International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Visual Saliency with Statistical Priors

Jia Li · Yonghong Tian · Tiejun Huang

Received: date / Accepted: date

Abstract Visual saliency is a useful cue to locate the conspicuous image content. To estimate saliency, many approaches have been proposed to detect the unique or rare visual stimuli. However, such bottom-up solutions are often insufficient since the prior knowledge, which often indicates a biased selectivity on the input stimuli, is not taken into account. To solve this problem, this paper presents a novel approach to estimate image saliency by learning the prior knowledge. In our approach, the influences of the visual stimuli and the prior knowledge are jointly incorporated into a Bayesian framework. In this framework, the bottom-up saliency is calculated to pop-out the visual subsets that are probably salient, while the prior knowledge is used to recover the wrongly suppressed targets and inhibit the improperly popped-out distractors. Compared with existing approaches, the prior knowledge used in our approach, including the foreground prior and the correlation prior, is statistically learned from 9.6 million images in an unsupervised manner. Experimental results on two public benchmarks show that such statistical priors are effective to modulate the bottom-up saliency to achieve impressive improvements when compared with 10 state-of-the-art methods.

Keywords Visual saliency · Prior knowledge · Image statistics · Bayesian framework

J. Li, Y. Tian and T. Huang
National Engineering Laboratory for Video Technology, Peking University, China.
Y. Tian is the the corresponding author
Tel.: +86-10-62753817
Fax: +86-10-62751638
E-mail: yhtian@pku.edu.cn

1 Introduction

Visual saliency estimation, which aims to detect the important content in images and videos, has become a popular research topic in recent years. In most cases, the salient stimuli have the capability to easily capture human visual attention and thus become interesting (Elazary and Itti 2008). By focusing on the salient content in images and videos, applications such as video retargeting, content-based advertising and image/video retrieval can generate results that can better meet human perception.

In existing studies on visual saliency estimation, rarity is a frequently-used criterion to quantify saliency. Usually, the unique or rare visual subsets are supposed to be salient. For example, Itti et al (1998) proposed a classical framework to estimate visual saliency by calculating the center-surround contrasts. Visual signals might become salient only if they could differ from their neighbors in multiple scales. Harel et al (2006) represented an image as a graph and adopted a random walker to detect the salient signals that were related to the less visited nodes. In (Riche et al 2012), saliency was estimated by detecting locally contrasted and globally rare features. Generally, these approaches can generate promising results but may have a severe problem since saliency is **not equivalent to** rarity. Although it is often safe to assume that salient signals are rare, the opposite assumption will not always hold since some background distractors may also become rare, either locally or globally (as shown in Fig. 1(a)-(b), some distractors are also rare and will be popped-out in the competition). Moreover, existing computing methodologies on rarity still have some drawbacks (e.g., computing local contrasts in improper scales) and some salient targets may be wrongly suppressed (as shown in Fig. 1(c)-(d), only the borders of the large salient targets can pop-out while their inner smooth parts are wrongly suppressed). Therefore, one of the most important problem in visual saliency estimation is to **recover the wrongly suppressed targets and inhibit the improperly popped-out distractors**.

To solve this problem, incorporating the prior knowledge could probably be a feasible solution. Actually, the prior knowledge can bias the competition between the input visual signals by favoring a specific category of visual stimuli (Frith 2005). In this process, such biased selectivity can help to pop-out the real targets and suppress the real distractors. For example, Cerf et al (2008) assigned high saliency values to human faces, while Meur et al (2006) proposed that visual stimuli around image centers should be emphasized. Beyond these predefined priors, some supervised approaches such as (Li et al 2010; Zhao and Koch 2012) tried to learn the feature prior, while (Torralba et al 2006; Chikkerur et al 2010) proposed to learn the task-dependent prior that can be integrated into the Bayesian framework to pop-out the objects corresponding to specific search or recognition tasks. However, the priors learned from a limited number of images usually have the over-fitting risk, while the task related priors prevent their further usage in generic scenarios. To sum up, effective and robust prior knowledge is inevitable for visual saliency estimation.

Following this idea, we propose a novel Bayesian approach for image saliency estimation by jointly capturing the influences of the input visual stimuli and the prior knowledge unsupervisedly learned from millions of images. In particular, we focus on estimating saliency under free-viewing conditions and the learned prior knowledge is task-independent. As shown in Fig. 2, our approach first calculates the bottom-up saliency by only considering the unbiased competitions between visual signals. After that, the bottom-up saliency is modulated by the prior knowledge statistically learned from millions of images. In particular, the foreground prior is learned by inferring the spatial distributions of all kinds of image patches, and can be used to identify whether an image patch belongs to foreground. The correlation prior is learned by mining the patch co-occurrence characteristics, which can be used to model the mutual influence between different image patches. These two priors are then used to bias the competition between visual signals by recovering the wrongly suppressed targets and inhibiting the improperly popped-out distractors. Finally, the estimated saliency maps can be improved by simultaneously using the cues from the visual signal and the prior knowledge.

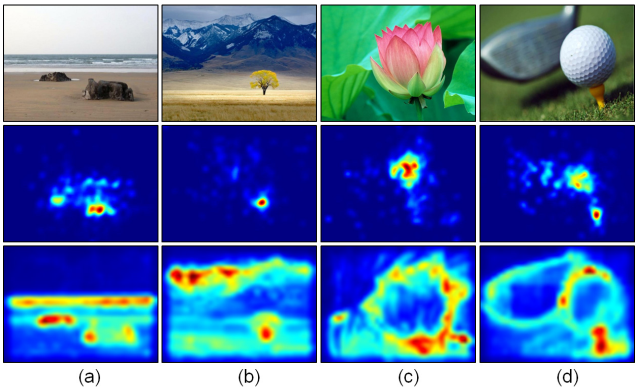


Fig. 1 Examples of wrongly suppressed targets and improperly popped-out distractors. Images in the first row are selected from the benchmark proposed by Li et al (2013). The second row presents the fixation density maps from 21 subjects and the third row illustrates the saliency maps calculated using the classical model proposed by Itti et al (1998). (a)-(b) background distractors may be also rare and will be popped-out in the unbiased competition; (c)-(d) only the borders of the large salient targets can pop-out while their inner smooth parts are wrongly suppressed.

In the experiments, we compare our approach with 10 state-of-the-art approaches, including (Itti et al 1998; Bruce and Tsotsos 2006; Harel et al 2006; Hou and Zhang 2007, 2008; Zhang et al 2008; Achanta et al 2009; Wang et al 2010; Goferman et al 2010; Riche et al 2012). Experimental results on two public image benchmarks show that the learned statistical priors can effectively modulate the bottom-up saliency to better predict human fixations. Consequently, our approach achieves impressive improvements and demonstrates several advantages in utilizing the prior knowledge. Our main contributions are summarized as follows:

1. Two kinds of prior knowledge, including the foreground prior and the correlation prior, are presented for estimating saliency in the free-viewing scenario. By modeling both the spatial distributions and correlations of various visual stimuli, such priors can well adapt to various scenes in recovering the wrongly suppressed targets and inhibiting the improperly popped-out distractors.
2. We propose an effective learning algorithm to learn the prior knowledge from millions of images in an unsupervised manner. Such prior knowledge, which is learned from huge amounts of images, is statistically significant and avoids the over-fitting risk.
3. A Bayesian framework is proposed to jointly capture the influences of the visual stimuli and the prior knowledge for visual saliency estimation. Experimental results show that this framework is effective to modulate any kinds of bottom-up saliency to better predict human fixations.

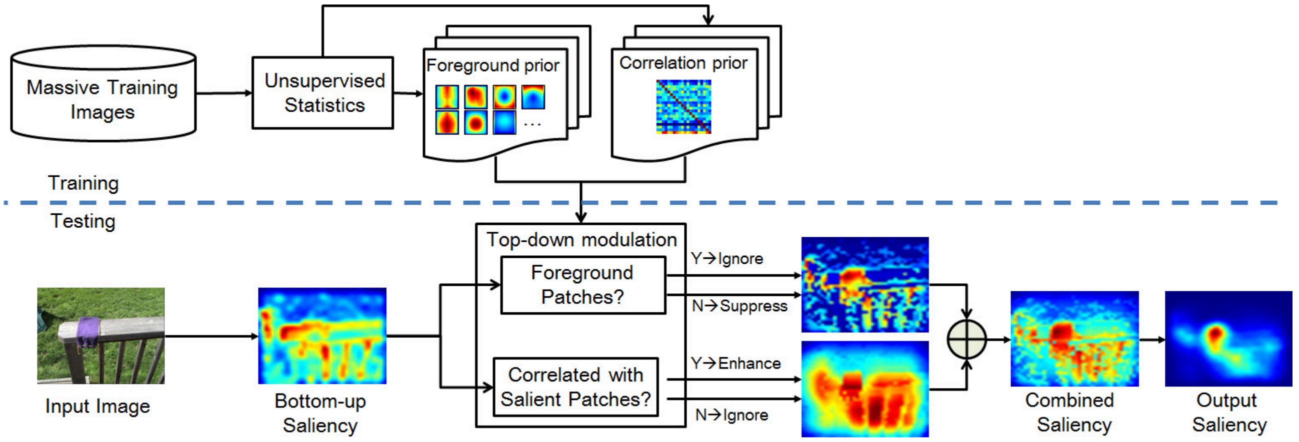


Fig. 2 The system framework of our approach. In this framework, the bottom-up saliency is first calculated using any existing stimulus-driven saliency model. After that, the top-down component will modulate the bottom-up saliency with the priors that are statistically learned from massive unlabeled images. In this process, the foreground prior is learned to evaluate whether a patch belongs to foreground, and background patches will be suppressed. Meanwhile, the correlation prior is learned to reveal the correlations between patches. Using the correlation prior, patches which have strong latent correlations with bottom-up salient patches are selectively enhanced, while irrelevant patches will be ignored. Finally, the saliency maps obtained from foreground and correlation priors are combined to generate the final saliency map.

