classes and the attention models associated with these classes are shown in Table I.

In Bayesian models, prior knowledge about the scene, and sensory information such as target features are employed to calculate salient image regions. The objective is to learn from past search experiences in similar scenes, and use those strategies that lead to a successful search (of the target). For instance, models such as those of Torralba,[47] Olivia et al.[39] and Zhang et al.[48] fall in this category.

Cognitive models are the ones that are strongly based on psychological and neuro-physiological findings from experiments. For instance, experiments have shown that difference of Gaussians (DOG) is a good approximation of how the receptive fields (i.e., basic units of our visual system) extract information from a scene before sending it to the brain.[49] This finding is the basis for many cognitive saliency models such as those of Itti et al.,[44] Walther,[50] Walther et al.,[51] Frintrop[52] and Borji and Itti.[53] Other models in this category include those of Meur et al.,[54] Rajashekar et al.,[55] Cerf et al.,[36] Erdem and Erdem[24] and Alsam et al.[56,57] We can see that a majority of saliency models belong to this category.

Decision theoretic models are based on the concept of identifying the optimal factors based on how people make decisions. Saliency is defined in terms of discrimination analysis, where salient features are those that best distinguish the target objects from all other visual classes of interest. For instance, models such as those of Gao and Vasconcelos,[58] Gao et al.,[59] Li et al.[60] and Wang et al.[61] are classified under this category.

A graphical model is a probabilistic model which represents a set of calculated image features as pairs connected by links, where the interconnected features are vertexes, and the links connecting some pairs of vertexes are called edges. For example, models such as those of Harel et al.,[35] Achanta et al.,[23] Avraham and Lindenbaum,[62] Chikkerur et al.[63] and Liu et al.[64] belong to this class.

Table I. The various visual attention models and their categories according to the study by Borji et al.[46] The classification made by Borji et al.[46] has been updated with the latest saliency models.

Bayesian models	Torralba,[47] Olivia et al.[39] and Zhang et al.[48]
Cognitive models	Itti et al.,[44] Walther,[50] Walther et al.,[51] Frintrop,[52] Meur et al.,[54] Rajashekar et al.,[55] Cerf et al.,[36] Murray et al.,[78] Erdem and Erdem,[24] Alsam et al.[56,57] and Borji and Itti[53]
Decision theoretic models	Gao and Vasconcelos,[58] Navalpakkam and Itti,[11] Gao et al.,[59] Li et al.[60] and Wang et al.[61]
Graphical models	Harel et al.,[35] Achanta et al.,[23] Avraham and Lindenbaum,[62] Chikkerur et al.[63] and Liu et al.[64]
Information theoretic models	Bruce & Tsotsos,[65] Mancas,[66] Seo and Milanfar,[67] Erdem and Erdem[24] and Borji and Itti[53]
Pattern classification models	Judd et al.[32] and Kienzle et al.[68]
Spectral analysis models	Hou and Zhang,[69] Guo et al.,[70] Achanta et al.,[23] Bian and Zhang[71] and Schauerte and Stiefelhagen[72]
Other models	Rao et al.,[73] Goferman et al.,[74] Kootstra et al.[86] and Garcia-Diaz et al.[75]

Information theoretic models are based on the concept that localized saliency computation serves to maximize information sampled from one's environment. In other words, these models select the most informative parts of the image and discard the rest. This class consists of models such as those of Bruce & Tsotsos,[65] Mancas,[66] Seo and Milanfar,[67] Erdem and Erdem[24] and Borji and Itti.[53]

In pattern classification models, a machine learning procedure is employed to model visual attention. For learning salient regions in images, typically the regions pertaining to eye fixation data or labeled salient regions are used as ground truth. For instance, models such as those of Judd et al.[32] and Kienzle et al.[68] are classified under this category.

Spectral analysis models calculate saliency in the frequency domain. This category consists of models such as those of Hou and Zhang,[69] Guo et al.,[70] Achanta et al.,[23] Bian and Zhang[71] and Schauerte and Stiefelhagen.[72]

The models that do not conform to the above categories are classified as other models. This class includes models such as those of Rao et al.,[73] Goferman et al.[74] and Garcia-Diaz et al.[75]

In this section, the saliency models are reviewed in terms of the above mentioned categories.

Cognitive Models

We start with cognitive models, as they were the earliest saliency models and they form the basis for many of the models in other categories.

The classic model of visual saliency proposed by Itti et al.[44] calculates salient regions by decomposing the input image into three different channels, namely, color,

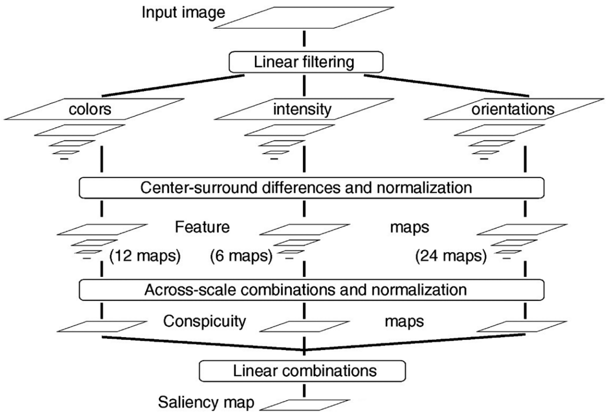


Figure 1. The general architecture of the saliency model by Itti et al.[44]

intensity and orientation, as shown in Figure 1. The color channel consists of two maps, red/green and blue/yellow color opponencies, the intensity channel consists of a gray-scale representation of the input image, and the orientation channel contains four local orientation maps associated with angles of 0, 45, 90, and 135 degrees. For each channel map, nine spatial scales are created by repeatedly low-pass filtering and subsampling the input channel. After that, feature maps are computed by using center-surround operations, which are inspired by vision studies such as Refs. 12, 76. The center-surround operations are defined as the difference between fine and coarse scales. For example, if the center is a pixel at scale $c \in \{2, 3, 4\}$, the surround is the corresponding pixel at scale $s = c + d$, with $d \in \{3, 4\}$, and $\ominus$ denotes the across-scale difference, then the center-surround feature maps for a channel I are represented as

$$I(c, s) = |I(c) \ominus I(s)|.$$

These operations generate 42 feature maps: six for intensity, 12 for color opponencies and 24 for orientation. Next, the maps associated with each channel are normalized and combined to generate three conspicuity maps (i.e., intensity, color and orientation). Finally, the resulting conspicuity maps are normalized and combined linearly to obtain the so-called saliency map. The VOCUS model proposed by Frintrop[52] and the saliency toolbox implemented by Walter et al.[50,51] are based on this saliency model.

Rajashekar et al.[55] proposed a bottom-up model that calculates salient image regions based on four foveated low-level image features, namely, luminance, contrast, luminance-bandpass and contrast-bandpass. The input image is divided into uniform regions, and the feature maps associated with the four low-level features are calculated. Finally, the four maps are linearly combined using a weighted average to obtain the saliency map. For evaluation, they used 101 static gray-scale images that contained no high-level features such as animals, faces or other items of high-level semantic interest.

Meur et al.[54] presented a saliency model inspired by various properties of the human visual system such as contrast sensitivity function, visual masking and perceptual grouping. This model is based on the saliency framework proposed in Ref. 12, and the saliency map is built by linearly combining the different feature maps. The authors showed that their model outperforms the saliency model proposed by Itti et al.[44]

Cerf et al.[36] proposed a model that combined the bottom-up feature channels of color, intensity and orientation, from Ref. 44, with a face-detection channel, based on the algorithm in Ref. 77. Their results showed that the combined model improves the correspondence between the fixated and salient image regions.

Murray et al.[78] calculated salient image regions in three steps. First, the input image is processed according to operations consistent with early visual pathway (color-opponent and luminance channels, followed by a multi-scale decomposition). Second, a simulation of the inhibition mechanisms present in cells of the visual cortex is performed; this step effectively normalizes their response to stimulus contrast. Third, the model integrates information at multiple scales by performing an inverse wavelet transform directly on weights computed from the non-linearization of the cortical outputs. Their saliency model showed better correspondence with the fixations than the saliency models in Refs. 65, 67.

Borji and Itti[53] presented a model based on combining local and global saliency. Local saliency is described as the rarity of an image region with respect to its neighboring regions. It is calculated as the average weighted dissimilarity between the center and L neighboring regions as $S_l = \frac{1}{L}\sum_{j=0}^{L} W_{ij}^{-1} D_{ij}^{c}$, where W_{ij} is the Euclidean distance between the center region (i) and the neighboring region (j), and D_{ij} is obtained by the basis vectors associated with the sparse coding of image regions. Global saliency is defined as the rareness of a region to be selected over the entire image and is calculated based on the information theoretic approach proposed by Bruce & Tsotsos.[65] In addition, the authors use both RGB and Lab color spaces to calculate the saliency maps. The results suggest that their model outperforms ten state-of-the-art saliency models.

Erdem and Erdem[24] introduced a model that calculates saliency based on covariance image descriptors proposed by Tuzel et al.[79] In their model, the image is decomposed into square regions and each region is represented in terms of a covariance descriptor as $C_R = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \mu)(f_i - \mu)^T$, where C_R is a d by d covariance matrix of a region R inside the image, f_i denotes the d-dimensional points associated with features such as intensity, color, orientation and spatial attributes, and μ is their mean. Saliency is measured as the rarity of an image region with respect to its neighboring regions, by using a dissimilarity measure similar to that in Ref. 53. Furthermore, mean feature information is added to make salient regions pop out, and center bias is added to improve its correspondence with eye fixations. The maps are calculated at different scales and combined to form the final saliency map.

