Image saliency estimation via random walk guided by informativeness and latent signal correlations

Jia Li [a,b], Shu Fang [c,1], Yonghong Tian [c,*], Tiejun Huang [c], Xiaowu Chen [a]

[a] *State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China*
[b] *International Research Institute for Multidisciplinary Science, Beihang University, China*
[c] *National Engineering Laboratory for Video Technology, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China*

ARTICLE INFO

Available online 29 May 2015

Keywords:
Image saliency
Random walk
Informativeness
Latent correlation

ABSTRACT

Visual saliency is an effective tool for perceptual image processing. In the past decades, many saliency models have been proposed by primarily considering visual cues such as local contrast and global rarity. However, such *explicit* cues derived only from input stimuli are often insufficient to separate targets from distractors, leading to noisy saliency maps. In fact, the *latent* cues, especially the latent signal correlations that link visually distinct stimuli (*e.g.*, various parts of a salient target), may also play an important role in saliency estimation. In this paper, we propose a graph-based approach for image saliency estimation by incorporating both explicit visual cues and latent signal correlations. In our approach, the latent correlations between various image patches are first derived according to the statistical prior obtained from 10 million reference images. After that, the informativeness of image patches and their latent correlations are jointly considered to construct a directed graph, on which a random walking process is performed to generate saliency maps that pop-out only the most salient locations. Experimental results show that our approach achieves impressive performances on three public image benchmarks.

1. Introduction

Perceptual image processing, which aims to analyze images as human being does, is now becoming a hot research topic in the field of computer vision. In perceptual analysis, a key step is to locate important image content that demonstrates strong ability in capturing human visual attention. Toward this end, visual saliency can be estimated to quantize the importance of various image contents. By processing *visually salient* content with high priority, images can be efficiently analyzed, and the analysis results can better meet human perception.

In the past decades, hundreds of approaches have been proposed for visual saliency estimation. Among these approaches, most of them computed visual saliency as a kind of visual *rarity*, which were often measured by using explicit visual cues such as center-surround contrast and regional dissimilarity. For example, Itti et al. [1] proposed to estimate visual saliency by fusing multi-scale center-surround contrasts from multiple features. Harel et al. [2] represented images as graphs and detected salient pixels by defining the weights of graph edges as pixel dissimilarities. Moreover, some approaches tried to learn an optimal mapping mechanism from explicit visual cues to real-valued saliency scores. For instance, Navapakkam and Itti [3] proposed an approach to optimally combine local contrasts by

[*] Corresponding author.
E-mail address: yhtian@pku.edu.cn (Y. Tian).
[1] First two authors contributed equally.

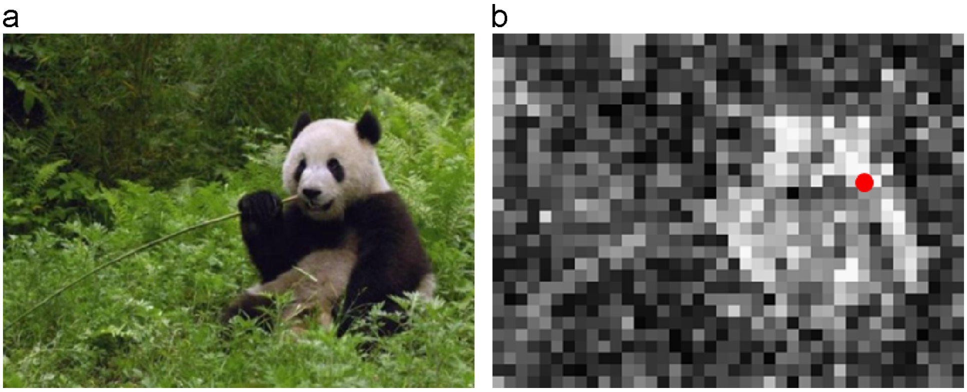

Fig. 1. Visual stimuli with distinct visual appearances may be inherently correlated. (a) A panda that consists of visually distinct parts; (b) the latent correlations between the 8×8 patch marked with red and all other patches.

maximizing the signal–noise-ratio. Zhao and Koch [4] proposed a boosting approach to train visual saliency model by fusing various visual features and their local contrasts. By focusing on explicit visual cues such as contrast and dissimilarity, all these approaches can pop-out targets and suppress distractors to some extent. However, such explicit cues are often insufficient to perfectly separate targets from distractors that share certain visual attributes. As a result, high saliency values may be wrongly assigned to distractors, resulting in "noisy" saliency maps.