The rest of this paper is organized as follows: Section 2 reviews the related work and Section 3 states the problems that should be solved. In Section 4, we describe the details of our approach in visual saliency estimation. Experimental results are presented in Section 5 and the paper is concluded in Section 6.

2 Related Work

In the past decades, many approaches have been proposed to estimate image/video saliency, segment salient objects, explore the neurobiological evidences, etc. In this survey, we will mainly focus on the computational approaches on *image* saliency estimation.

2.1 The Bottom-up Approaches

In the bottom-up approaches, visual signals will compete fairly to pop-out. Inspired by this idea, existing bottom-up approaches often aim to detect the unique or rare visual subsets which are supposed to be the winner in the unbiased competition. Therefore, the main difference between these approaches lies in the way to quantify such visual rarity.

The most popular way to quantify rarity is to calculate the difference between various visual subsets. For example, Itti et al (1998) proposed a classical framework in which high saliency values were assigned to the visual subsets with high center-surround contrasts. Riche et al (2012) assumed that locally contrasted and globally rare features were salient and adopted a sequential framework to extract various features and estimate visual saliency. In (Achanta et al 2009), image saliency was determined by the difference between Gaussian blurred features and mean image features, while Vikram et al (2012) estimated image saliency by integrating the local differences over random rectangular regions. Generally, the difference-based approach can pop-out most of the rare targets (i.e., the recall can be relatively high). However, the simple difference calculation often fails to recognize the distractors which may be also rare (i.e., the precision may be low).

Instead of calculating the difference between various visual stimuli, some approaches adopted intuitive definitions on rarity. For instance, Lu et al (2011) proposed that regions on the convex side of curved boundaries were probably salient and detected salient targets by using such concavity contexts. In (Harel et al 2006), an image was first represented by a weighted graph and a random walker was then adopted to pop-out the visual subsets corresponding to the less visited nodes. Hou and Zhang (2007) proposed an approach to estimate visual saliency by calculating the spectral residuals using the Fourier transform. Visual irregularities were first detected in the transform domain, which were then transformed back to the spatial domain to locate the salient targets.

Generally speaking, the bottom-up approaches can work well in many cases. As mentioned above, however, visual rarity is not equivalent to visual saliency. Some background distractors may be wrongly popped-out and some foreground targets may be wrongly suppressed when using improper features, scales and computing methods to quantify visual rarity. If the input visual signals are fairly treated without any bias, these fake targets (i.e., the wrongly popped-out distractors) and false distractors (i.e., the wrongly suppressed targets) will be inevitable in the estimated saliency maps.

2.2 The Knowledge-based Approaches

To solve the problems in existing bottom-up approaches, incorporating the prior knowledge into visual saliency estimation could probably be a feasible solution. Actually, the prior knowledge can often bias the competition between visual signals by favoring a specific category of visual stimuli (Frith 2005). According to the ways that the prior knowledge is obtained, existing knowledge-based approaches can be divided into three groups: ad-hoc group, learning-based group and statistical group.

The approaches in the ad-hoc group aim to bias the competition between various visual signals by using the *predefined* prior knowledge. For instance, Cerf et al (2008) and Goferman et al (2010) assumed that faces were inherently salient. In their approaches, human faces were detected with the face detection algorithms and high saliency values were directly assigned to the related visual signals. By observing that subjects often stared at the center of the scene to start the eye-tracking experiments, Meur et al (2006) adopted such center-bias as the predefined prior to enhance patches near to image centers with an anisotropic Gaussian. Some approaches such as (Cheng et al 2011; Aziz and Mertsching 2008; Liu et al 2007a) proposed that saliency values should be assigned to objects, instead of spatial locations. The latent assumption was that the input visual signals were inherently correlated and such signals should be treated as a whole in the competition. Consequently, the manually fine-tuned parameters were used as the prior knowledge to group different visual signals into objects (or super-pixels). Often, these approaches can work well on simple images. However, they may have difficulties to process images with rich contents since it is usually difficult to obtain the required cues from the complex scenes (e.g., segment all the objects with one set of predefined parameters, detecting the side faces, etc.).

Instead of using the predefined priors, the approaches in the learning-based group aim to *learn* the prior knowledge in a *supervised* manner. They often try to learn the optimal "stimuli-saliency" mapping models, which can emphasize the effective feature channels (e.g., with high weights) and inhibit the useless feature channels (e.g., with low weights). For instance, Kienzle et al (2007) adopted a Support Vector Machine (SVM) to model the correlations between high-dimensional features and visual saliency values, while Judd et al (2009) also utilized the SVM with linear kernels to optimize the mapping from low-level, mid-level and high-level features to visual saliency. Similarly, Zhao and Koch (2011) and Li et al (2010) proposed to model such "stimuli-saliency" mapping by using linear functions which were optimized by least square algorithm or quadratic programming. In (Liu et al 2007b), several novel features were proposed and the Conditional Random Field (CRF) was adopted to combine these features for salient object detection. In (Navalpakkam and Itti 2007), the weights of various visual feature channels were optimized by maximizing the signal-noise-ratio. Instead of optimizing the weights for various feature channels, Peters and Itti (2007) tried to learn a direct mapping from the global feature matrix to the fixation density map. Zhao and Koch (2012) first established a feature pool with 88 features and the AdaBoost algorithm was then adopted to train a set of weak classifiers by iteratively training weak classifier, estimating classifier weight and updating sample weights. These weak classifiers were then combined to build the saliency model. Beyond the models that mainly focus on estimating saliency in free-viewing conditions, Torralba et al (2006) proposed a Bayesian approach to estimate task-dependent saliency. In their approach, the global scene context served as a cue to reveal the probable locations to search specific targets (e.g., searching painting, mug and people). By learning the relationship between global features and target locations, the bottom-up saliency can be modulated to adapt to various search tasks. Similarly, Chikkerur et al (2010) learned both the feature and location priors about specific object categories. These priors were then integrated with the bottom-up factors using a Bayesian inference framework to pop-out the objects corresponding to specific search or recognition tasks.

Generally speaking, these learning-based approaches can demonstrate promising performance on small benchmarks. The parameters trained and fine-tuned on part of the benchmark can usually achieve high performance on the rest of the benchmark. However, the most severe drawback of these approaches is that they require the supervised learning process. In this process, all the training data should be labeled with eye tracking devices or manual labeling activities, which is really time-consuming. Consequently, existing benchmarks are usually very small (with only hundreds or thousands of images), which is far from sufficient to cover all possible cases. Therefore, the trained models often have the over-fitting risk. For instance, the model trained on a limited number of images can bias to specific feature channels and locations to generate promising results on similar testing scenes, but such model may fail when encountering unknown scenes. That also hampers the further usage of these learning-based models in actual applications.

To avoid the over-fitting risk, the approaches in the statistical group try to *learn* the prior knowledge from massive images in an *unsupervised* manner. For these approaches, a common process is to train a set of visual words (or namely the independent components, basis functions, sparse codes, dictionaries, etc.) from massive image statistics (Bruce and Tsotsos 2006; Hou and Zhang 2008; Zhang et al 2008; Wang et al 2010; Borji and Itti 2012; Yang and Yang 2012; Sun et al 2012). Image patches are then projected to these visual words to get more compact visual representations. By representing image patches with the projection coefficients, Bruce and Tsotsos (2006) estimated visual saliency by maximizing the information sampled from a scene, while Hou and Zhang (2008) proposed the Incremental Coding Length (ICL), which was used as the criterion to redistribute the limited energy (saliency) amongst features. Borji and Itti (2012) proposed an approach to estimate visual saliency by using the projection coefficients to quantify the local center-surround difference and the global rarity. In (Wang et al 2010), a set of sub-band feature maps were first extracted using the learned sparse codes. These feature maps were then represented as fully-connected graphs, on which random walkers were used and visual saliency was defined by the average information transmitted during the random walk. Different from these approaches, Yang and Yang (2012) proposed a novel algorithm for learning the visual words. In their approach, the visual words were treated as the latent variables of CRF. By jointly learning the CRF and the dictionary, the overall performance was greatly improved and the estimated saliency maps were much clear.

To sum up, the latent assumption in these statistical approaches is that foreground targets and background distractors are more distinguishable in the new subspace formed by the learned visual words. These visual words, which are often learned unsupervisedly from thousands of images, can be viewed as some kinds of prior knowledge. However, these approaches still have the same problem as the bottom-up approaches since the visual subsets are also equally treated in the new subspace and no bias is applied. By projecting the image patches onto the new subspace, the problem of wrongly suppressing targets and improperly popping-out distractors can be mitigated but remains unsolved. Moreover, these statistical approaches also failed to consider the influence of the latent correlations between various visual stimuli. Actually, inherently correlated stimuli can usually excite each other to become salient, while irrelevant stimuli may compete to inhibit each other. Without modeling the prior knowledge on such latent correlations, it is often difficult to perfectly process the scenes with massive objects. Therefore, it is necessary to unsupervisedly learn the *biased* prior knowledge such as the foreground prior and the correlation prior to adaptively process different visual stimuli when considering their mutual correlations.