Alsam et al.[56,57] presented a model that uses asymmetry as a measure of saliency. For this, the authors use the dihedral group D_4, which is the symmetry group of the

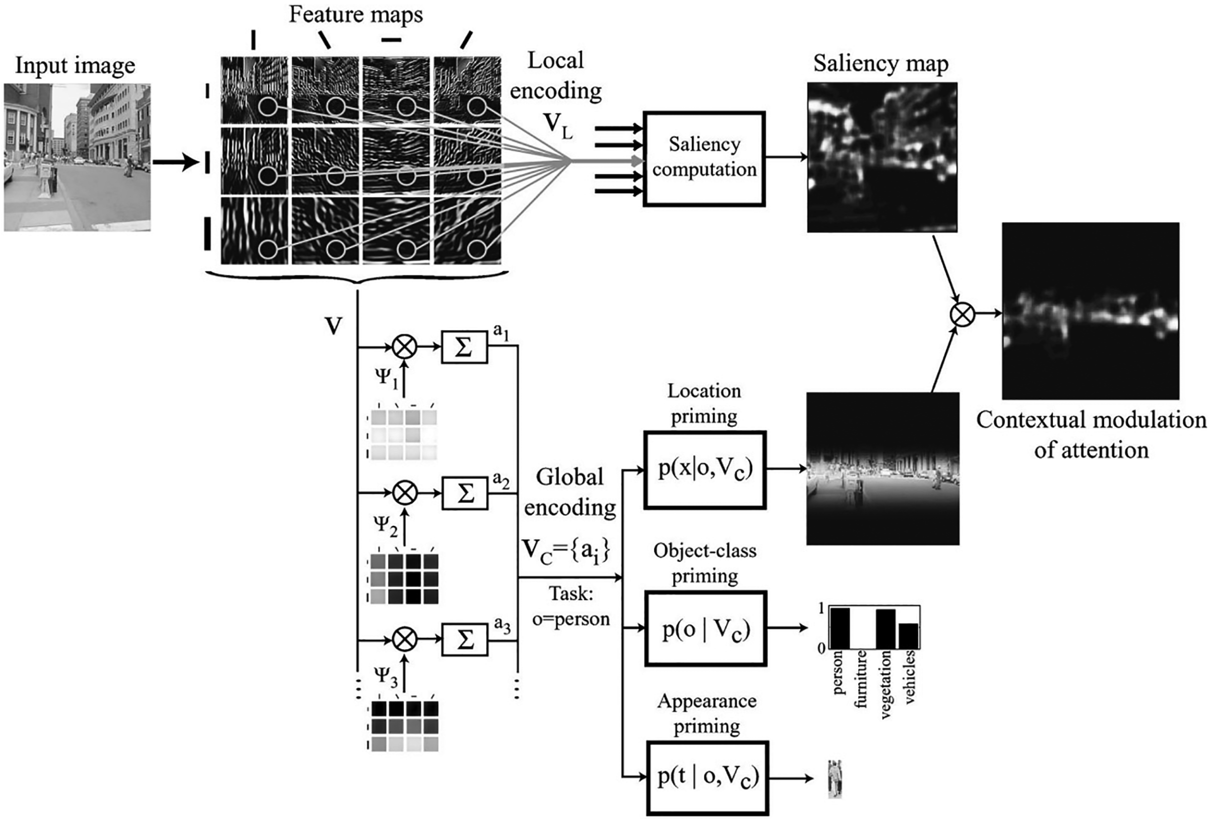


Figure 2. The general architecture of the saliency model by Torralba.[47]

square image grid and includes two types of symmetries, i.e., rotation and reflection. To calculate saliency, the input image is decomposed into square blocks, and for each block the absolute difference between the block itself and the result of the D_4 group element acting on the block is calculated. The mean of the absolute difference for each block is used as a measure of asymmetry for the block. The asymmetry values for all the blocks are then collected in an image matrix and scaled up to the size of the original image using bilinear interpolation. In order to capture both the local and the global salient details in an image, three scales are used. All maps are combined linearly to obtain a single saliency map.

Bayesian Models

Torralba[47] and Olivia et al.[39] defined a model that combines three factors: bottom-up saliency, object likelihood and contextual prior. The local saliency is calculated as $S(x) = \frac{1}{p(v_L/v_C)}$, where v_L encodes local features in the neighborhood of a location that is represented by the outputs of multi-scale oriented bandpass filters, and v_C represents the contextual properties of the scene or background, which include global image statistics, color histograms and wavelet histograms. In the object likelihood factor (represented as priming in Figure 2), the locations corresponding to features different from the target object are suppressed, and the locations with similar features are maintained. The contextual priors stage modifies the two-dimensional saliency map based on past search experience from similar images and uses the strategies that were successful in finding the target.

The SUN model by Zhang et al.[48] defined saliency as a combination of three components. The first contains self-information, which depends only on the visual features at a location. Here, rarer features are considered more informative. In the second, top-down information such as the knowledge about the attributes of the target is used to obtain a log likelihood. The third component consists of the probability associated with the knowledge of the location of the target. In their algorithm, the saliency map was calculated using difference of Gaussians and independent component analysis derived features.

Decision Theoretic Models

Navalpakkam and Itti[11] introduced a model that combines top-down and bottom-up aspects of attention. The bottom-up component is calculated by using the saliency model by Itti et al.,[44] and the top-down component uses the information about the target and the background objects to maximize the ratio between the saliency values of the targets and that of the background objects. This model was evaluated using a search task, i.e., the observers were instructed to search for a specific object in the scene. Their results showed that a combined top-down and bottom-up model yields a faster search than a bottom-up model.

Gao et al.[59] defined saliency as equivalent to discrimination, i.e., they state that the most salient features are the

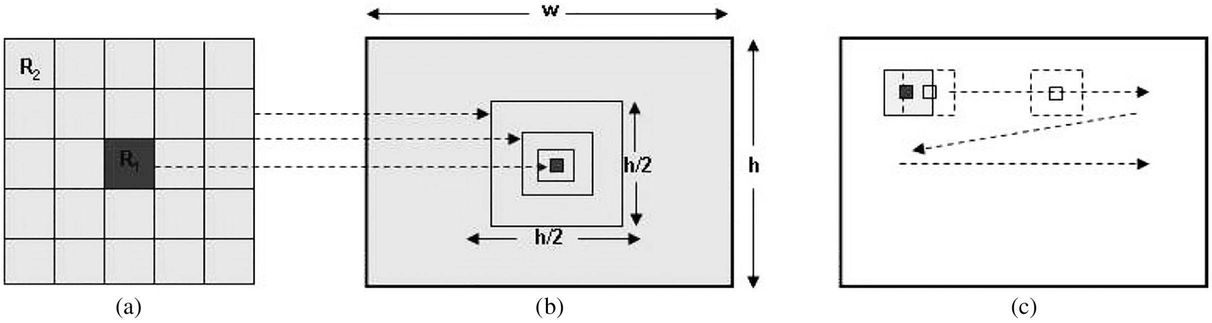


Figure 3. (a) Contrast detection filter showing inner square region R_1 and outer square region R_2. (b) The width (w) of R_1 remains constant while that of R_2 ranges from $w/2$ to $w/8$. (c) The image is filtered at one of the scales in a raster scan fashion (figure used from the article by Achanta et al.[23]).

ones that best separate the target class from all others. In their model, saliency is represented by two components: feature selection and saliency detection. The best feature subset is selected by computing the marginal mutual information as

$$I(X;Y) = \sum_i P_Y(i) D_{KL}(P_{X|Y}(x|i) \parallel P_X(x)),$$

where X is a set of features and Y is a class label with prior probabilities $P_Y(i)$, such that the probability density of X_k given class i is $P_{X_{k|Y}}(x|i)$, and D_{KL} is the Kullback–Leibler divergence.[80] In the saliency detection, the features that are considered highly non-salient are eliminated by employing the Barlow principle of inference.[81]

Li et al.[60] introduced a model that measures saliency as minimum conditional entropy. In their model, the minimum conditional entropy represents the uncertainty of the center-surround local region, when the surrounding area is given and the perceptional distortion is considered. The authors state that the larger the uncertainty the more salient the center is, and vice verse. The minimum conditional entropy is approximated by the lossy coding length of Gaussian data. Finally, the saliency map is segmented by thresholding to detect the salient objects. In their results it was shown that their model outperforms the saliency model in Ref. 44.

Graphical Models

Harel et al.[35] proposed a bottom-up model that uses graph algorithms for saliency computations. In their model, the first step consists of calculating feature maps using a procedure similar to that of Itti et al.[44] After that, a fully connected graph for the locations of the feature maps is built. A graph comprises nodes or vertexes connected by links or edges. The weights between two nodes are calculated based on their dissimilarity and their closeness. Given two locations (i, j) and (p, q) in the feature map M, the dissimilarity between their respective nodes $M(i, j)$, $M(p, q)$ is defined as

$$d((i,j) \parallel (p,q)) \triangleq \left| \log \frac{M(i,j)}{M(p,q)} \right|.$$

Next, the graphs obtained are treated as Markov chains, and the equilibrium distributions of these chains are adopted as the activation maps. Finally, these activation maps are normalized using another Markovian algorithm to highlight the conspicuity, and admitting their combination to form the saliency map.

Achanta et al.[23] presented a model that represents saliency as the local contrast of an image pixel with respect to its neighborhood at different scales. For a given scale, the saliency value at a pixel (i, j) is calculated as the distance D between the mean vectors of pixel features of the inner region R_1 and the outer region R_2 as

$$c_{i,j} = D\left[\left(\frac{1}{N_1}\sum_{p=1}^{N_1} v_p\right), \left(\frac{1}{N_2}\sum_{p=1}^{N_2} v_q\right)\right],$$

where N_1 and N_2 are the numbers of pixels associated with the regions R_1 and R_2 as depicted in Figure 3. In their model, *CIELAB* color space is used to generate feature vectors for color and luminance. The final saliency map is obtained by summing the saliency values across the different scales.

Chikkerur et al.[63] presented a Bayesian model of attention based on the concept that the task of the visual system is to recognize what is where, and this is archived by localizing sequentially, i.e., one object at a time. Their model extends the template-based approach used in the model in Ref. 73, in the following ways. First, both feature and object priors are included, which allows top-down feature-based attention and spatial attention to be combined. Second, this model allows a combination of N feature vectors that share common spatial modulation. Third, in the spatial attention, scale/size information is used in addition to the location information. The authors state that their model combines bottom-up, feature-based and context-based attention mechanisms, and in so doing it is able to explain part of the basic functional anatomy of attention.