To further separate targets from distractors, a feasible solution is to refer to additional saliency cues beyond the input image. By observing massive images, we find that various image contents can be inherently correlated, and such *latent signal correlations* actually link visually distinct stimuli (*e.g.*, various parts of a salient target, see Fig. 1 for an example). In visual saliency estimation, such latent signal correlations can work as a kind of prior knowledge that helps to further distinguish targets from distractors. Therefore, it is necessary to take such latent signal correlations into account in image saliency estimation.

Inspired by this idea, we propose an approach to estimate image saliency via random walk guided by informativeness and latent signal correlations. In our approach, the latent correlations between various image patches are first mined according to the statistical prior learned from 10 million reference images. These latent correlations between image patches, together with the patch informativeness, are then jointly considered to build a fully connected graph with directed edges and asymmetric weights. Under the guidance of informativeness and latent signal correlations, we divide image patches into three categories and adopt different random walking strategies between different types of patches. In this manner, the estimated saliency maps pop-out only the most salient locations while distractors can be well suppressed. Experimental results show that our approach achieves impressive performances in the comparisons with 13 approaches on three public image benchmarks.

Our main contributions are summarized as follows:

1. We incorporate latent signal correlations to facilitate the separation of targets and distractors. By exploiting the statistical prior obtained from 10 million reference images, the latent relationship between image patches can be well characterized so as to pop-out targets and suppress distractors.
2. We propose a graph-based approach that estimates image saliency via random walk guided by informativeness and latent signal correlations. In this approach, we divide image patches into three categories and apply different random walking strategies between different types of patches. In this manner, estimated saliency maps can pop-out only the most salient image locations while distractors can be well suppressed.

The rest of this paper is organized as follows: Section 2 briefly reviews related work, and Section 3 presents the learning process of latent signal correlations. In Section 4, we describe the details of the proposed saliency model. Experimental results are shown 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 saliency in images and videos. In this review, we mainly focus on image saliency estimation. According to the visual cues used to estimate saliency, existing approaches can be roughly grouped into two categories, including the bottom-up category and the top-down category. We will briefly review approaches in these two categories from the perspective of visual cues they ever used.

2.1. The bottom-up approaches

The approaches in the bottom-up category mainly focus on the explicit visual cues that can be directly extracted from the input visual stimuli (*e.g.*, local contrast, dissimilarity, entropy). For example, Itti et al. [1] first extracted the multi-scale center-surround contrasts from multiple features. These contrasts were then fused to estimate image saliency. Similarly, Gao et al. [5] computed image saliency as the discriminant power between center and surround regions. Vikram et al. [6] proposed to estimate visual saliency by calculating the local difference over randomly sampled rectangular regions in the Lab color space. The visual cue used here was the difference between the feature value of one pixel and the average feature value of a local, which can be also considered as a center-surround cue. Sen et al. [7] proposed to estimate image saliency by measuring the perceptual distinctness of patterns in the patches around pixels.

Instead of using these “local” features, some approaches tried to estimate visual saliency in a “global” manner. In [2], images were represented as fully-connected graphs while the edge weights were defined as pixel dissimilarities. By adopting a random walker on the graph, less visited nodes can pop-out and become salient. Duan et al. [8] estimated image saliency by jointly considering the dissimilarity, spatial distance and center bias of all image patches. Riche et al. [9] first extracted various visual features such as YCbCr color channels and Gabor orientations. By assuming that locally contrasted and globally rare features were salient, they adopted a sequential framework to estimate visual saliency. In [10], boolean maps were first generated via simple thresholding operations. Impressive saliency maps were then generated by analyzing the topological structure of boolean maps. Lu et al. [11] proposed that the convexity was a good indicator of saliency. By assuming that the region on the convex side of a curved image boundary was salient, they proposed a hierarchical framework to segment salient image regions from the background. In [12], image co-occurrence histograms were computed to capture the patch unusualness, which can be viewed as global uncommonness or local discontinuity.