3 Problem Statement

To estimate visual saliency, one of the most important problem is to simultaneously model the influences of the visual stimuli and the prior knowledge. In human vision system, various visual stimuli will compete to become salient, while the prior knowledge may bias the competition in two ways: recovering the foreground targets that are wrongly suppressed and inhibiting the background distractors that are improperly popped-out. Following this idea, we propose a Bayesian framework to jointly capture the influences of the visual stimuli and the prior knowledge. In this framework, we focus on modulating the visual saliency acquired through bottom-up competition with various top-down priors. Let $\mathbf{s}_n$ be the event that an image patch $\mathcal{B}_n$ (e.g., 8×8 macro blocks) pops-out after the bottom-up competition and $\mathbf{r}_n$ be the event that $\mathcal{B}_n$ becomes salient after the top-down modulation, we can assume that:

$$\begin{aligned} P(\mathbf{r}_n) &= \sum_{m=1}^{M} P(\mathbf{s}_m)P(\mathbf{r}_n|\mathbf{s}_m) \\ &= P(\mathbf{s}_n)P(\mathbf{r}_n|\mathbf{s}_n) + \sum_{m\neq n}^{M} P(\mathbf{s}_m)P(\mathbf{r}_n|\mathbf{s}_m). \end{aligned} \tag{1}$$

where M is the total number of patches in the same image. From (1), we can see that the problem of estimating $P(\mathbf{r}_n)$ can be divided into three sub-problems:

1. Estimating $P(\mathbf{s}_n)$, which is the probability that $\mathcal{B}_n$ pops-out in the bottom-up competition between the input visual stimuli;
2. Estimating $P(\mathbf{r}_n|\mathbf{s}_n)$, which is the probability that $\mathcal{B}_n$ becomes salient after the top-down modulation if it already pops-out in the bottom-up competition (as shown in Fig. 3(a)); and
3. Estimating $P(\mathbf{r}_n|\mathbf{s}_m)$, which is the probability that $\mathcal{B}_n$ becomes salient after the top-down modulation if another patch $\mathcal{B}_m$ pops-out in the bottom-up competition (as shown in Fig. 3(b)).

Among all these three sub-problems, the first one has been well studied in the past decades and there already exist many feasible solutions to this sub-problem

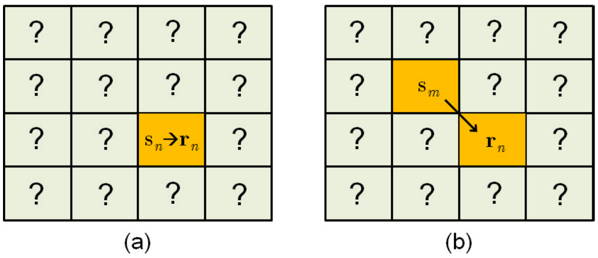


Fig. 3 Problem statement. Note that $\mathbf{s}_n$ is the event that $\mathcal{B}_n$ pops-out after the bottom-up competition and $\mathbf{r}_n$ is the event that $\mathcal{B}_n$ becomes salient after the top-down modulation. Here we incorporate the top-down priors to solve two problems: (a) whether $\mathcal{B}_n$ can become salient after the top-down modulation when only considering its bottom-up saliency; (b) whether $\mathcal{B}_n$ can become salient after the top-down modulation when only considering the bottom-up saliency of $\mathcal{B}_m$.

(e.g., Itti et al (1998); Bruce and Tsotsos (2006); Harel et al (2006); Hou and Zhang (2007); Parikh et al (2008)). Here we denote the estimated bottom-up saliency as $S_b(n)$ for a patch $\mathcal{B}_n$ and thus can assume:

$$P(\mathbf{s}_n) \propto S_b(n). \tag{2}$$

In the following study, we will mainly focus on the last two sub-problems and the main difficulty is to modulate bottom-up saliency using various kinds of prior knowledge. In this process, only the appearances and locations of the patches are available. Therefore, **the prior knowledge related to visual attributes and positional information** could probably be an effective key to solve the proposed two sub-problems.

4 Saliency with Statistical Priors

In this section, we will address the proposed two sub-problems and modulate the bottom-up saliency with the learned top-down priors. First, we investigate what kinds of prior knowledge should be learned. After that, we describe the details on how to learn the required prior knowledge through massive image statistics. Finally, we present how to estimate visual saliency using the learned priors.

4.1 What to Learn?

Generally speaking, there are numerous kinds of prior knowledge and it is impossible to learn all of them. According to the problems stated in Fig. 3, we have to learn the prior knowledge that demonstrates a biased selectivity on the visual attributes and positional information, which are the only cues in conducting the top-down modulation. Therefore, we will mainly focus on two kinds of prior knowledge, including the **foreground prior** and the **correlation prior**. The foreground prior aims to identify whether an image patch belongs to foreground using its visual attributes and positional information. This prior knowledge can be helpful to estimate $P(\mathbf{r}_n|\mathbf{s}_n)$. The correlation prior aims to model the mutual correlations between image patches. This prior knowledge can be helpful to estimate $P(\mathbf{r}_n|\mathbf{s}_m)$ by taking the correlation between image patches into account. With these two kinds of prior knowledge, bottom-up saliency can be modulated to recover the wrongly suppressed targets and inhibit the improperly popped-out distractors.

In order to learn such prior knowledge and ensure the learned priors are statistically significant, we collect 9.6 million images that are randomly crawled from Flicker. Each image is resized to have a max side length of no more than 320 pixels while keeping the width-to-height ratio. Each image is then divided into a set of non-overlapping 8×8 patches, and each patch is characterized by its preattentive features (Wolfe 2005). That is, we represent a pixel v in an image patch $\mathcal{B}_n$ by its intensity I_v, red-green opponency RG_v, blue-yellow opponency BY_v and four orientation features $\{O_v^\theta\}, \theta \in \{0°, 45°, 90°, 135°\}$. The intensity and color opponencies can be calculated as:

$$\begin{aligned} I_v &= \frac{r_v + g_v + b_v}{3}, \\ RG_v &= \frac{r_v - g_v}{\max(r_v, g_v, b_v)}, \\ BY_v &= \frac{b_v - \min(r_v, g_v)}{\max(r_v, g_v, b_v)}, \end{aligned} \tag{3}$$

where r_v, g_v, b_v are the red, green and blue components for pixel v, respectively. Here the red-green and blue-yellow opponencies are calculated as in (Walther and Koch 2006), which will be set to zero if $\max(r_v, g_v, b_v) < 0.1$ to avoid large fluctuations at low luminance.

The orientation feature O_v^θ can be calculated by convolving I_v with Gabor filters:

$$O_v^\theta = \parallel I_v * G_0(\theta) \parallel + \parallel I_v * G_{\pi/2}(\theta) \parallel , \tag{4}$$

where $G_0(\theta)$ and $G_{\pi/2}(\theta)$ are two Gabor filters oriented at θ with phase 0 and $\pi/2$, respectively. After calculating these features for all pixels in $\mathcal{B}_n$, we further quantize each feature into 4 bins to acquire 7 histograms from an image patch $\mathcal{B}_n$. Finally, each patch is characterized by a feature vector with 7×4=28 components with the same dynamic range of [0,1].

From 9.6 million images, we can obtain billions of image patches (i.e., billions of 28d feature vectors). To efficiently learn prior knowledge from such a huge number of image patches, we have to further reduce the feature dimension to obtain a more compact patch representation. Toward this end, a feasible solution could be generating a set of visual words by performing k-means clustering on all the image patches and represent each patch with the nearest visual word. However, it is very difficult to directly perform the k-means clustering on billions of image patches due to the limitation of computational resource. Therefore, we use the affinity propagation algorithm proposed in (Frey and Dueck 2007) to select a set of representative patches (i.e., exemplars) from each image. In this process, the number of exemplars is automatically determined according to the complexity of image content. Since such exemplars are much fewer than the original patches, we can perform k-means clustering on their feature vectors to form a vocabulary of N_w visual words, denoted as $\{\mathbf{w}_i\}_{i=1}^{N_w}$ (in the experiments, we will show the influence of N_w). With these visual words, each patch can be quantized to the nearest visual word using the Euclidian distance measure. Finally, we can represent a patch $\mathcal{B}_n$ with only one integer label $l_n \in \{1, ..., N_w\}$.

4.2 Learning the Foreground Prior

To calculate the foreground prior, we simply count the times that a visual word $\mathbf{w}_i$ appears at any probable locations. After that, we can get N_w distribution maps $\{\mathcal{D}_i\}_{i=1}^{N_w}$, which are then normalized to let the location with the highest frequency corresponds to 1. Some representative visual words and their distribution maps are shown in Fig. 4.

From Fig. 4, we can see that many visual words demonstrate center-biased distributions, while several visual words distribute around image edges. Since people often snap photos by intentionally placing the targets near to image centers (i.e., the photographer bias in Tseng et al (2009)), foreground targets often appear around image centers while background distractors usually appear near to image edges (as shown in Fig. 5). Therefore, we can safely assume that **visual words have higher probabilities to appear in the foreground than in the background if they distribute around image centers**.

Following this assumption, we can use the distribution maps as the foreground criterion. As shown in Fig. 6(a), we divide a distribution map $\mathcal{D}_i$ into two regions with equivalent area and use Ω_i to quantify its center-bias property as the percentage of energy in the "center" region. From Fig. 6(b), we can see that the quantified center-bias properties are high on many distribution maps, while some maps are obviously edge-biased. Thus the corresponding patches can be treated as distractors and should be suppressed.