Liu et al.[64] introduced a supervised approach to calculating salient image regions. The salient object detection is formulated as an image segmentation problem, where the objective is to separate the salient object from the image background. To do this in their model, ground truth salient objects are obtained from the regions labeled by the observers as salient. After that, a set of features including multi-scale contrast, center-surround histogram and color spatial distribution are used to describe a salient object

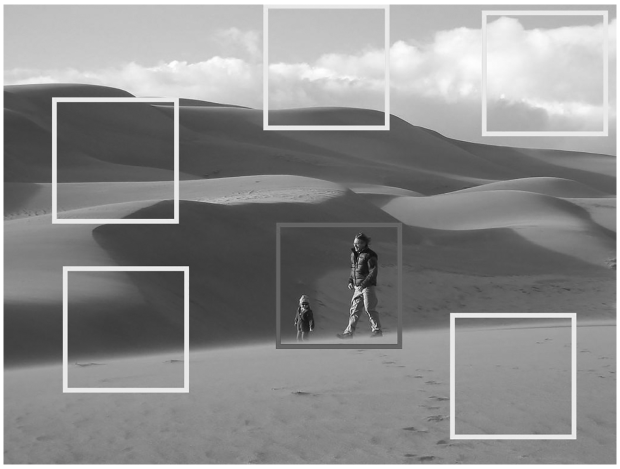


Figure 4. For the given image (used from Judd et al.[32]), the information carried by the image patch at the center (represented by the red square) is quite different from all the other patches (represented by yellow squares), and hence has the greatest self-information (as per Shannon's criterion).[65]

locally, regionally and globally. Finally, these features are optimally combined through conditional random field (CRF) learning. The CRF was trained and evaluated for a large dataset containing 20,840 labeled images by multiple users.

Avraham and Lindenbaum[62] presented a stochastic model of visual saliency. In their model, first, the input image is segmented into regions that are considered as candidates for attention. An initial probability for each candidate is set using preferences such as small number of expected targets. After that, each candidate is represented by a feature vector, and visual similarity between every two candidates is evaluated using a Pearson correlation coefficient. Next, a tree-based Bayesian network is employed for clustering the candidates. Finally, the saliency map is obtained by selecting the most likely candidates.

Information Theoretic Models

Lee & Yu[82] proposed a theoretical model based on the assumption that our visual system operates on the principle of information maximization, i.e., we fixate at a location in the image that provides the maximum amount of information. They proposed that mutual information among cortical representations of the retinal image, the priors constructed from our long-term visual experience and a dynamic short-term internal representation constructed from recent saccades provides the map for the guidance of eye movements. Based on this approach, a similar model was defined in Ref. 83.

Bruce & Tsotsos[65] introduced a saliency model based on the principle of maximizing information that uses Shannon's self-information measure. The saliency is defined by the self-information associated with each local image region.

For instance, as shown in Figure 4, the information carried by the image patch at the center (represented by the red square) is quite different from all the other patches (represented by yellow squares), and hence has the greatest self-information. The self-information is given by $I(X) = -\log(p(X))$, where X is an n-dimensional feature vector extracted from an image region, and $p(X)$ is the probability of observing X based on its surround. The authors state that there are insufficient data in a single image to provide a reasonable estimate of the probability distribution. To address this issue, they employ independent component analysis (ICA) in order to learn the bases from a large database of natural images. After that, the probability of observing an image region is calculated for each basis coefficient. Finally, for a given image region the likelihood of observing it is represented by the product of corresponding ICA basis probabilities for that region.

Seo and Milanfar[67] introduced a bottom-up model based on a self-resemblance measure. In their model, image features are obtained by using local regression kernels, which are quite robust to noise and efficient at capturing the underlying structure of the image. After that, matrix cosine similarity is used to compute the resemblance of each location to its surroundings. The saliency for a given location i is represented as

$$S_i = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{-1+\rho(F_i,F_j)}{\sigma^2}\right)},$$

where σ is a weight parameter and $\rho(F_i, F_j)$ is the matrix cosine similarity between two feature maps F_i and F_j. Here, the matrix cosine similarity is defined as the Frobenius inner product between two normalized matrices F_i and F_j. The authors showed that their model predicts fixations better than the models in Refs. 48, 65.

Mancas[66] defined saliency as a measure of two components, contrast and rarity, i.e., rare features in an image are interesting. To account for contrast two methods are proposed: global and local. Global contrast is measured using a histogram, and local contrast is calculated using center-surround operations similar to that of Ref. 44. The rarity is quantified by employing Shannon's self-information measure. First, a low-level saliency map is calculated by describing each location by the mean and the variance of its neighborhood. After that, rarity is measured based on the features such as size and orientation, where smaller areas and lines corresponding to the orientations get higher saliency values on the saliency map. Finally, high-level methods such as Gestalt laws of grouping are employed to find the salient regions.

Wang et al.[61] proposed a computational model based on the principle of information maximization. Their model considers three key factors, namely, reference sensory responses, fovea-periphery resolution discrepancy and visual working memory. In their model, first, three multi-band filter response maps are calculated as a coherent representation for the three factors. After that, the three filter response maps are combined into multi-band residual filter response maps. Finally, the saliency map is obtained by calculating the residual perceptual information at each location. The results from the authors showed that their model performs significantly better than the saliency model in Ref. 44.

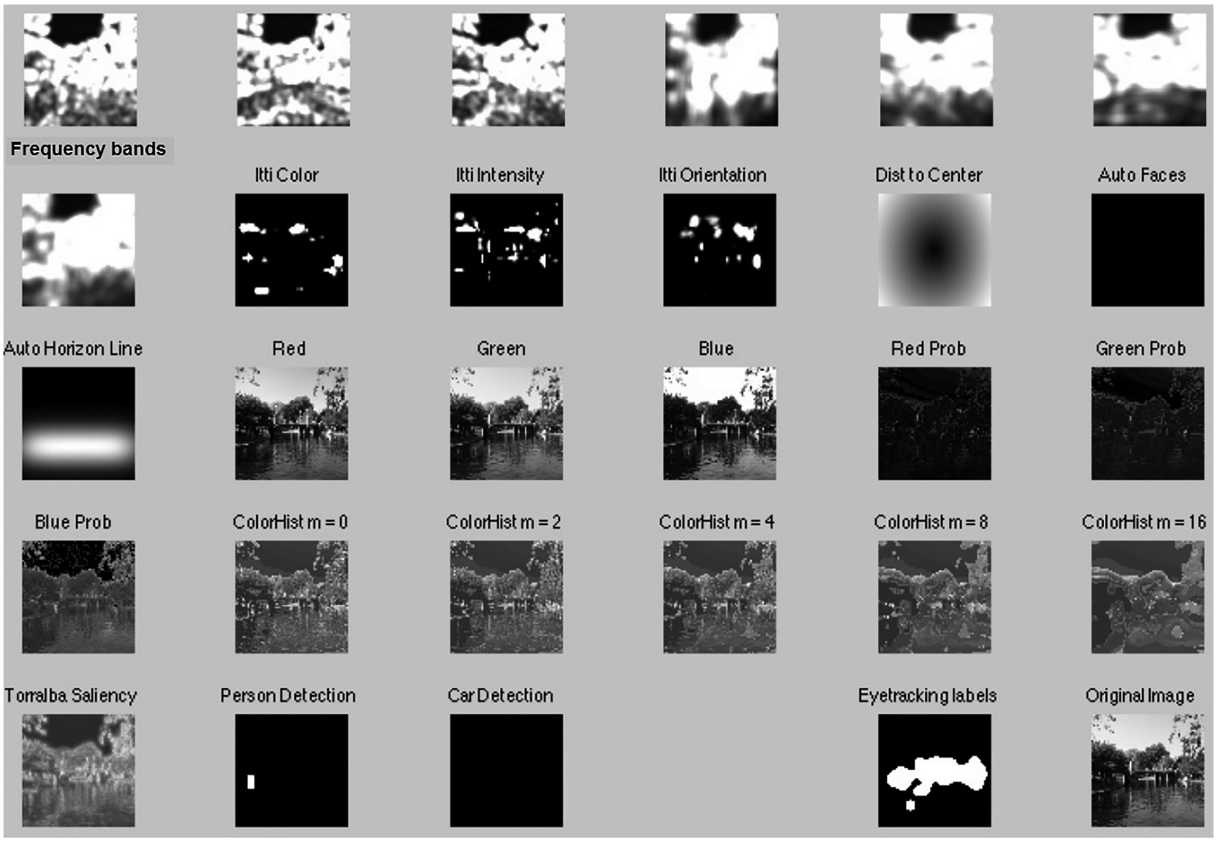


Figure 5. The features used for saliency calculation by Judd et al.[32] These include subband features, Itti and Koch saliency channels, distance to the center, color features and automatic horizon, face, person and car detectors.

Pattern Classification Models

Judd et al.[32] used a machine learning approach to train a combined bottom-up, top-down model based on low-, mid- and high-level image features. As shown in Figure 5, the low-level features such as intensity, orientation and contrast are described by models such as Refs. 44, 45, 84, the mid-level features are represented by a horizon line detector, and the high-level features consist of people and face detectors. The authors collected eye fixations of 15 observers from a comprehensive dataset (with 1003 images) which was also used for evaluation. The model proposed by the authors showed better correspondence with the fixations than several other models such as Refs. 36, 44, 45, 84.

Kienzle et al.[68] proposed a non-linear machine learning approach for calculating saliency. In their model, the intensities pertaining to local image regions are used as feature vectors. The authors employ a support vector machine to train the feature vectors of fixated regions to yield positive values and the feature vectors of randomly selected regions to yield negative values. The resultant saliency is modeled with four perceptive fields, two most likely image structures and two least likely patterns for driving fixations. For the training and evaluation, a dataset of 200 gray-scale images was used.

Spectral Analysis Models

Hou and Zhang[69] proposed a saliency model based on analyzing the log spectrum of the input image. First, the log spectrum is defined as $L(f) = \log(A(f))$, where $A(f)$ is the amplitude of the Fourier spectrum of the image. After computing the log spectrum, the spectral residue is calculated as $R(f) = L(f) - A(f)$. Finally, the spectral residue is transformed to the spatial domain to obtain the saliency map. The results from the authors suggested that their model predicts the fixations better than the saliency model in Ref. 44.

Guo et al.[70] calculated saliency in a manner similar to the spectral residue approach in Ref. 69, with the exception that this model excludes the computation of the spectral residue in the amplitude spectrum. They state that by excluding the amplitude computation the saliency map is obtained faster. For a given image $I(x, y)$, the saliency map is defined as

$$sM(x, y) = g(x, y) * \|F^{-1}[e^{i \cdot p(x,y)}]\|^2,$$

such that $f(x, y) = F(I(x, y))$ and $p(x, y) = P(f(x, y))$, where F and F^{-1} represent the Fourier transform and inverse Fourier transform, respectively. $P(f)$ denotes the phase spectrum of the image and $g(x, y)$ is a two-dimensional Gaussian filter.

Bian and Zhang[71] adopted a spectral approach similar to Ref. 70 for calculating salient image regions. In their model, the input image is resized to a fixed scale, and a windowed Fourier transform of the image is calculated to obtain a spectral response. The spectral response, denoted by $f(u, v)$, is then normalized as $n(u, v) = f(u, v)/\|f(u, v)\|$. After that, $n(u, v)$ is transformed to the spatial domain by using an inverse Fourier transform followed by squaring to promote the salient regions. The resultant saliency map is convolved with a Gaussian filter g to model the spatial pooling operations of complex cells as $S(x, y) = g(u, v) * \|F^{-1}[n(u, v)]\|$, where F^{-1} denotes the inverse Fourier transform.