Beyond spatial saliency models, some approaches tried to estimate saliency in the transform domain. For example, Hou and Zhang [13] proposed to extract the spectral residual over image intensity channel for estimating saliency. By using the Fourier transform, visual irregularities could be effectively extracted from the transform domain. In the later work [14], such an approach was further extended to detect salient locations from the sign function of DCT coefficients. In recent studies, the quaternion (or hypercomplex) Fourier transform was widely adopted in [15–18] to detect the visual irregularities from the spectra of multiple visual features. Besides, some approaches tried to estimate regional saliency. For instance, Cheng et al. [19] proposed an approach to segment images into regions and computed visual saliency using the regional contrasts. Perazzi et al. [20] first abstracted images into elements and then estimated their rarities and spatial distributions using the *Lab* color and Gaussian blurring kernels. Finally, these cues were combined to estimate regional saliency. Wo et al. [21] proposed to detect salient regions by measuring the aggregation degree of color and texture. Wang et al. [22] proposed to non-linearly combine various visual cues so as to highlight the locations around salient objects. In [23], the random forest regressor was trained to map regional features to saliency scores, which achieved impressive performance in detecting salient objects.

To sum up, the visual cues used in the bottom-up approaches can be described as “explicit.” That is, these approaches can estimate visual saliency using only the information from the target image. Generally, the explicit visual cues often have difficulties to distinguish targets from distractors, which may also have rare visual appearance (*e.g.*, tv logos, ad banners), and it is very difficult to suppress this kind of distractors without using the prior knowledge.

2.2. The top-down approaches

Compared with the approaches in the bottom-up category, the top-down approaches also utilize the visual cues derived from the prior knowledge (*e.g.*, face detection, visual dictionary and feature fusion strategy). The main characteristic of top-down approaches is that some latent cues are learned *before* processing the input visual signals. For instance, many approaches adopted the face detector and assumed human faces to be salient. For instance, Cerf et al. [24] integrated the face conspicuity map with other saliency maps to improve the performance of visual saliency estimation. In [25,26], human face was treated as a specific feature channel for estimating image saliency. In fact, the face detector can be viewed as a specific latent correlation model which can correlate the signals from nose, eyes and mouth. However, such domain-specific knowledge may not always work well in all scenes.

Another latent cue used in visual saliency estimation is the visual dictionary. The intrinsic assumption is that targets and distractors can be better distinguished after projecting the input visual signals onto some new subspaces described by the basis functions in the visual dictionary. In most approaches (*e.g.*, [27–33]) such visual dictionary is statistically learned from a large amount of images, while Yang and Yang [34] proposed an approach to jointly learn the visual saliency model and the visual dictionary. Given the learned dictionary, the original visual features can be transformed into the new subspaces (usually followed by the dimension reduction). By using the new features, visual saliency can be estimated by the same process in bottom-up approaches, *e.g.*, graph representations and random walker [28], information maximization [27], local and global rarities [30], coding length increment [32]. Actually, it is sure that certain subspaces (*i.e.*, kernels, basis functions and codewords) have some advantages in distinguishing targets and distractors. However, whether the learned subspaces are optimal or not is still unknown, especially for those subspaces learned by applying independent component analysis (*e.g.*, [27,28]) to only thousands of images.