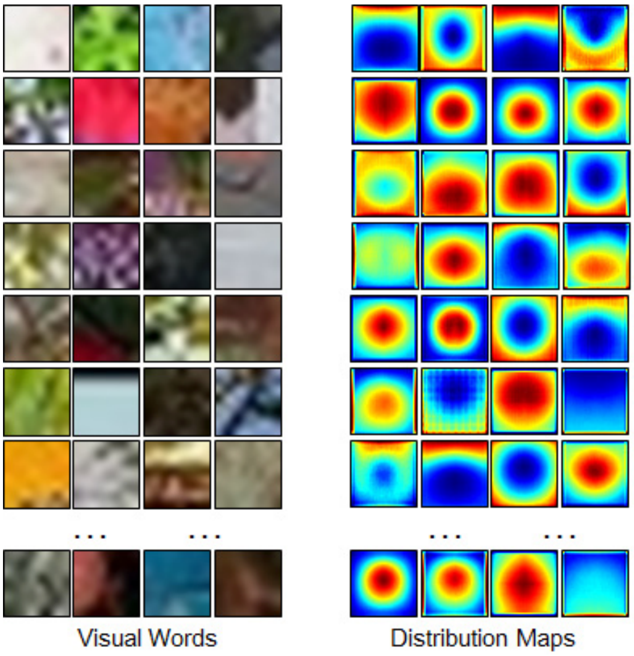


Fig. 4 Some representative visual words and their distribution maps learned from millions of images. Many maps demonstrate the strong center-bias property, while several maps are obviously edge-biased. These maps may contain cues to evaluate whether a patch belongs to foreground and thus can be used as the prior knowledge in visual saliency estimation.

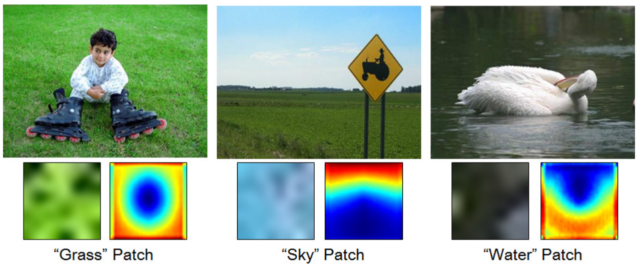


Fig. 5 The distribution maps of typical background patches. Salient targets are often intentionally placed near to image centers by photographers while background patches often distribute around the edges. These three visual words and distribution maps correspond to the last three samples from the first row in Fig. 4.

Given the quantified center-bias property of each visual word, we can infer the probability that an image patch belongs to foreground using its visual attributes and positional information. Suppose that $\mathcal{B}_n$ is classified to the visual word $\mathbf{w}_{l_n}$ and let $\mathbf{f}_n$ be the event that $\mathcal{B}_n$ is a foreground patch, we can estimate $P(\mathbf{f}_n)$ as:

$$P(\mathbf{f}_n) \propto [\Omega_{l_n} \geq 0.5]_{\mathbf{I}} \cdot \mathcal{D}_{l_n}(n), \tag{5}$$

where $[\Omega_{l_n} \geq 0.5]_{\mathbf{I}}$ equals to 1 if $\Omega_{l_n} \geq 0.5$ and 0 otherwise. $\mathcal{D}_{l_n}(n)$ indicates the frequency that the visual

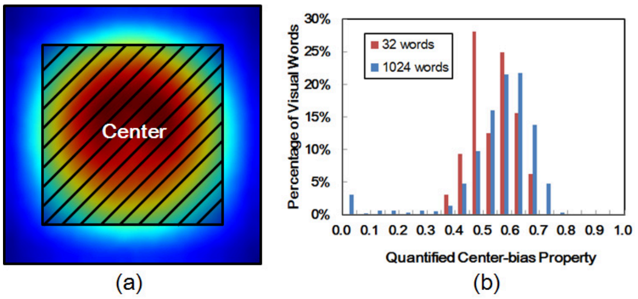


Fig. 6 Quantified Center-biased Property. (a) a distribution map is divided into two regions with equivalent area and the center-bias property Ω_i is quantified as the percentage of energy in the "center" area. (b) the histogram of the quantified center-bias properties $\{\Omega_i\}_{i=1}^{N_w}$ of the distribution maps. Without loss of generality, we use 32 and 1024 visual words.

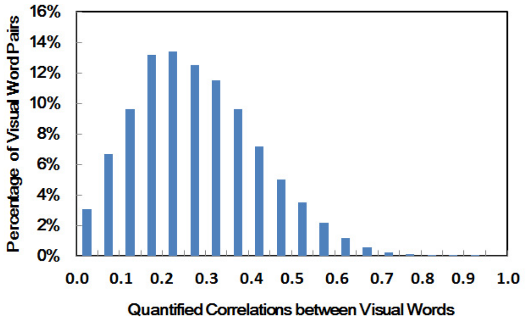


Fig. 7 The histogram of quantified correlations $\{\Upsilon_{mn}\}$ between N_w=1024 visual words. We can see that most visual words demonstrate weak correlations, while some visual words can demonstrate strong co-occurrence properties even on millions of images.

word $\mathbf{w}_{l_n}$ appears at the location of $\mathcal{B}_n$. From (5), we can see that the probability that $\mathcal{B}_n$ belongs to foreground could become high if: 1) $\mathbf{w}_{l_n}$ distributes around image centers (i.e., $\Omega_{l_n} \geq 0.5$) and thus has a higher probability to appear in the foreground than in the background; and 2) $\mathcal{B}_n$ appears at recurring locations that a probable foreground visual word $\mathbf{w}_{l_n}$ can be frequently observed in millions of images.

4.3 Learning the Correlation Prior

The objective of learning the correlation prior is to model the mutual influence between any two patches. Generally speaking, there exist two kinds of typical mutual influences: 1) if two image patches $\mathcal{B}_m$ and $\mathcal{B}_n$ are correlated, they will probably excite each other. Once we observe one patch, we may expect the other one; 2) on the contrary, irrelevant image patches will compete to inhibit each other. Here we use Υ_{mn} to quantify the correlation strength between two visual words $\mathbf{w}_m$ and $\mathbf{w}_n$. In calculating Υ_{mn}, two visual words that frequently co-occur in the same images may be tightly correlated. That is, if we can observe one visual word, we can expect another one with a high probability. Moreover, the probability of expecting $\mathbf{w}_n$ when $\mathbf{w}_m$ is observed should be different from that of expecting $\mathbf{w}_m$ when $\mathbf{w}_n$ is observed. For example, many images may contain a common visual word $\mathbf{w}_n$. When some other visual words in these images are observed, $\mathbf{w}_n$ can be expected with high probabilities. However, we can hardly expect other specific visual words when $\mathbf{w}_n$ is observed.

Following this idea, we can estimate Υ_{mn} using massive image statistics. First, we count the frequency F_n which indicates the total times that the visual word $\mathbf{w}_n$ appears in all the training images. Meanwhile, we also count the frequency F_{mn} which represents the total times that two visual words $\mathbf{w}_m$ and $\mathbf{w}_n$ appear in the same images (note that F_{nn}=F_n). After that, we can calculate Υ_{mn} as:

$$\Upsilon_{mn} = \frac{F_{mn}}{F_n}. \tag{6}$$

From (6), we can see that Υ_{mn} is unequal to Υ_{nm} and a higher co-occurrence frequency will lead to a stronger correlation. Fig. 7 shows the histogram of such quantified correlation strength between 1024 visual words. We can see that most visual words have weak correlations, while some visual words demonstrate strong co-occurrence properties even in millions of images.

Given the quantified correlation strength between visual words, we can infer the probability that one image patch is tightly correlated with another patch using their visual attributes and positional information. Suppose that $\mathcal{B}_m$ and $\mathcal{B}_n$ are classified to visual words $\mathbf{w}_{l_m}$ and $\mathbf{w}_{l_n}$ and let $\mathbf{o}_{mn}$ be the event that $\mathcal{B}_n$ is correlated to $\mathcal{B}_m$, we can estimate $P(\mathbf{o}_{mn})$ as:

$$P(\mathbf{o}_{mn}) \propto \Upsilon_{l_m l_n} \cdot \mathcal{N}(d_{mn}; 0, \sigma_c), \tag{7}$$

where $\mathcal{N}$ is the Gaussian distribution and σ_c is empirically set to 0.3 in this study. d_{mn} is the distance between $\mathcal{B}_m$ and $\mathcal{B}_n$, which is normalized by the distance from image corner to image center. The Gaussian term is important to ensure that only the correlations between the patches in a local area are considered to increase the computational efficiency. From (7), we can see that the probability that $\mathcal{B}_n$ is correlated to $\mathcal{B}_m$ will be high if: 1) $\mathbf{w}_{l_m}$ and $\mathbf{w}_{l_n}$ frequently co-occur in millions of images; and 2) $\mathcal{B}_m$ and $\mathcal{B}_n$ are near to each other.

4.4 Visual Saliency with Statistical Priors

Given the learned foreground prior and correlation prior, we can now turn to the two sub-problems proposed above: how to estimate $P(\mathbf{r}_n|\mathbf{s}_n)$ and $P(\mathbf{r}_n|\mathbf{s}_m)$?

To estimate $P(\mathbf{r}_n|\mathbf{s}_n)$, we have to first infer the foreground prior $P(\mathbf{f}_n)$ using (5) to see whether $\mathcal{B}_n$ is a foreground patch. With the foreground prior, we can rewrite $P(\mathbf{r}_n|\mathbf{s}_n)$ as:

$$P(\mathbf{r}_n|\mathbf{s}_n) = P(\mathbf{f}_n)P(\mathbf{r}_n|\mathbf{s}_n, \mathbf{f}_n) + P(\bar{\mathbf{f}}_n)P(\mathbf{r}_n|\mathbf{s}_n, \bar{\mathbf{f}}_n). \quad (8)$$

From (8), we can see that there are two probable combinations of events $\mathbf{s}_n$ and $\mathbf{f}_n$, including:

– $\mathbf{s}_n$ and $\mathbf{f}_n$: $\mathcal{B}_n$ is a target that is correctly popped-out by the bottom-up model.
– $\mathbf{s}_n$ and $\bar{\mathbf{f}}_n$: $\mathcal{B}_n$ is probably a distractor that is improperly popped-out by the bottom-up model.