Schauerte and Stiefelhagen[72] proposed a saliency model that extends the spectral residue approach of Ref. 69. Saliency is calculated using the inverse discrete cosine transform (DCT), quaternions are employed to represent color images, and quaternion DCT is used to calculate the saliency map. To model the influence of attention due to faces in an image, the resulting saliency map is combined with a face saliency map, calculated by using a modified census transform (MCT). The authors evaluated their model on a face images dataset from Ref. 85 and an image dataset from Ref. 65. The results show that for both datasets, their saliency model predicts eye fixations significantly better than the models in Refs. 35, 44, 48, 65.

Other Models

Rao et al.[73] introduced a model that uses a top-down search template matching approach to locate the salient regions. In their model, first, a saliency map is obtained from the input image by employing oriented spatiochromatic filters. After that, a template of the desired target object is moved across different regions of the saliency map, and the similarity between a selected region and the target is measured by calculating their Euclidean distance. Finally, the N most similar regions are represented as salient.

Kootstra et al.[86] proposed a model that calculates saliency on the basis of symmetry. In their model, three local symmetry operators, namely, isotropic symmetry,[87] radial symmetry[87] and color symmetry,[88] are defined. These three symmetry features are calculated at five image scales. The resulting saliency map is obtained by normalizing and combining the feature maps. For the evaluation of this model, the authors used a dataset containing 99 images belonging to different categories such as natural symmetries, animals, street scenes, buildings and natural environments. The authors showed that their symmetry model outperforms the saliency model in Ref. 44 in predicting the eye fixations.

Garcia-Diaz et al.[75] introduced a saliency model based on adaptive whitening of color image and feature maps. First, the input image is transformed from (r, g, b) to (z_1, z_2, z_3), a whitened representation. The whitening is done through decorrelation by employing principal component analysis. The feature maps are calculated for (z_1, z_2, z_3) using a bank of log-Gabor filters for orientations (0°, 45°, 90°, 135°), and seven scales are calculated for z_1 and only five for z_2 and z_3. Next, for each chromatic component the feature maps are whitened and contrast normalization is performed in several steps in a hierarchical manner. Saliency is computed as the square of the vector norm in the resulting representation. The authors showed that their model outperforms the state-of-the-art models in predicting fixations. These results were confirmed in an independent study by Borji et al.,[89] which concluded that the saliency model by Garcia-Diaz et al.[75] is the top performing model for natural images.

In this section, we briefly reviewed 34 different saliency algorithms in eight different categories. From the classification of saliency models we can see the following. First, a vast majority of the algorithms fall under the category of cognitive models. Second, some saliency models (such as Refs. 24, 32, 53) belong to several different categories. This can be explained by the fact that for calculating the initial saliency map, the saliency algorithms (e.g., Refs. 11, 35, 39, 47, 53) use similar features to that of the classic saliency model by Itti et al.[44] This initial saliency map is then modified by the various saliency models based on different criteria. For instance, in the saliency model by Torralba[47] (as shown in Fig. 2), the image locations of a saliency map (obtained by using traditional methods) are attenuated or amplified based on contextual priors to generate a new saliency map. This means that different saliency models share similar underlying concepts, which makes it difficult to classify a saliency model in one strict category. Third, while the first saliency models (e.g., Ref. 44) were bottom-up, the later models (e.g., Refs. 32, 36) propose adding top-down features such as faces, text and cars to the bottom-up model. Adding top-down features improves the performance of the bottom-up saliency models. As the performance is evaluated by how well the saliency algorithms predict where we look in real-world images, and real-world images typically comprise objects such as people, text, cars and mobile phones, adding these top-down features is seen as a natural step towards better prediction. However, this approach makes it challenging to analyze the performance of saliency algorithms from a purely bottom-up perspective.

EVALUATION OF SALIENCY MODELS

Image Database

For the analysis, we used the eye tracking database from the study by Judd et al.[32] As shown in Figure 6, the dataset consists of 1003 images selected randomly from different categories and different geographical locations. In the eye tracking experiment,[32] these images were shown to fifteen different users under free viewing conditions for a period of 3 s each. In the dataset, a majority of the images are 1024 pixels in width and 768 pixels in height. These landscape images were specifically used in the analysis.

Evaluation Metrics

In the literature, various metrics have been employed to measure the performance of saliency models. The performance is measured in terms of how well a bottom-up saliency model can predict where people look in images under free viewing conditions.

In this section, these metrics are briefly discussed.

Figure 6. Landscape images from the database by Judd et al.[32]

Pearson Correlation Coefficient

The Pearson correlation coefficient[80,90] is a measure of linear dependence between two variables. It is calculated as

$$r = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}},$$

where X and Y are the two variables, $\bar{X}$ and $\bar{Y}$ are the sample means, and r is the correlation coefficient. r returns a value in the range $[-1, 1]$. If r is 1 then it suggests a perfect prediction of the fixated regions by the saliency model, while a value of -1 implies that the predicted regions are the exact opposite of the fixations. A value of 0 suggests that there is no linear relation between the salient image regions and the fixated regions.

Eightieth Percentile Measure

To calculate the 80th percentile measure the saliency maps are thresholded to the top 20% of the salient image locations.[80,91] After that, the percentage of fixations falling inside these locations is calculated. In this way, this measure calculates the true positive rate of a classifier that uses the 80th percentile as the threshold for the saliency values.[80] This evaluation metric gives a scalar value in the range [0, 100].

Kullback–Leibler Divergence (D_{KL})

D_{KL}[80,92] is a measure of the logarithmic distance between two probability distributions. For evaluating saliency models, it is calculated as

$$D_{KL}(P \parallel Q) = \sum_{i} P(i) \ln\left(\frac{P(i)}{Q(i)}\right),$$

where P is the fixations probability distribution, i.e., the fixations map normalized in the interval [0, 1], and Q refers to the normalized saliency map. As D_{KL} is not a symmetric measure, i.e., $D_{KL} \neq D_{KL}$, a symmetric version of D_{KL} is calculated as

$$KL = D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P).$$

A KL value of zero indicates that the saliency model is perfect in predicting fixations. The KL metric does not have a well defined upper bound, thus its interval is $[0, \infty)$.

Earth Mover's Distance (EMD)

The earth mover's distance is a measure of similarity between two probability distributions or histograms.[93] In the literature, it is often described as comparing two piles of earth mass, and the minimum cost associated with turning one pile into the other pile, where cost is defined as the product of the amount of earth mass moved and the distance by which it is moved. According to Rubner et al.,[94] the earth mover's distance for two normalized histograms, P and Q, is calculated as $EMD(P, Q) = \min \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}}$, under the following constraints: $\sum_j f_{ij} \leq P_i$, $\sum_i f_{ij} \leq Q_j$, $\sum_{i,j} f_{ij} = \min(\sum_i P_i, \sum_j Q_j)$, $f_{i,j} \geq 0$, where f_{ij} denotes the amount of flow from bin i to j of the histograms, and d_{ij} is the ground distance between the two bins. An EMD value of zero indicates that the two distributions are the same, while a larger EMD value suggests that the two are quite dissimilar.

Normalized Scan-Path Saliency (NSS)

The NSS[80,95] is calculated by normalizing the saliency maps such that the maps have zero mean and unit standard deviation. From the resulting saliency maps, the mean of the saliency values for the fixated regions is used as a measure to judge the performance of the model. An NSS value ≥ 1 suggests that the correspondence between the saliency map and the fixations map is significantly better for the fixated locations than the non-fixated locations. If the NSS is less than or equal to zero then it implies that the prediction by the saliency model is not better than chance prediction. For detailed insight into the NSS metric, see the study by Peters et al.[95]

Chance Adjusted Salience

The chance adjusted salience[68,80] is calculated by the difference between the mean saliency values of two sets of image regions. The first set consists of parts that are fixated by an observer and the second consists of non-fixated parts. The non-fixated parts are selected from the fixations of the observer for an unrelated image. If the difference value obtained is greater than zero then it suggests that the saliency model is better than a random classifier. The range of this metric is governed by the interval of saliency values which can be arbitrary.

Ratio of Medians

To calculate the ratio of medians,[80,96] two sets of saliency values are selected. The first set consists of the saliency values of the fixated regions and second pertains to the saliency values of regions chosen from random points on the image. The saliency value for a fixation point is calculated as the maximum of the saliency values within a circular area of diameter 5.6° with the fixation point as the center. The saliency values for the random points are computed in the same manner as that of the fixation points. Next, for a given image the median of the saliency values for the fixated regions and the median of the saliency values for the randomly selected regions are calculated. The ratio of the two medians is used for the evaluation of the saliency model. A higher ratio implies that the prediction of fixations by the saliency model is better than the prediction by chance.

String Editing Distance

To calculate the string editing distance[46,97,98] for a given image, the fixations and the saliency values are clustered using methods such as *k*-means. After that, regions of interest (ROIs) are defined around these clusters which are labeled by alphabetic characters. Next, the ROIs are ordered based on the values assigned by the saliency model or the time sequence in which the ROIs were fixated on by the observer. The character strings obtained after ordering the ROIs for the saliency model and the fixations are then compared by using a string editing similarity index S_s, which is defined by the cost associated with performing operations such as deletion, insertion and substitution on the strings. An S_s value of zero implies that the saliency model perfectly predicts the fixated regions and their temporal sequence. For a detailed description of the string editing distance, see the study by Privitera & Stark.[98]

Area Under the Receiver-Operating-Characteristic Curve (AUC)

The AUC[46,99] is commonly employed in vision studies to evaluate the correspondence between fixated regions and salient image regions predicted by visual saliency models. For this, the fixations pertaining to a given image are averaged into a single two-dimensional map which is then convolved with a two-dimensional Gaussian filter. The resultant fixations map is then thresholded to yield a binary map with two classes—the positive class consisting of fixated regions and the negative class consisting of non-fixated regions. Next, from the two-dimensional saliency map, we obtain the saliency values associated with the positive and negative classes. Using the saliency values, a receiver-operating-characteristic (ROC) curve is drawn which plots the true positive rate against the false positive rate. The area under the ROC curve gives us a measure of the performance of the classifier. The AUC gives a scalar value in the interval [0, 1]. If the AUC is 1 then it indicates that the saliency model is perfect in predicting fixations. An AUC of 0.5 implies that the performance of the saliency model is not better than a random classifier or by chance prediction. For a detailed description of the AUC, see the study by Fawcett.[99]

Shuffled AUC

The shuffled AUC metric was proposed by Tatler et al.[43] and later used by Zhang et al.[48] to mitigate the effect of center bias in fixations. To calculate the shuffled AUC metric for a given image and one observer, the locations fixated by the observer are associated with the positive class in a manner similar to the regular AUC; however, the locations for the negative class are selected randomly from the fixated locations of other unrelated images, such that they do not coincide with the locations from the positive class.

Robust AUC

The robust AUC metric was proposed in a recent study by Alsam & Sharma.[100] This metric is derived from statistical analysis of eye fixation data with the aim of separating content driven fixations from content independent fixations (mainly defined as center bias). In order to calculate the robust metric, the negative class locations (non-fixated locations) are selected from the first eigenvector of the correlation matrix obtained from the fixations data of all images and observers. The authors state that the first principal component provides a reasonable basis for representing the content independent regions likely to be fixated upon; thus, choosing a non-fixated region from within it would indeed counteract the influence of center bias.

Criteria for Selecting an Evaluation Metric

The factors influencing the performance of saliency models include range, sample size, size of Gaussian, center bias and edge effect.

Range

As outlined in the study by Wilming et al.,[80] the range of a metric should be interpretable to make an objective judgment on its performance. It is evident from Table II that metrics such as the Pearson correlation coefficient, 80th percentile

Table II. Evaluation metrics and their attributes.

Metric	Robustness to center bias and edge effect	Sample size	Range
Pearson correlation coefficient	No	Large	[−1, 1]
Eightieth percentile measure	No	Large	[0, 100]
Kullback–Leibler divergence	No	Large	Arbitrary
Earth mover's distance	No	Large	Arbitrary
Normalized scan-path saliency	No	Large	Arbitrary
Chance adjusted salience	Yes	Small	Arbitrary
Ratio of medians	No	Large	Arbitrary
String editing distance	No	Large	Arbitrary
Area under the ROC curve (AUC)	No	Small	[0, 1]
Shuffled AUC	Yes	Small	[0, 1]
Robust AUC	Yes	Small	[0, 1]

measure, area under the ROC curve (AUC), shuffled AUC and robust AUC have a fixed range, which makes them more intuitive than the metrics that have an arbitrary scale such as Kullback–Leibler divergence, earth mover's distance, normalized scan-path saliency, chance adjusted salience, ratio of medians and string editing distance.

Sample Size

Sample size refers to the number of locations used to compare the correspondence between the fixated locations on a given image and the salient image locations for the associated image. While evaluation metrics such as Kullback–Leibler divergence, earth mover's distance, normalized scan-path saliency, Pearson correlation coefficient and 80th percentile measure use a large number of locations to calculate probability distributions, metrics such as chance adjusted salience, AUC, shuffled AUC and robust AUC need few locations.

Size of Gaussian

Studies[43,48] have shown that the performance of saliency algorithms is influenced by the size of the Gaussian used for smoothing the saliency maps. We know that different saliency models use different image scales to calculate salient image regions, resulting in noticeable disparities among saliency maps associated with a given image; this can be observed in Figure 7. To this end, researchers have suggested optimizing parameters such as σ (i.e., the standard deviation of the Gaussian distribution) and the size of the Gaussian for each saliency algorithm. For the analysis discussed in the fourth section, the size of the Gaussian for each saliency model was selected by optimizing for 50 test images (from the dataset by Judd et al.[32]) and using the ordinary AUC metric.

Center Bias

While viewing images, observers tend to look at the center regions more than peripheral regions. As a result of this, a majority of fixations fall at the image center. This effect is known as center bias and is well documented in vision studies.[41,43] The two main reasons for this are as follows. First, the tendency of photographers to place the objects at the center of the image. Second, the viewing strategy employed by observers, i.e., to look at center locations more in order to acquire the most information about a scene.[101] The presence of center bias in fixations makes it difficult to analyze the correspondence between the fixated regions and the salient image regions. This can be explained by the fact in a study by Judd et al.,[32] it was observed that a dummy classifier consisting of a two-dimensional Gaussian shape drawn at the center of the image outperformed all saliency models.

Edge effect

The center bias is implicitly linked with the so-called edge effect discussed by Zhang et al.[48] The edge effect[89] is defined as adding a varied image border of zeros to a saliency map, as a result of which it can yield different values from evaluation metrics. For example, in the study by Zhang et al.,[48] it was observed that a dummy saliency map consisting of all ones with a four-pixel image border consisting of zeros gave an AUC value of 0.62. Meanwhile, an AUC of 0.73 was obtained with a dummy saliency map using an eight-pixel border.

In the presence of center bias and the edge effect, a fair comparison of the performance of the saliency algorithms becomes a challenging task. To a certain extent this can be addressed by the following methods. First, weighting the salient regions at the center more than the peripheral regions, as done in the saliency models by Harel et al.[35] and Erdem and Erdem.[24] Second, explicitly adding a center Gaussian blob to the saliency map, as in the model by Judd et al.[32] Third, carefully selecting images with least center bias for the purpose of evaluation, as in the study by Borji et al.[89] However, these methods do not provide an optimal solution to this problem. As a result, the development of evaluation metrics that can compensate for the influences of center bias and the edge effect is seen as the next step towards addressing this issue. Recently, it has been observed that the influence of the center bias and the edge effect can be alleviated by using metrics such as chance adjusted salience, shuffled AUC by Tatler et al.[43] and robust AUC by Alsam & Sharma.[100] However, the range of chance adjusted salience is arbitrary, while the shuffled and robust AUC metrics have a more intuitive scale in the range [0, 1], which makes them quite suitable for evaluating the performance of saliency algorithms.

ANALYSIS

For analysis, ten latest state-of-the-art saliency models, namely, AIM by Bruce & Tsotsos,[65] AWS by Garcia-Diaz et al.,[75] Erdem by Erdem & Erdem,[24] Hou by Hou & Zhang,[69] Spec by Schauerte & Stiefelhagen,[72] GBA by Alsam et al.,[56,57] GBVS by Harel et al.,[35] Itti by Itti et al.,[44] Judd by Judd et al.[32] and LG by Borji & Itti[53] were used. In line with the study by Borji et al.,[89] two models were selected to provide a baseline for the evaluation. Gauss is defined as a two-dimensional

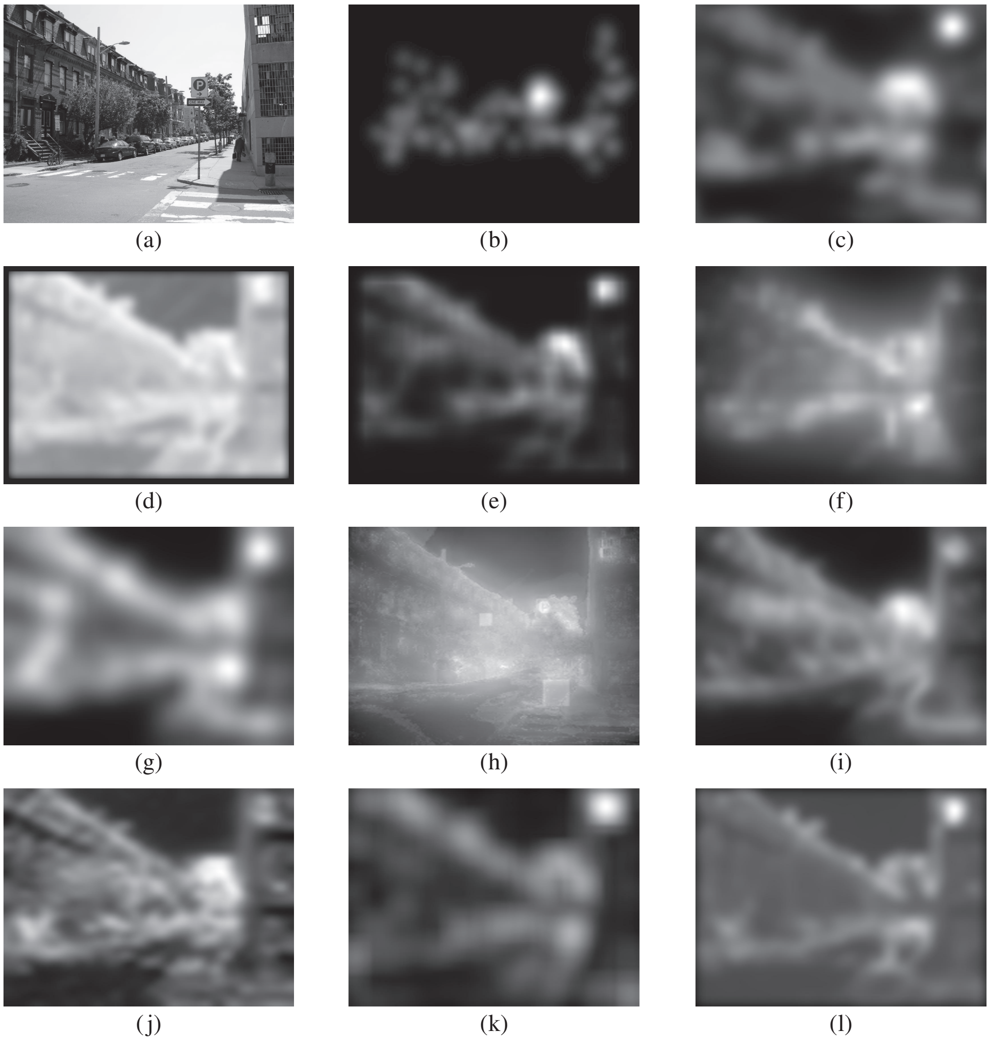


Figure 7. A test image, the associated fixations map (obtained from the fixations of 15 different observers) and the saliency maps from the different saliency algorithms used in the article: (a) image from database,[32] (b) fixations map, (c) AWS, (d) AIM, (e) Hou, (f) GBVS, (g) Itti, (h) Judd, (i) GBA, (j) LG, (k) Spec, (l) Erdem.

Gaussian blob at the center of the image. Different radii of the Gaussian blob are tested, and the radius that corresponds best with human eye fixations is selected.