When modulating the bottom-up saliency with the foreground prior, we can maintain the correctly popped-out targets and suppress the improperly popped-out distractors by setting:

$$P(\mathbf{r}_n|\mathbf{s}_n, \mathbf{f}_n) \approx 1, \; P(\mathbf{r}_n|\mathbf{s}_n, \bar{\mathbf{f}}_n) = e^{-\alpha_b}. \quad (9)$$

where $\alpha_b{\geq}0$ is a predefined constant to fuse the conflict predictions made by the bottom-up saliency model and the foreground prior. Smaller $e^{-\alpha_b}$ indicates the foreground prior is more reliable (we will show the influence of α_b in the experiment). By incorporating (9) into (8), we can estimate $P(\mathbf{r}_n|\mathbf{s}_n)$ as:

$$P(\mathbf{r}_n|\mathbf{s}_n) = e^{-\alpha_b} + (1 - e^{-\alpha_b})P(\mathbf{f}_n), \quad (10)$$

where the foreground prior $P(\mathbf{f}_n)$ can be estimated using (5). From (10), we can see that the bottom-up saliency can be selectively modulated by the foreground prior. In this process, the real targets, which are predicted as salient by both the bottom-up saliency model and foreground prior, will become salient. On the contrary, the distractors, which pop-out in the bottom-up competition, will be suppressed if the foreground prior classifies them as distractors.

To estimate $P(\mathbf{r}_n|\mathbf{s}_m)$, we have to first infer the correlation prior $P(\mathbf{o}_{mn})$ to find whether $\mathcal{B}_n$ is tightly correlated with $\mathcal{B}_m$. Inspired by this idea, we have:

$$\begin{aligned} P(\mathbf{r}_n|\mathbf{s}_m) =& P(\mathbf{o}_{mn})P(\mathbf{r}_n|\mathbf{s}_m, \mathbf{o}_{mn}) \\ &+ P(\bar{\mathbf{o}}_{mn})P(\mathbf{r}_n|\mathbf{s}_m, \bar{\mathbf{o}}_{mn}). \end{aligned} \quad (11)$$

From (11), we can also find two probable combinations of events $\mathbf{s}_n$ and $\mathbf{o}_{mn}$, including:

– $\mathbf{s}_m$ and $\mathbf{o}_{mn}$: $\mathcal{B}_n$ is tightly correlated with a patch that pops-out in the bottom-up competition. In this case, $\mathcal{B}_n$ will be excited by $\mathcal{B}_m$.
– $\mathbf{s}_m$ and $\bar{\mathbf{o}}_{mn}$: $\mathcal{B}_n$ is irrelevant with a patch that pops-out in the bottom-up competition. In this case, $\mathcal{B}_n$ will be inhibited by $\mathcal{B}_m$.

When modulating the bottom-up saliency with the correlation prior, two tightly correlated patches will excite each other while irrelevant patches will inhibit each other by setting:

$$P(\mathbf{r}_n|\mathbf{s}_m, \mathbf{o}_{mn}) \approx 1, \; P(\mathbf{r}_n|\mathbf{s}_m, \bar{\mathbf{o}}_{mn}) \approx 0. \quad (12)$$

By incorporating (12) into (11), we can thus estimate $P(\mathbf{r}_n|\mathbf{s}_m)$ as:

$$P(\mathbf{r}_n|\mathbf{s}_m) = P(\mathbf{o}_{mn}), \quad (13)$$

where the correlation prior $P(\mathbf{o}_{mn})$ can be estimated using (7). From (13), we can see that the wrongly suppressed targets may become salient after the top-down modulation if it is tightly correlated with the targets that pop-out in the bottom-up competition. As shown in Fig. 8, a patch that pops-out in the bottom-up competition (e.g., the head and tail of the cow) can then selectively enhance nearby patches with strong correlations (e.g., the body of the cow) in the following top-down modulation. In this manner, we can pop-out the salient target as a whole, especially for those objects with large smooth regions.

After estimating $P(\mathbf{r}_n|\mathbf{s}_n)$ and $P(\mathbf{r}_n|\mathbf{s}_m)$, the saliency value of $\mathcal{B}_n$ after the top-down modulation, denoted as $S_r(n) \propto P(\mathbf{r}_n)$, can thus be calculated by incorporating (2), (10) and (13) into (1):

$$\begin{aligned} S_r(n) \propto\, & S_b(n) \cdot (e^{-\alpha_b} + (1 - e^{-\alpha_b})P(\mathbf{f}_n)) \\ & + \sum_{m \neq n}^{M} S_b(m) \cdot P(\mathbf{o}_{mn}), \end{aligned} \quad (14)$$

where $P(\mathbf{f}_n)$ and $P(\mathbf{o}_{mn})$ are foreground and correlation priors that can be derived using (5) and (7), respectively. To facilitate the computation of $S_r(n)$, we first obtain two saliency maps using the first and the second terms in (14), respectively. These two saliency maps are then normalized into the same range of [0,1] and fused with equal weights.

Given a testing image, its saliency map can be easily estimated through four major steps (as shown in Fig. 2):

1. Resize the image to have a max side length of no more than 320 pixels and divided the image into non-overlapping 8×8 patches;
2. Use any existing bottom-up model to estimate a bottom-up saliency value for each patch;
3. Estimate the foreground and correlation priors using (5) and (7) and then use them to generate the modulated saliency map using (14); and
4. Convolve the saliency map with a disk filter (with the radius of 3) to fill in the “holes” generated by applying inconsistent foreground priors on adjacent

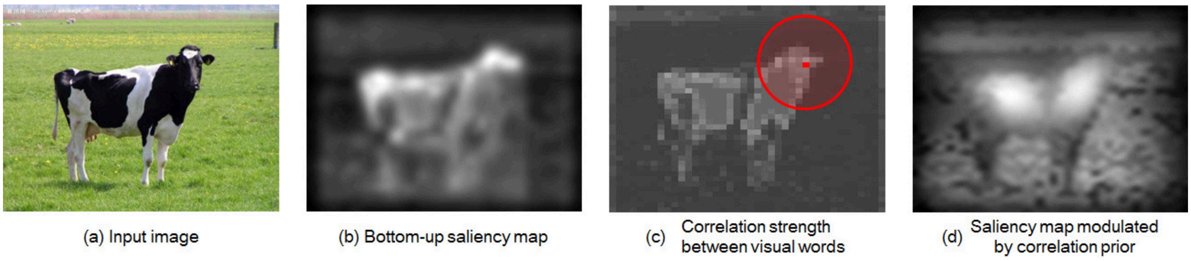

(a) Input image
(b) Bottom-up saliency map
(c) Correlation strength between visual words
(d) Saliency map modulated by correlation prior

Fig. 8 The correlation prior can help to recover the wrongly suppressed targets and to pop-out large salient target as a whole. Given an image (a), its bottom-up saliency map sometimes only pops-out the borders of the large salient target, while the inner smooth parts are ignored. For example, the saliency map in (b) is calculated using the model in (Itti et al 1998), which mainly pops-out the head and tail of the cow. To recover the wrongly suppressed targets (e.g., the inner smooth parts of the cow), we first estimate the correlation strength between visual words. The map in (c) shows the correlation strength Υ_{mn} between the patch marked in red and all the other patches. To increase the computational efficiency, we only consider the correlations between nearby patches (e.g., patches in the red circle controlled by a Gaussian term). Finally, the border patch marked with red, which successfully pops-out in the bottom-up competition, will help to recover the wrongly suppressed targets (i.e., large $P(\mathbf{s}_m)$ and $P(\mathbf{r}_n|\mathbf{s}_m)$ will lead to large $P(\mathbf{r}_n)$).

patches in smooth regions. Then conduct an exponential operation $S_r^*(n)=S_r(n)^3$ to remove the fuzzy background generated by using the additive Bayesian formulation.

From these processes, we can see that our proposed approach is biologically plausible since neurobiological evidences show that the bottom-up factors in human vision system act faster than the top-down factors (Wolfe et al 2000; Henderson 2003). Visual signals will first compete fairly to generate the bottom-up saliency, while a slower recall or recognition process is conducted to load the related prior knowledge into the working memory to bias the competition. In this process, the bottom-up saliency maps are modulated by various top-down priors to pop-out the real targets and suppress the real distractors. Moreover, we can also see that our approach exhibits good generalization abilities and can be easily extended. On one hand, we can plug any state-of-the-art bottom-up saliency model into the proposed framework, no matter how it detects saliency. On the other hand, if we can learn more kinds of prior knowledge (e.g., the task-dependent priors), we can easily incorporate them into our framework by calculating more kinds of top-down saliency maps, leading to a more accurate estimation of visual saliency.

5 Experiments

In this section, several experiments are conducted to prove the effectiveness of our approach. The main objectives are two folds: 1) to evaluate whether the prior knowledge is useful in estimating visual saliency and 2) to explore how the prior knowledge works in the estimation processes. Toward this end, we adopt two datasets in the experiments, including:

- **Toronto-120**. This popular dataset was first proposed in (Bruce and Tsotsos 2006) and has been used in many recent studies on visual saliency. It contains 120 color images. On each image, the fixations from 20 different subjects were recorded under the free-viewing conditions to reveal the locations of the salient targets.
- **MIT-1003**. This dataset was provided by Judd et al (2009). It consists of 1,003 images in total, most of which are color images. The eye tracing data were recorded from 15 subjects who free viewed these images. Compared with **Toronto-120**, this dataset is more challenging since images in this dataset are usually more complex and most of them contain a lot of targets and distractors.