This model corresponds well with the fixations falling at the image center. The IO model is based on the fact that an observer's fixations can be predicted best by the fixations of other observers viewing the same image. In this model, the map for an observer is calculated as follows. First, the fixations corresponding to a given image from all the observers except the one under consideration are averaged into a single two-dimensional map. Having done that, the fixations are spread by smoothing the map using a Gaussian filter. The IO model gives us an upper bound on the level of correspondence that is expected between the saliency models and the fixations. For comparing the performance of the different saliency algorithms, 10 evaluation metrics, namely, the Pearson correlation coefficient, normalized scan-path saliency, Kullback–Leibler divergence, earth mover's distance, 80th percentile measure, chance adjusted salience, ratio of medians, area under the ROC curve (AUC), shuffled AUC and robust AUC, were selected.

Metrics such as the Pearson correlation coefficient, earth mover's distance, 80th percentile measure and Kullback–Leibler divergence typically use the average of the probability distributions of fixations across different observers, while metrics such as the area under the ROC curve (AUC), shuffled AUC, robust AUC, ratio of medians and chance adjusted salience use the fixated and non-fixated locations for each observer—this means that the IO model can only be used for the latter metrics.

As the string editing distance metric is used to compare the order of fixations in time, it was not used for this examination. In the analysis, 463 landscape images of size 1024 by 768 pixels were used from the study by Judd et al.[32] Fig. 7 shows a test image, the associated fixations map (obtained from the fixations of 15 different observers) and the saliency maps from the different saliency algorithms.

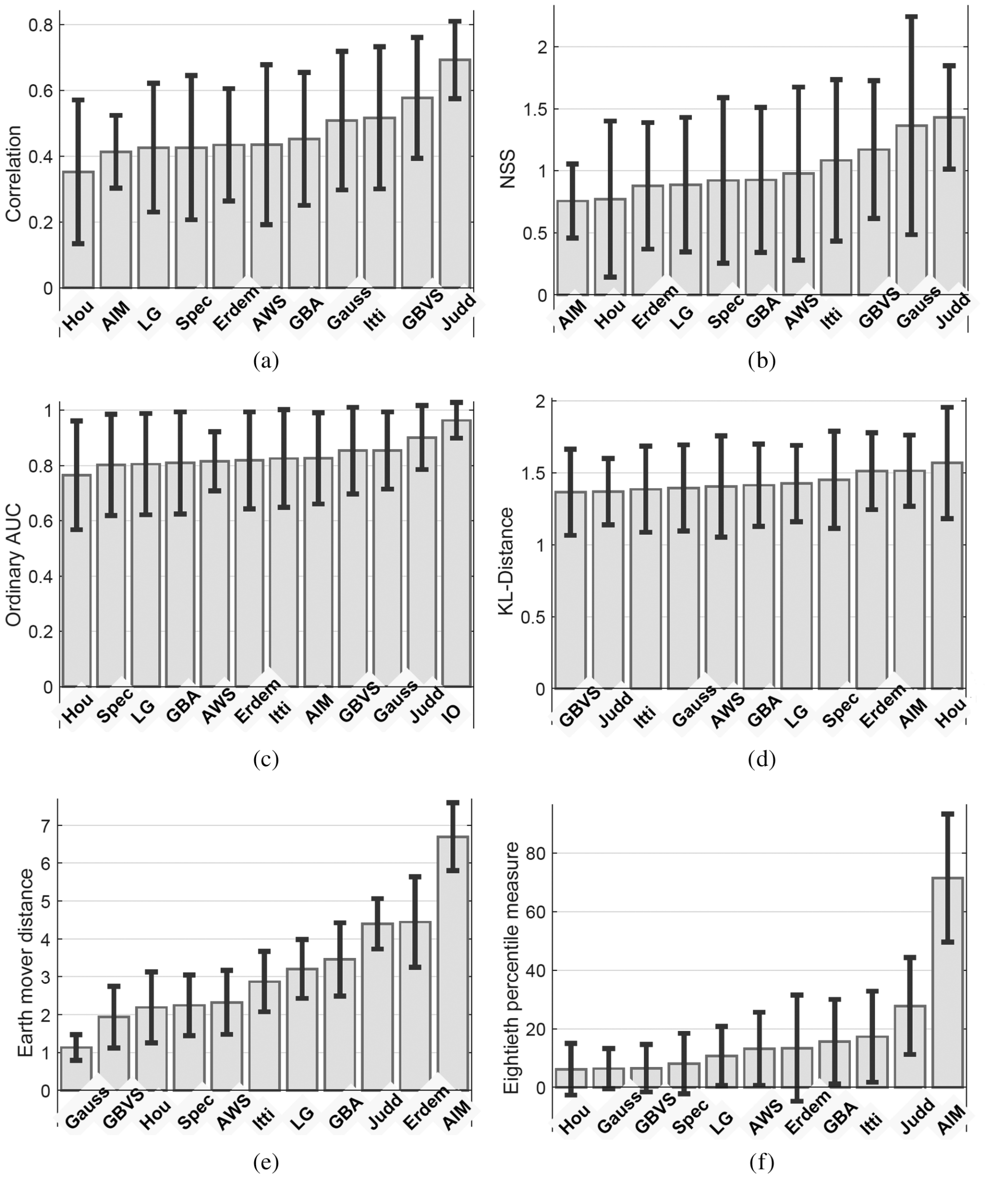


Figure 8. Ranking of visual saliency models using the Pearson correlation coefficient (correlation), normalized scan-path saliency (NSS), ordinary AUC, Kullback–Leibler divergence (KL-distance), earth mover's distance, and 80th percentile measure. The results are obtained from the fixations data of 463 landscape images and 15 observers.

The main objective of this analysis is to highlight how the ranking of saliency models is influenced by using different evaluation metrics.

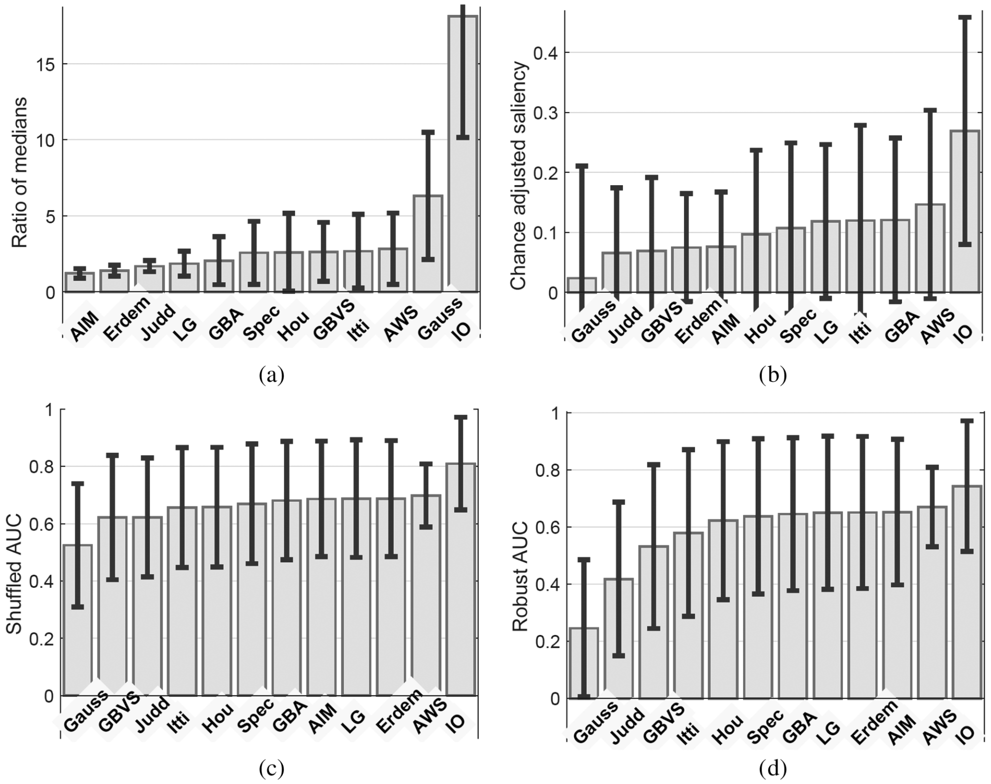
Results & Discussion

Comparison of Saliency Models With Human Eye Fixations

Figure 8(a) shows the ranking of saliency models obtained by using the Pearson correlation coefficient.

The vertical axis shows the average correlation coefficient for 463 test images along with the error bars which depict one standard deviation from the mean. We observe that Hou, AIM, LG, Spec, Erdem, AWS and GBA perform worse than the Gauss model, with GBVS and Judd being the two best models. This finding can be explained by the fact that the center regions are weighted more in the GBVS and Judd models. As a dummy classifier such as the Gauss model outperforms a majority of the saliency models used in this article, this indicates that the Pearson correlation coefficient metric is not able to counter the effects of fixations associated with center bias.

Next, the saliency algorithms are compared by using the normalized scan-path saliency metric. In line with the other metrics, the error bars represent one standard deviation from the mean. From the results in Fig. 8(b), it can be observed that the ranking obtained is similar to that of the correlation metric, with Gauss outperforming a majority of the saliency algorithms (including GBVS) and Judd being the best. A similar trend is observed when the evaluation is done by using the ordinary AUC metric (see Fig. 8(c)); in addition, it can be noted that all saliency models perform above chance.



(a)

(b)

(c)

(d)

Figure 9. Ranking of visual saliency models using the ratio of medians, chance adjusted salience, shuffled AUC and robust AUC metrics. The results are obtained from the fixations data of 463 landscape images and 15 observers.

In Fig. 8(d), we can see the ranking obtained by using the Kullback–Leibler divergence metric. In this metric, a higher value indicates greater differences between the saliency maps and the fixations maps. We note that the correspondence between the saliency and fixations maps is best for the GBVS algorithm, and is closely followed by the Judd model. In addition, we observe that the Hou model performs the worst and again the Gauss model performs better than a majority of the saliency models.

Next, the different algorithms are compared by using the earth mover's distance metric. In this case (as shown in Fig. 8(e)), we observe that the Gauss model corresponds best with the fixations maps, followed by the GBVS model. On the other hand, the AIM model performs the worst. These results suggest that the earth mover's distance metric is not able to reduce the effects associated with center bias.

Next, the saliency algorithms are compared by using the 80th percentile measure (shown in Fig. 8(f)). We can see that using this metric significantly improves the ranking of the AIM model. In this case, we observe that the Hou model is ranked the lowest. We can also note that the Gauss model performs poorly, but the Judd model ranks as the second best model. From the saliency maps in Fig. 7, we observe that the maps from algorithms such as AIM and Judd that are blurred more rank higher. This indicates that in its current form the 80th percentile measure metric cannot be used for the evaluation of saliency algorithms.

Figure 9(a) depicts the performance of the saliency models using the ratio of medians metric. This metric changes the ranking of some models significantly. For instance, it changes the ranking of the AIM model to the lowest and the AWS model to second best. However, again the Gauss model is the best (outranking all other saliency models), indicating the influence of center bias.

Fig. 9(b) shows the ranking of the saliency models obtained by using the chance adjusted salience. It is evident that using chance adjusted salience changes the ranking of the saliency models significantly. The Gauss model changes from being one of the best to the worst. In addition, the models with center bias (such as Judd and GBVS) are ranked low. In this case, models such as Itti, GBA and AWS perform the best. The results suggest that this metric can be used to alleviate the influences of the center bias and the edge effect. Fig. 9(b) also shows that the (one) standard deviations of the mean values exceed the mean values themselves–this along with the fact that the range of the chance adjusted salience metric is arbitrary makes it unsuitable for the evaluation of saliency models.

Finally, we compare the ranking of the saliency models using the shuffled AUC metric as depicted in Fig. 9(c) and the robust AUC as shown in Fig. 9(d). We note that in both of these metrics the Gauss model is ranked the worst and the AWS model is ranked the best. In the case of the robust AUC metric, the AUC value for the Gauss model is lower than that obtained by using the shuffled AUC metric. This suggests

Table III. The AUC metric matrix.

	AIM	Erdem	GBA	Hou	GBVS	Itti	Judd	LG	Spec	AWS
AIM	1.0	1.0	0.9	0.9	0.8	0.9	0.8	0.9	0.9	0.9
Erdem	0.9	1.0	0.9	0.9	0.8	0.9	0.8	0.9	0.9	0.9
GBA	0.9	0.9	1.0	0.9	0.8	0.9	0.8	0.9	0.9	0.9
Hou	0.8	0.9	0.9	1.0	0.8	0.9	0.8	0.9	1.0	0.9
GBVS	0.8	0.8	0.8	0.8	1.0	0.9	0.9	0.8	0.8	0.8
Itti	0.9	0.9	0.9	0.9	0.8	1.0	0.8	0.9	0.9	0.9
Judd	0.8	0.8	0.8	0.8	0.9	0.9	1.0	0.8	0.8	0.8
LG	0.9	0.9	0.9	0.9	0.8	0.9	0.8	1.0	0.9	0.9
Spec	0.9	0.9	0.9	1.0	0.8	0.9	0.8	0.9	1.0	0.9
AWS	0.9	0.9	0.9	0.9	0.8	0.9	0.8	0.9	0.9	1.0

that the robust AUC metric is better at countering the effects associated with the center bias. Furthermore, we also note that the correspondence of the IO model is lower in the case of the shuffled AUC and robust AUC metrics as compared with the ratio of medians and the chance adjusted salience metrics. This indicates that the saliency algorithms are not far from human performance represented by the IO model. The results show that a majority of the state-of-the-art saliency models such as Spec, GBA, LG, Erdem, AIM and AWS are quite close to each other in terms of their performance (in both the shuffled AUC and robust AUC metrics). This raises the question of whether the different saliency algorithms are similar to one another.

Inter-Comparison of Different Saliency Models

In order to measure the similarity of the different saliency models to each other, the models were compared using the ordinary AUC metric for 463 landscape images (the same as before). In Table III, the first row and first column represent the different saliency models and the rest of the rows and columns show their associated AUC values. For instance, the diagonal elements of the table show correspondence of a saliency model with itself. Based on the results, we can see that the maps generated by different saliency algorithms are quite similar to one another.

CONCLUSION

In this article, the state of the art of saliency algorithms for still images is discussed. As the number of saliency models has increased significantly in the past two decades, we are faced with the challenge of finding a metric that can be used to objectively compare the performance of different saliency algorithms. To understand this, first, we study the important factors that influence the comparison of saliency algorithms with human eye fixation data. From the discussion in the section on criteria for selecting an evaluation metric, we note that the center bias and the edge effect are the two main factors. Next, the performance of ten different saliency algorithms is analyzed by using ten different metrics. The results suggest that the evaluation metrics such as the shuffled AUC proposed by Tatler et al.[43] and the robust AUC by Alsam & Sharma[100] are better suited to mitigating the influences of the center bias and the edge effect. The shuffled and robust AUC metrics can be calculated for human eye fixations associated with static images—the aim is to extend this analysis to videos in the future.

Based on the results obtained from the shuffled and robust AUC metrics, we note the following. First, the saliency model by Garcia-Diaz et al.[75] outperforms all the other saliency algorithms discussed in this article. Second, we observe that the state-of-the-art saliency models (such as those in Refs. 24, 53, 57, 65, 72) are statistically close to each other in their correspondence with human eye fixations. This is further supported by the results in the section on inter-comparison of different saliency models, which indicate that the saliency maps from different models are quite similar to each other. Third, the results suggest that the saliency models are not far from achieving the upper limit of performance—represented by the inter-observer model. This indicates that the saliency algorithms can account well for the bottom-up factors associated with visual attention; however, further studies are needed to study and develop models that can account for individual differences between different observers.

Typically, visual saliency algorithms are evaluated by comparing the saliency maps with fixations maps—which are obtained by showing an image to an observer for a fixed duration (usually 3 s[32]). The saliency map for a real-world scene comprises a number of salient locations; this number exceeds the number of fixations in nearly all cases. In other words, the saliency maps have more basis vectors than human fixations data from a typical eye tracking experiment.[100,102] This can be addressed by showing the images to the observers for longer durations of time and increasing the number of observers.

Although it is well known that visual attention is a combination of different mechanisms including top-down, bottom-up and spatial bias (towards human faces and body parts), recent attempts in Refs. 42, 100 at separating the fixations data into content driven and content independent fixations is seen as the next step towards improving the robustness of evaluation metrics.

REFERENCES

[1] R. A. Rensink, J. K. O'Regan, and J. J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," Psychol. Sci. **8**, 368–373 (1997).

[2] D. J. Simons and D. T. Levin, "Failure to detect changes to people during a real-world interaction," Psychon. Bull. Rev. **5**, 644–649 (1998).

[3] J. K. O'Regan, R. A. Rensink, and J. J. Clark, "Change-blindness as a result of 'mudsplashes'," Nature **398**, 34 (1999).

[4] C. H. Anderson, D. C. Van Essen, and B. A. Olshausen, "Directed visual attention and the dynamic control of information flow," *Neurobiology of Attention* (Elsevier, 2005), pp. 11–17.

[5] J. Braun and D. Sagi, "Vision outside the focus of attention," Percept. Psychophys. **48**, 45–58 (1990).

[6] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," Annu. Rev. Neurosci. **18**, 193–222 (1995).

[7] S. B. Steinman and B. A. Steinman, "Vision and attention I: Current models of visual attention," Optom. Vis. Sci. **75**, 146–155 (1998).

[8] M. C. Mozer and M. Sitton, *Computational Modeling of Spatial Attention* (Psychology Press, 1998), pp. 341–393, chapter 9.

[9] K. Suder and F. Worgotter, "The control of low-level information flow in the visual system," Rev. Neurosci. **11**, 127–146 (2000).

[10] L. Itti and C. Koch, "Computational modelling of visual attention," Nature Rev. Neurosci. **2**, 194–203 (2001).

[11] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. **2**, 2049–2056 (2006).

[12] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," Hum. Neurobiol. **4**, 219–227 (1985).

[13] M. M. Chun and J. M. Wolfe, "Visual attention," *Blackwell Handbook of Sensation and Perception*, edited by E. B. Goldstein (Blackwell Publishing Ltd, Oxford, UK, 2001), pp. 272–310.

[14] J. M. Wolfe, S. J. Butcher, C. Lee, and M. Hyle, "Changing your mind: On the contributions of top-down and bottom-up guidance in visual search for feature singletons," J. Exp. Psychol.: Hum. Percept. Perform. **29**, 483–502 (2003).

[15] H. Jasso and J. Triesch, "Learning to attend—from bottom-up to top-down," *Top-Down Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, edited by L. Paletta and E. Rome (Springer, Berlin, Heidelberg, 2008), pp. 106–122.

[16] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," Perception **28**, 1311–1328 (1999).

[17] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," Exp. Brain Res. **139**, 266–277 (2001).

[18] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," Vis. Res. **40**, 1489–1506 (2000).

[19] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," IEEE Trans. Image Process. **13**, 1304–1318 (2004).

[20] S. X. Yu and D. A. Lisin, "Image compression based on visual saliency at individual scales," *Advances in Visual Computing*, edited by G. Bebis, R. Boyle, B. Parvin, D. Koracin, Y. Kuno, J. Wang, J.-X. Wang, J. Wang, R. Pajarola, P. Lindstrom, A. Hinkenjann, M. Encarnação, C. T. Silva and D. Coming, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2009), Vol. 5875, pp. 157–166.

[21] A. Alsam, H. Rivertz, and P. Sharma, "What the eye did not see—a fusion approach to image coding," *Advances in Visual Computing*, edited by G. Bebis, R. Boyle, B. Parvin, D. Koracin, C. Fowlkes, S. Wang, M.-H. Choi, S. Mantler, J. Schulze, D. Acevedo, K. Mueller and M. Papka, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2012), Vol. 7432, pp. 199–208.

[22] A. Alsam, H. J. Rivertz, and P. Sharma, "What the eye did not see—a fusion approach to image coding," Int. J. Artif. Intell. Tools **22**, 13 (2013).

[23] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," *Proc. 6th Int'l Conf. on Computer Vision Systems* (Springer, Berlin, 2008), pp. 66–75.

[24] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," J. Vis. **13**, 1–20 (2013).

[25] C. Siagian and L. Itti, "Biologically-inspired robotics vision Monte-Carlo localization in the outdoor environment," *IEEE Int'l Conf. on Intelligent Robots and Systems (IROS'07)* (IEEE, Piscataway, NJ, 2007).

[26] S. Frintrop, P. Jensfelt, and H. I. Christensen, "Attentional landmark selection for visual slam," *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS'06)* (IEEE, Piscataway, NJ, 2006).

[27] T. Kadir and M. Brady, "Saliency, scale and image description," Int. J. Comput. Vis. **45**, 83–105 (2001).

[28] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency based objective quality assessment of decoded video affected by packet losses," *15th IEEE Int'l Conf. on Image Processing (ICIP 2008)* (IEEE, Piscataway, NJ, 2008), pp. 2560–2563.