On these two datasets, we adopt 10 approaches for comparison. All the source codes or executables can be found on the Internet. These approaches can be roughly categorized into two groups, including:

- **BU Group**. This group contains six bottom-up approaches, including CS[1] (Itti et al 1998), GB (Harel et al 2006), SR (Hou and Zhang 2007), FT (Achanta et al 2009), CA[2](Goferman et al 2010) and RA (Riche et al 2012). These bottom-up approaches only utilize the input visual signals to generate the bottom-up saliency maps. By comparing our approach with them, we wish to prove that incorporating the learned statistical priors can improve the performance of visual saliency estimation by modulating bottom-up saliency.
- **STAT Group**. This group contains four statistical approaches, including AIM (Bruce and Tsotsos

[1] The “winner-take-all” competition is not used in CS.

[2] The face detection component is not activated and here we can treat CA as a bottom-up approach.

2006), ICL (Hou and Zhang 2008), SUN (Zhang et al 2008) and SER (Wang et al 2010). These approaches also utilize the statistical image priors. By comparing our approach with them, we wish to prove that our framework is more effective in utilizing the learned prior knowledge.

In the comparison, we use the Area Under the ROC Curve (**AUC**) for performance evaluation. Since different saliency models often generate saliency maps with different resolutions, we resize all these saliency maps to the original resolutions of the input images for fair comparison. Suppose that the estimated saliency value for each pixel is in [0,1], a saliency model can be treated as a binary classifier by using all probable saliency values as the thresholds. On each threshold, a pixel can be classified as “fixated” or “non-fixated” using its saliency value. The classification results are then validated by the eye fixations to obtain the numbers of true positives, true negatives, false positives and false negatives. Consequently, we can calculate the True Positive Rate (TPR) and the False Positive Rate (FPR) on each threshold. Finally, the ROC curve can be built by plotting all the (TPR, FPR) points and the Area Under the ROC Curve can be used to quantify the performance of the saliency model. A perfect saliency model corresponds to an **AUC** of 1.0, while a random model will have an **AUC** of 0.5.

When computing **AUC**, the central fixation and salience bias is an important issue. That is, human fixations are often biased to image centers while non-fixated pixels usually distribute around image edges. However, the different distributions of fixated and non-fixated pixels often lead to unfair comparisons by favoring the saliency models that mainly emphasize the targets around image centers (e.g., using center-bias re-weighting) or ignore distractors near to image borders (e.g., using border cut). Inspired by the approach used in (Tatler et al 2005), we randomly re-sample the non-fixated pixels according to the distribution of fixations on all the images in the same benchmark. In the re-sampling process, we mainly refer to the fixation density maps that are usually generated by summing up a set of 2D Gaussians centered at each fixation point. For the sake of simplicity, we assume that each pixel in the fixation density map is assigned a score between [0, 1]. As shown in Fig. 9, we only re-sample the non-fixated pixels from those with scores lower than 0.05. In this manner, we can avoid possible ambiguities such as simultaneously selecting fixated and non-fixated pixels from the same object. For these candidate pixels, we generate a reference map by summing up all the fixation density maps from all the images in the same benchmark to guide the re-sampling process. Note that different benchmarks may have different reference maps due to different experimental settings (e.g., viewing distance, angle and image/screen resolution). A non-fixated pixel will be selected with high probability if the corresponding pixel in the reference map has a high score. Finally, only the selected non-fixated pixels, which are also biased to image centers, will be used for performance evaluation.

Actually, the proposed re-sampling strategy is quite reasonable, making the comparisons much fairer than using unbiased re-sampling. For instance, Judd et al (2009) randomly selected 10 fixated and 10 non-fixated pixels from the top 20% and the bottom 70% salient pixels on 100 images of **MIT-1003**. They further divided each image into center region and peripheral region, while the center region lies in a circle around image center whose radius equals to 42% of the distance from image center to image corner. After the division, the center region contains 78.8% fixated pixels and 24.5% non-fixated pixels, while the numbers change to 21.2% and 75.5% in the peripheral region, respectively. In this case, a model that simply emphasizes the center region will pop-out most of the fixated pixels and suppress most of the non-fixated pixels, leading to unfair comparisons. To address this problem, we re-sample the non-fixated pixels according to fixation density maps. After the re-sampling, the center region contains 71.0% fixated pixels and 64.1% non-fixated pixels, while the numbers change to 29.0% and 35.9% in the peripheral region, respectively. When the ratios of fixated and non-fixated pixels are comparable in each region, emphasizing only the center region will no-longer obtain much gain, making the comparisons much fairer.

Moreover, there are usually two ways to evaluate the overall performance on multiple images: 1) calculate the **AUC** score on each image first and then compute the mean and standard deviation of all the **AUC** scores; and 2) summing up the numbers of true positives, true negatives, false positives and false negatives on all images and generate a unique ROC curve, leading to a unique **AUC** score. Both ways can make sense and we will adopt the first way in the following experiments.

5.1 Whether It Works

In the first experiment, the main objective is to see whether our approach can really work. Toward this end, we adopt 6 bottom-up models to see whether the prior knowledge learned by our approach is effective to modulate the bottom-up saliency. In this process, we use 32 visual words and set $e^{-\alpha_b} \approx 0$ (the influences of these parameters will be discussed in other experiments). We also compare the modulated saliency maps with those

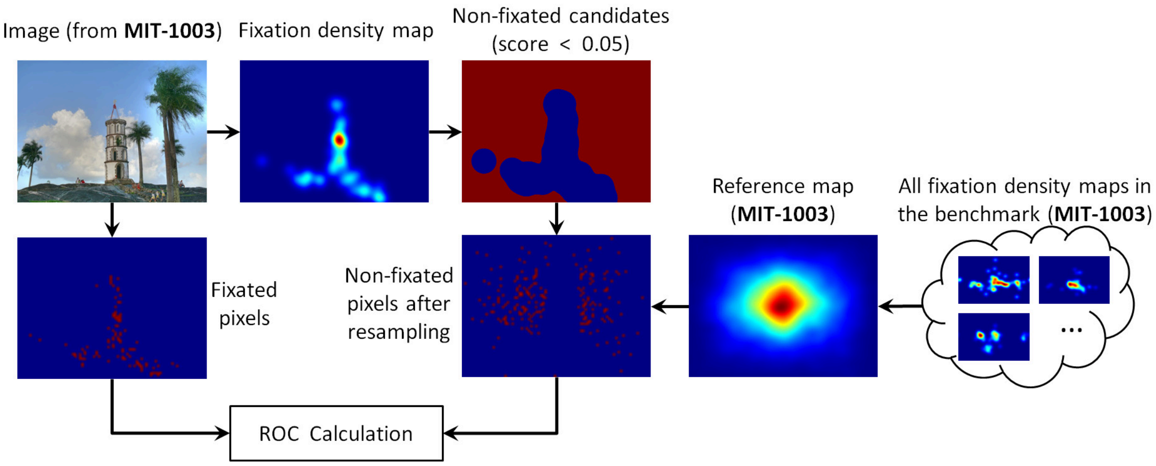


Image (from **MIT-1003**)
Fixation density map
Non-fixated candidates (score < 0.05)
Reference map (**MIT-1003**)
All fixation density maps in the benchmark (**MIT-1003**)
Fixated pixels
Non-fixated pixels after resampling
...
ROC Calculation

Fig. 9 Non-fixated pixels are re-sampled for fair comparison. In the re-sampling process, only the non-fixated pixels that are away from the fixated ones will be used as candidates to avoid possible ambiguities (i.e., pixels in the fixation density maps should have scores less than 0.05). Moreover, these non-fixated pixels are re-sampled with respect to a reference map generated by summing up all fixation density maps in the same benchmark. In this manner, most non-fixated pixels around image borders are ignored to avoid favoring saliency models that emphasize targets near image centers (e.g., using center-bias re-weighting) or ignore distractors around image borders (e.g., using border cut).

Table 1 Performance of various approaches on the two image benchmarks. The **OUR Group** illustrates the **AUC** scores of our approach when modulating the bottom-up saliency maps generated by different models.

Approaches		**Toronto-120**	**MIT-1003**
BU Group	CS	0.731 ± 0.123	0.678 ± 0.134
	GB	0.762 ± 0.134	0.700 ± 0.152
	SR	0.763 ± 0.122	0.693 ± 0.142
	FT	0.575 ± 0.126	0.554 ± 0.130
	CA	0.797 ± 0.100	0.713 ± 0.140
	RA	0.821 ± 0.090	0.722 ± 0.135
STAT Group	AIM	0.758 ± 0.109	0.700 ± 0.123
	ICL	0.787 ± 0.112	0.708 ± 0.153
	SUN	0.705 ± 0.129	0.667 ± 0.136
	SER	0.786 ± 0.113	0.704 ± 0.152
OUR Group	**Our**+CS	0.794 ± 0.118	0.714 ± 0.143
	Our+GB	0.804 ± 0.122	0.710 ± 0.153
	Our+SR	0.797 ± 0.112	0.706 ± 0.148
	Our+FT	0.710 ± 0.148	0.637 ± 0.168
	Our+CA	0.816 ± 0.102	0.725 ± 0.140
	Our+RA	**0.834 ± 0.086**	**0.738 ± 0.13**

maps generated by 4 approaches in the statistical group to see whether our framework can utilize the learned prior knowledge in a more effective manner. The **AUC** scores are shown in Table. 1. Some representative examples are illustrated in Fig. 10. Note that the fixation density maps are generated by filtering the pixel-wise fixation maps using a Gaussian kernel to account for inaccurate tracking results and the decreasing visual accuracy with increasing eccentricity from the fovea.

From Table. 1, we can see that the priors learned by our approach can improve the saliency maps generated by all the 6 bottom-up approaches. No matter how the bottom-up competitions are conducted in these approaches, our learned prior knowledge can effectively recover the wrongly suppressed targets and inhibit the improperly popped-out distractors. As shown in Fig. 10, a salient patch will *selectively* excite the tightly correlated patches using the correlation prior, while the distractors, especially the common background patches, can be effectively suppressed by using the foreground prior. In traditional bottom-up models, high saliency values are usually assigned to unique or rare visual subsets. However, the assumption that visual rarity corresponds to high saliency may not always hold since the background patches can sometimes become unique or rare (e.g., the building in Fig. 10(b) and Fig. 10(d)). These patches, which are already very familiar to the subjects, will be easily ignored. However, the bottom-up approaches will equally treat all the input signals since they have no prior knowledge on what the patch is. In our approach, we find that such common distractors often distribute around image edges. Therefore, we learn the distribution maps and quantify their center-bias properties to determine whether a patch is a common background patch or not. Then these patches will be effectively recognized and suppressed.

From Table 1, we can also see that the modulated saliency maps from CS, GB, SR, CA and RA can better predict human fixations than those saliency maps generated by another four approaches in the statistical group. This is mainly due to two reasons. First, most of the parameters used in our approach (e.g., the visual words, foreground and correlation priors) are learned from millions of images. Therefore, our approach can well handle the outliers. Second, we have adopted an

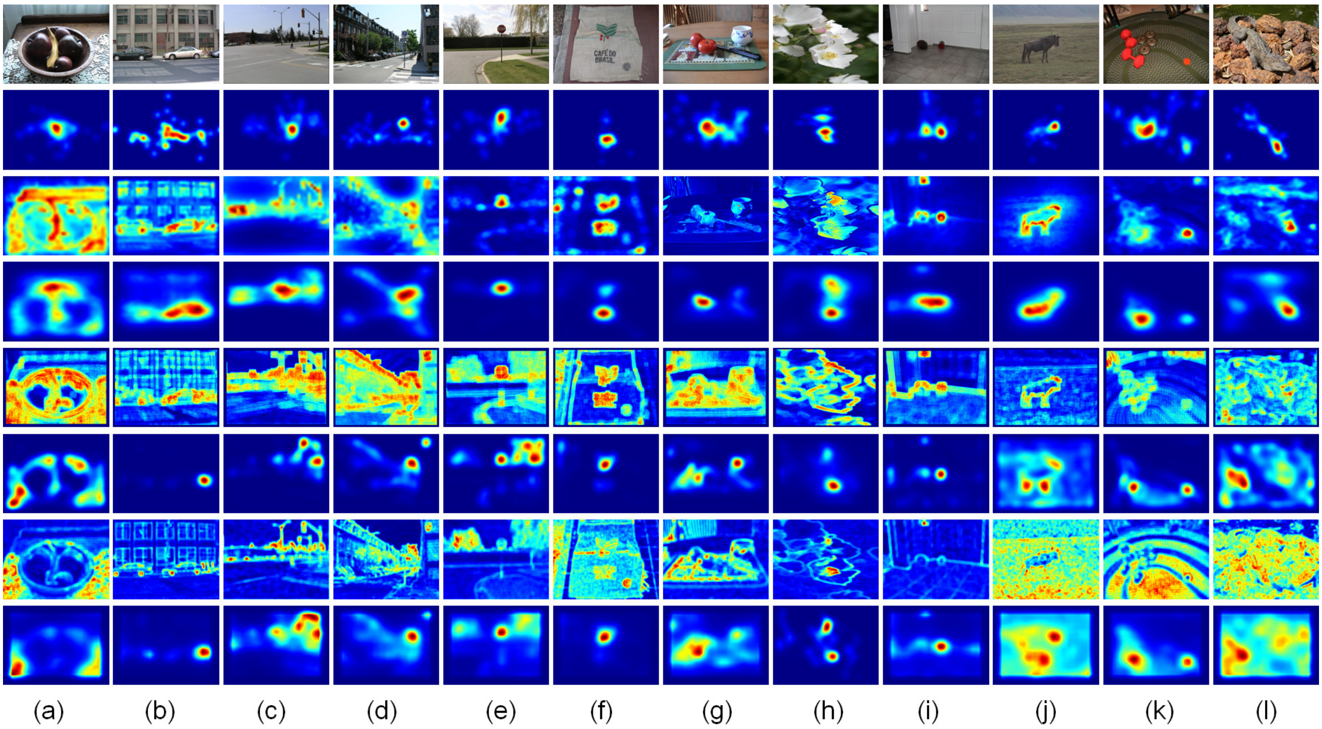


Fig. 10 Some representative saliency maps generated by various saliency models. The first row shows the input images and the second row illustrates the corresponding fixation density maps. The third row contains bottom-up saliency maps calculated by (a)-(b) CS; (c)-(d) GB; (e)-(f) SR; (g)-(h) FT; (i)-(j) CA; (k)-(l) RA. The 4th row demonstrates our results acquired by modulating the bottom-up saliency maps with the learned priors. The last four rows are results from AIM, ICL, SUN and SER, respectively.

opposite way to use the learned priors. In our approach, each patch is quantified to the nearest visual word and represented only by an integer label. In this process, many details are discarded but the integer label can work well since its main role is to retrieve the related prior knowledge. On the contrary, the other approaches will map the patch into a subspace with much higher dimensions and then estimate visual saliency in that subspace. Since the subspace may be not optimal, there may generate rich redundancies in the mapping. As illustrated in Fig. 10, these redundancies may generate many "noise" in the estimated saliency maps since it can be very difficult to distinguish targets from distractors when projecting all the signals onto specific basis. Therefore, these approaches achieve lower **AUC** scores.

Generally speaking, the main difference between our approach and all the other approaches discussed above lies in that we treat the input signals with bias. That is, each kind of prior knowledge will demonstrate a specific kind of biased selectivity in visual saliency estimation. For instance, the foreground prior will selectively suppress the patches that are judged as distractors, while the correlation prior will selectively enhance the patches around existing salient patches. Actually, such selectivity is well supported by biological evidences, which have proved that the top-down factors can bias the competition between the neurons linked with visual stimuli by favoring a specific category of stimuli.

In particular, the priors used in our approach are statistically learned from massive images in an unsupervised manner. Compared with (Torralba et al 2006) and (Chikkerur et al 2010) that mainly focused on incorporating the task-dependent priors, our approach can be used in much more scenarios to predict human fixations under free-viewing conditions since we have no assumption on the probable target-of-interests. Another advantage of learning the prior knowledge from millions of images is that the over-fitting risk can be largely avoided. Compared with the models trained on hundreds of images, our model often demonstrates impressive generalization ability. For instance, Judd et al (2009) selected 903 images from **MIT-1003** and extracted a set of low-, mid- and high-level features as well as the center prior to train a linear SVM model as the saliency model. Even with such a large feature pool, the **AUC** only reached 0.725 on the rest 100 images. Actually, if we adopt the same center-surround contrast features used in CS to train the linear SVM model, the

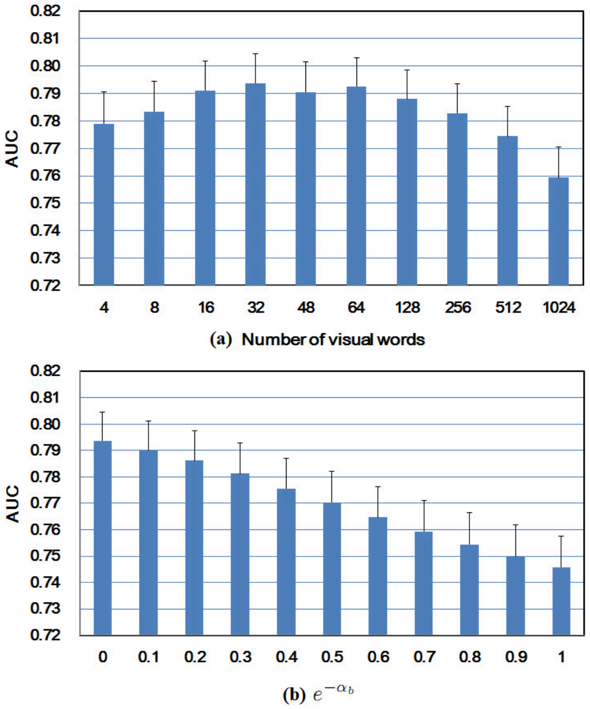
AUC will decrease to 0.684. This is natural since simple linear weights often lack the ability to model complex priors. Actually, people may attend to the salient targets in limited training images by only focusing on some specific features (e.g., human face as a special case). However, these features, which can be mined through supervised learning algorithms, may not always work well on the testing images (i.e., over-fitting). Therefore, it is necessary to learn the prior knowledge from massive unlabeled training images, probably by using unsupervised learning algorithms.

5.2 How It Works

To further investigate how the learned priors work in the top-down modulation, we conduct several experiments on the **Toronto-120** dataset to see the influence of various parameters and top-down priors. In these experiments, we adopt the bottom-up saliency maps generated by CS, which is treated as a baseline approach with **AUC**=0.731.

First, we conduct an experiment to see the influence of the number of visual words. In the experiment, we test 4, 8, 16, 32, 48, 64, 128, 256, 512 and 1024 visual words, and the **AUC** scores are shown in Fig. 11(a). From Fig. 11(a), we can see that our approach performs the best when using 32 visual words. When using more visual words, the performance gradually decreases. Although more visual words can better describe the details of the input images, they will also become more sensitive to noise and small fluctuations. For instance, two image patches with similar contents may be mistakenly quantized to different visual words. Due to the probable increase of such classification errors when using more visual words, the learned prior knowledge will become less reliable. Moreover, when using visual words less than 32, the influence of foreground prior will greatly decrease. For instance, when using 4 visual words, each visual word appears at each specific location with almost the same frequency. In this case, it is difficult to identify whether a patch belongs to foreground or not.