[29] Q. Ma and L. Zhang, "Saliency-based image quality assessment criterion," *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, edited by D.-S. Huang, D. C. Wunsch II, D. S. Levine and K.-H. Jo, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2008), Vol. 5226, pp. 1124–1133.

[30] M. S. El-Nasr, A. Vasilakos, C. Rao, and J. Zupko, "Dynamic intelligent lighting for directing visual attention in interactive 3-d scenes," IEEE Trans. Comput. Intell. AI in Games **1**, 145–153 (2009).

[31] R. Rosenholtz, A. Dorai, and R. Freeman, "Do predictions of visual perception aid design?," ACM Trans. Appl. Percept. **8**, 12 (2011).

[32] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *Proc. 2009 IEEE Int'l Conf. on Computer Vision (ICCV)* (IEEE, Piscataway, NJ, 2009), pp. 2106–2113.

[33] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," *IJCAI* (1999), pp. 1146–1153.

[34] M. Ajallooeian, A. Borji, B. N. Araabi, M. N. Ahmadabadi, and H. Moradi, "An application to interactive robotic marionette playing based on saliency maps," *The 18th IEEE Int'l. Symposium on Robot and Human Interactive Communication, 2009. RO-MAN 2009* (IEEE, Piscataway, NJ, 2009), pp. 841–847.

[35] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Proc. Neural Information Processing Systems (NIPS)* (MIT Press, 2006), pp. 545–552.

[36] M. Cerf, J. Harel, W. Einhauser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *Advances in Neural Information Processing Systems (NIPS)* (MIT Press, 2007), Vol. 20, pp. 241–248.

[37] J. M. Henderson, J. R. Brockmole, M. S. Castelhano, and M. Mack, "Visual saliency does not account for eye movements during visual search in real-world scenes," *Eye Movements: A Window on Mind and Brain* (Elsevier, 2007), pp. 537–562.

[38] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," Vis. Res. **42**, 107–123 (2002).

[39] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," *ICIP 2003: Proc. 2003 Int'l Conf. on Image Processing, 2003* (IEEE, Piscataway, NJ, 2003), Vol. 1, pp. 253–256.

[40] J. M. Henderson, "Human gaze control during real-world scene perception," Trends Cogn. Sci. **7**, 498–504 (2003).

[41] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," J. Vis. **7**, 1–17 (2007).

[42] A. Alsam and P. Sharma, "Analysis of eye fixations data," *Proc. IASTED Int'l Conf., Signal and Image Processing (SIP 2011)* (IASTED, ACTA Press, Dallas, USA, 2011), pp. 342–349.

[43] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," Vis. Res. **45**, 643–659 (2005).

[44] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1254–1259 (1998).

[45] R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," Vis. Res. **39**, 3157–3163 (1999).

[46] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," IEEE Trans. Pattern Anal. Mach. Intell. **35**, 185–207 (2013).

[47] A. Torralba, "Modeling global scene factors in attention," J. Opt. Soc. Am. A **20**, 1407–1418 (2003).

[48] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," J. Vis. **8**, 1–20 (2008).

[49] I. P. Howard, *Seeing in Depth: Volume 1 Basic Mechanisms* (I Porteous, Toronto, 2002).

[50] D. Walther, "Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics," Ph.D. thesis (California Institute of Technology, Pasadena, California, 2006).

[51] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition—a gentle way," *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02)*, Lecture Notes in Computer Science (Springer, Berlin, 2002), Vol. 2525, pp. 472–479.

[52] S. Frintrop, "VOCUS: A visual attention system for object detection and goal-directed search," Ph.D. thesis (University of Bonn, 2006).

[53] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2012), pp. 1–8.

[54] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," IEEE Trans. Pattern Anal. Mach. Intell. **28**, 802–817 (2006).

55 U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "Gaffe: A gaze-attentive fixation finding engine," IEEE Trans. Image Process. **17**, 564–573 (2008).

56 A. Alsam, P. Sharma, and A. Wrålsen, "Asymmetry as a measure of visual saliency," *SCIA 2013, Lecture Notes in Computer Science (LNCS)* (Springer, Berlin, Heidelberg, 2013), Vol. 7944, pp. 591–600.

57 A. Alsam, P. Sharma, and A. Wrålsen, "Calculating saliency using the dihedral group d4," J. Imaging Sci. Technol. **58**, 010504 (2014).

58 D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," *Proc. NIPS* (MIT Press, 2004), pp. 481–488.

59 D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," IEEE Trans. Pattern Anal. Mach. Intell. **31**, 989–1005 (2009).

60 Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, "Visual saliency based on conditional entropy," *Computer Vision—ACCV 2009*, edited by H. Zha, R.-I. Taniguchi and S. Maybank, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2010), Vol. 5994, pp. 246–257.

61 W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating human saccadic scanpaths on natural images," *2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2011), pp. 441–448.

62 T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," IEEE Trans. Pattern Anal. Mach. Intell. **32**, 693–708 (2010).

63 S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and where: A Bayesian inference theory of attention," Vis. Res. **50**, 2233–2247 (2010).

64 T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," IEEE Trans. Pattern Anal. Mach. Intell. **33**, 353–367 (2011).

65 N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," *Proc. Neural Information Processing Systems Conf. (NIPS 2005)* (MIT Press, 2005), pp. 155–162.

66 M. Mancas, "Computational attention: towards attentive computers," Ph.D. thesis (Faculté Polytechnique de Mons—FPMs, 2007).

67 H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," J. Vis. **9**, 1–27 (2009).

68 W. Kienzle, M. O. Franz, B. Schölkopf, and F. A. Wichmann, "Center-surround patterns emerge as optimal predictors for human saccade targets," J. Vis. **9**, 1–15 (2009).

69 X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," *IEEE Conf. on Computer Vision and Pattern Recognition, 2007 (CVPR'07)* (IEEE, Piscataway, NJ, 2007), pp. 1–8.

70 C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," *2008 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2008), pp. 1–8.

71 P. Bian and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency," *Advances in Neuro-Information Processing*, edited by M. Köppen, N. Kasabov and G. Coghill, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2009), Vol. 5506, pp. 251–258.

72 B. Schauerte and R. Stiefelhagen, "Predicting human gaze using quaternion dct image signature saliency and face detection," *Proc. IEEE Workshop on the Applications of Computer Vision (WACV)* (IEEE, Breckenridge, CO, 2012).

73 R. P. N. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard, "Eye movements in iconic visual search," Vis. Res. **42**, 1447–1463 (2002).

74 S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2010), pp. 2376–2383.

75 A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," Image Vis. Comput. **30**, 51–64 (2012).

76 A. Treisman and G. Gelade, "A feature-integration theory of attention," Cogn. Psychol. **12**, 97–136 (1980).

77 P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proc. 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR), Vol. 1* (IEEE, Piscataway, NJ, 2001), vol. 1, p. I.

78 N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," *Proc. 2011 IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'11* (IEEE Computer Society, Washington, DC, 2011), pp. 433–440.

79 O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," *Computer Vision ECCV 2006*, edited by A. Leonardis, H. Bischof and A. Pinz, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2006), Vol. 3952, pp. 589–600.

80 N. Wilming, T. Betz, T. C. Kietzmann, and P. Konig, "Measures and limits of models of fixation selection," PLoS ONE **6**, 1–19 (2011).

81 H. Barlow, "Cerebral cortex as model builder," *Matters of Intelligence*, edited by L. M. Vaina, Synthese Library (Springer, Netherlands, 1987), Vol. 188, pp. 395–406.

82 T. S. Lee and S. Yu, "An information-theoretic framework for understanding saccadic behaviors," *Advance in Neural Information Processing Systems*, edited by K.-R. Muller, S. A. Solla and T. K. Leen (MIT Press, 2000), Vol 12.

83 L. W. Renninger, J. Coughlan, P. Verghese, and J. Malik, "An information maximization model of eye movements," Adv. Neural Inform. Process. Sys. **17**, 1121–1128 (2005).

84 A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," Int. J. Comput. Vis. **42**, 145–175 (2001).

85 M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," J. Vis. **9**, 1–15 (2009).

86 G. Kootstra, B. de Boer, and L. R. B. Schomaker, "Predicting eye fixations on complex visual stimuli using local symmetry," Cogn. Comput. **3**, 223–240 (2011).

87 D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context free attentional operators: The generalized symmetry transform," Int. J. Comput. Vis. **14**, 119–130 (1995).

88 G. Heidemann, "Focus-of-attention from local color symmetries," IEEE Trans. Pattern Anal. Mach. Intell. **26**, 817–830 (2004).

89 A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," IEEE Trans. Image Process. **22**, 55–69 (2013).

90 A. D. Hwang, E. C. Higgins, and M. Pomplun, "A model of top-down attentional control during visual search in complex scenes," J. Vis. **9**, (2009).

91 A. Torralba, M. S. Castelhano, A. Oliva, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," Psychol. Rev. **113**, 1–23 (2006).

92 L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," Vis. Res. **49**, 1295–1306 (2009) (Top cited article 2008–2010 award from Vision Research).

93 O. Pele and M. Werman, "A linear time histogram metric for improved SIFT matching," *Computer Vision—ECCV 2008*, edited by D. Forsyth, P. Torr and A. Zisserman, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2008), Vol. 5304, pp. 495–508.

94 Y. Rubner, C. Tomasi, and L. J. Guibas., "The earth mover's distance as a metric for image retrieval," Int. J. Comput. Vis. **40**, 99–121 (2000).

95 R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," Vis. Res. **45**, 2397–2416 (2005).

96 N. Parikh, L. Itti, and J. Weiland, "Saliency-based image processing for retinal prostheses," J. Neural Eng. **7**, 016006 (2010).

97 S. A. Brandt and L. W. Stark, "Spontaneous eye movements during visual imagery reflect the content of the visual scene," J. Cogn. Neurosci. **9**, 27–38 (1997).

98 C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 970–982 (2000).

99 T. Fawcett, "ROC graphs with instance-varying costs," Pattern Recognit. Lett. **27**, 882–891 (2004).

100 A. Alsam and P. Sharma, "Robust metric for the evaluation of visual saliency algorithms," J. Opt. Soc. Am. A **31**, 1–9 (2014).

101 P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," J. Vis. **9**, 1–16 (2009).

102 P. Sharma and A. Alsam, "A robust metric for the evaluation of visual saliency models," *Proc. 9th Int'l Conf. on Computer Vision Theory and Applications*, edited by J. Braz and S. Battiato (SciTePress, 2014), pp. 654–661.