Second, we conduct an experiment to see the influence of α_b (i.e., $e^{-\alpha_b}= P(\mathbf{r}_n|\mathbf{s}_n,\bar{\mathbf{f}}_n)$), which indicates whether to trust the foreground prior when it makes conflict prediction with the bottom-up saliency model. When α_b is large (i.e., $e^{-\alpha_b}$ is small), we choose to trust the foreground prior, and vice versa. In the experiment, we vary α_b from $+\infty$ to 0. Equivalently, $e^{-\alpha_b}$ changes from 1 to 0 and the **AUC** scores are shown in Fig. 11(b). From Fig. 11(b), we can see that setting $e^{-\alpha_b}\approx 0$ can guarantee the best performance for **Our**+CS, which proves the effectiveness on the learned foreground prior.



(a) Number of visual words

(b) $e^{-\alpha_b}$

Fig. 11 The **AUC** scores of our approach on the **Toronto-120** dataset when using different parameters such as (a) different number of visual words and (b) different α_b. Note that here the error bar corresponds to $\frac{\sigma}{\sqrt{N}}$, where σ is the standard derivation of **AUC** and N is the total number of images in **Toronto-120**.

Third, we conduct an experiment to show the influences of foreground and correlation priors. By setting $P(\mathbf{r}_n|\mathbf{s}_m)$=0 in (1), we find that the **AUC** can reach 0.771 when only using the foreground prior. In contrast, the **AUC** can reach 0.746 when using only the correlation prior by setting $P(\mathbf{r}_n|\mathbf{s}_n)$=1 in (1). By combining these two kinds of prior knowledge, the overall **AUC** can reach 0.794. In particular, we find that the post-smoothing also contributes to the overall performance, while the exponential operation, which can often provide a “cleaner” viewing effect, has almost no influence on the **AUC** scores since it will not change the order of patch saliency values. When the post-smoothing operations are not used, the **AUC** can only reach 0.782. The reason is that there may exist some “holes” when modulating the bottom-up saliency using the foreground prior since adjacent patches in smooth regions sometimes are wrongly classified to different visual words. The overall **AUC** will probably decrease without filling such “holes” using the post-smoothing operation.

To sum up, the proposed approach can work well in visual saliency estimation and demonstrate several advantages in utilizing the prior knowledge. Actually, the whole framework can be uniquely characterized by two main phases, one *fast* bottom-up phase and one *slow* top-down phase. The bottom-up phase is mainly driven by data and transfers signals in a feed-forward manner. In the transmission, certain attributes of the data will be gradually extracted to active the related prior knowledge to generate feed-backward control signals. Compared with the models that contain pure bottom-up or top-down phase or parallel bottom-up/top-down phases, such framework has been proved to be consistent with the neurobiological mechanisms demonstrated in human perception experiments and takes advantage of optimizing each phase separately (Han and Zhu 2009; Wu and Zhu 2011).

Moreover, the framework in our approach can be easily extended. Once we learn some new kinds of prior knowledge, we can easily add them into our framework like the foreground and correlation priors. With this additive framework, we believe that the performance of visual saliency estimation can be gradually improved and a "perfect" model is expectable. Furthermore, our approach can be easily distributed on multiple computing units. This is very important since the learned knowledge database could become extremely large in the future (e.g., thousand kinds of prior knowledge). In our framework, different kinds of prior knowledge can be deployed on different computers, each of which can bias the competition of the input stimuli to generate a specific top-down saliency map and numerous top-down saliency maps can be fused to better predict human fixations.

6 Conclusion

This paper presents a novel approach for visual saliency estimation by using the statistical prior knowledge. We find that the bottom-up saliency estimated by existing stimulus-driven models can be further improved in top-down modulation. Thus we adopt a Bayesian framework to incorporate the influence of the prior knowledge, while such prior can be learned unsupervisedly from massive image statistics. From the experimental results, we can see that such statistical priors are very effective in recovering the wrongly suppressed targets and removing the improperly popped-out distractors.

In the future work, we will extend our approach by incorporating several new kinds of prior knowledge. We will also try to bring in some other top-down factors such as the task prior and global context prior. Since the proposed framework can be easily extended, we believe that its performance can be gradually improved by modulating the bottom-up saliency with more and more top-down factors.

Acknowledgements This work was supported in part by grants from the Chinese National Natural Science Foundation under contract No. 61370113, the National Basic Research Program of China (No. 2009CB320906) and the Supervisor Award Funding for Excellent Doctoral Dissertation of Beijing (No. 20128000103). This research was also partially supported by the Singapore National Research Foundation under its IDM Futures Funding Initiative and administered by the Interactive & Digital Media Programme Office, Media Development Authority.

References

Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition

Aziz MZ, Mertsching B (2008) Fast and robust generation of feature maps for region-based visual attention. IEEE Transactions on Image Processing 17(5):633–644

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

Bruce ND, Tsotsos JK (2006) Saliency based on information maximization. In: Advances in Neural Information Processing Systems, pp 155–162

Cerf M, Harel J, Einhauser W, Koch C (2008) Predicting human gaze using low-level saliency combined with face detection. In: Advances in Neural Information Processing Systems, pp 241–248

Cheng MM, Zhang GX, Mitra NJ, Huang X, Hu SM (2011) Global contrast based salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition

Chikkerur S, Serre T, Tan C, Poggio T (2010) What and where: a bayesian inference theory of attention. Vision Research 50(22):2233–2247

Elazary L, Itti L (2008) Interesting objects are visually salient. Journal of Vision 8(3):1–15

Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315:972–976

Frith C (2005) The top in top-down attention. Neurobiology of Attention pp 105–108

Goferman S, Zelnik-Manor L, Tal A (2010) Context-aware saliency detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition

Han F, Zhu SC (2009) Bottom-up/top-down image parsing with attribute grammar. IEEE Trans Pattern Anal Mach Intell 31(1):59–73

Harel J, Koch C, Perona P (2006) Graph-based visual saliency. In: Neural Information Processing Systems, pp 545–552

Henderson JM (2003) Human gaze control during real-world scene perception. Trends in Cognitive Sciences 7(11):498–504

Hou X, Zhang L (2007) Saliency detection: A spectral residual approach. In: IEEE Conference on Computer Vision and Pattern Recognition

Hou X, Zhang L (2008) Dynamic visual attention: Searching for coding length increments. In: Neural Information Processing Systems

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, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. In: IEEE International Conference on Computer Vision

Kienzle W, AWichmann F, Scholkopf B, Franz MO (2007) A nonparametric approach to bottom-up visual saliency. In: Advances in Neural Information Processing Systems, pp 689–696

Li J, Tian Y, Huang T, Gao W (2010) Probabilistic multi-task learning for visual saliency estimation in video. International Journal of Computer Vision 90(2):150–165

Li J, Levine MD, An X, Xu X, He H (2013) Visual saliency based on scale-space analysis in the frequency domain. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(4):996–1010

Liu H, Jiang S, Huang Q, Xu C, Gao W (2007a) Region-based visual attention analysis with its application in image browsing on small displays. In: ACM International Conference on Multimedia, pp 305–308

Liu T, Sun J, Zheng NN, Tang X, Shum HY (2007b) Learning to detect a salient object. In: IEEE Conference on Computer Vision and Pattern Recognition

Lu Y, Zhang W, Lu H, Xue X (2011) Salient object detection using concavity context. In: IEEE International Conference on Computer Vision, pp 233–240

Meur OL, Callet PL, Barba D, Thoreau D (2006) A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(5):802–817

Navalpakkam V, Itti L (2007) Search goal tunes visual features optimally. Neuron 53:605–617

Parikh D, Zitnick C, Chen T (2008) Determining patch saliency using low-level context. In: European Conference on Computer Vision

Peters RJ, Itti L (2007) Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: IEEE Conference on Computer Vision and Pattern Recognition

Riche N, Mancas M, Gosselin B, Dutoit T (2012) Rare: a new bottom-up saliency model. In: IEEE International Conference on Image Processing

Sun X, Yao H, Ji R (2012) What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp 1552–1559

Tatler BW, Baddeley RJ, Gilchrist ID (2005) Visual correlates of fixation selection: effects of scale and time. Vision Research 45:643–659

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

Tseng PH, Carmi R, Cameron IGM, Munoz DP, Itti L (2009) Quantifying center bias of observers in free viewing of dynamic natural scenes. Journal of Vision 9(7):1–16

Vikram TN, Tscherepanow M, Wrede B (2012) A saliency map based on sampling an image into random rectangular regions of interest. Pattern Recognition pp 3114–3124

Walther D, Koch C (2006) Modeling attention to salient proto-objects. Neural Networks 19(9):1395–1407

Wang W, Wang Y, Huang Q, Gao W (2010) Measuring visual saliency by site entropy rate. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition

Wolfe JM (2005) Guidance of visual search by preattentive information. Neurobiology of Attention pp 101–104

Wolfe JM, Alvarez GA, Horowitz TS (2000) Attention is fast but volition is slow. Nature 406:691

Wu T, Zhu SC (2011) A numerical study of the bottom-up and top-down inference processes in and-or graphs. Int J Comput Vision 93(2):226–252

Yang J, Yang MH (2012) Top-down visual saliency via joint crf and dictionary learning. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp 2296–2303

Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) Sun: A bayesian framework for saliency using natural statistics. Journal of Vision 8(7):1–20

Zhao Q, Koch C (2011) Learning a saliency map using fixated locations in natural scenes. Journal of Vision 11(3):1–15

Zhao Q, Koch C (2012) Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. Journal of Vision 12(6):1–15