Beyond the face detector and visual dictionary, the most popular latent cue is the feature fusion strategy. The intrinsic assumption is, when high-dimensional visual features are used, it is infeasible to manually adjust the contribution of each feature dimension to visual saliency. Instead, the optimal feature fusion strategy should be learned from the user data in a supervised manner. For example, Navalpakkam and Itti [3] proposed to learn the optimal feature weights by maximizing the signal-to-noise ratio. Kienzle et al. [35] adopted a Support Vector Machine (SVM) to seek a mapping model from local intensity to saliency value, while Judd et al. [25] introduced multiple low-level, mid-level and high-level features to train the SVM. Similarly, Lu et al. [36] collected a large number of visual features (*e.g.*, local energy, saliency values of existing bottom-up models, car and pedestrian detectors, face detectors, convexity maps). These features were then used to train a context-aware model for image saliency estimation. Zhao and Koch [4] proposed a boosting approach to iteratively train weak classifiers from a large feature pool. These weak classifiers are then fused with linear weights to generate a visual saliency model. In [37], the scene-specific feature fusion strategies were trained for estimating visual saliency. Peters et al. [38] assumed that there was a direct mapping from global features to saliency maps. Thus they simply trained a parameter matrix to transform global features directly to the fixation density map. In general, these learning-based approaches demonstrated promising performance by automatically fine-tuning parameters w.r.t. user data. However, the latent signal correlation is often ignored in these top-down approaches. Without considering such inherent correlations between various visual signals, targets and distractors are difficult be separated when they may share certain visual attributes.

To solve this problem, we propose a novel approach for image saliency estimation by jointly considering explicit saliency cues (*i.e.*, informativeness of image patch) and latent saliency cues (*i.e.*, latent correlations between image patches). The system framework of our approach is shown in Fig. 2. The main objective of our approach is to pop-out *only the most salient locations* and suppress the others via random walk guided by informativeness and latent signal correlations. In the next two sections, we will first demonstrate how to mine the latent signal correlations, followed by the details of incorporating such correlation with explicit visual cues in visual saliency estimation.

3. Mining the latent signal correlations

In visual saliency estimation, the latent correlations between input signals may play an important role to aggregate various visual cues from inherently related signals so as to separate targets from distractors. In this study, we aim to infer such latent correlations in an unsupervised manner by using the statistical prior derived from massive images. We assume that the concurrence attribute of visual signals can be an effective cue for deriving such latent correlations. Toward this end, we collect $M = 10{,}240{,}000$ non-duplicated images from Flicker. Each image is resized to have a maximum side length of no more than 320 pixels and the aspect ratio is preserved. For each 8×8 macro-block in these images, we extract the histogram of oriented gradients (HOG) using the approach in [39] to characterize its visual appearance. Given all the HOG descriptors extracted from all patches in 10 million images, we group them using the k-means algorithm to obtain $N_w = 1000$ visual words, denoted as $\{\mathcal{W}_i\}_{i=1}^{N_w}$.

Intuitively, two visual words can be inherently correlated if they frequently appear (and disappear) in the same images. To quantize such concurrence attribute of visual words, we use $\mathcal{P}_{ij} = \{\mathcal{W}_i, \mathcal{W}_j\}$ to denote a pair of visual words. Thus the probabilities of observing visual word $\mathcal{W}_i$ and visual word pair $\mathcal{P}_{ij}$ in all the M images can be calculated as

$$Pr(\mathcal{W}_i) = \frac{\mathcal{N}(\mathcal{W}_i)}{M}, \quad Pr(\mathcal{P}_{ij}) = \frac{\mathcal{N}(\mathcal{P}_{ij})}{M}, \tag{1}$$

where $\mathcal{N}(\mathcal{W}_i)$ is the number of images that contain visual word $\mathcal{W}_i$ and $\mathcal{N}(\mathcal{P}_{ij})$ can be derived by counting the number of images that contain both $\mathcal{W}_i$ and $\mathcal{W}_j$.

Given the individual probabilities $Pr(\mathcal{W}_i)$, $Pr(\mathcal{W}_j)$ and their joint probability $Pr(\mathcal{P}_{ij})$, we quantize the affinity relationship (namely, the correlation coefficient) of two visual words $\mathcal{W}_i$ and $\mathcal{W}_j$ using the approach in [40]

$$\varphi(\mathcal{W}_i, \mathcal{W}_j) = \log\left(\frac{Pr(\mathcal{P}_{ij})}{Pr(\mathcal{W}_i)Pr(\mathcal{W}_j)}\right), \quad \forall i \neq j. \tag{2}$$

From (2), we can see that two different visual words can be positively correlated ($\varphi(\mathcal{W}_i, \mathcal{W}_j) > 0$), independent

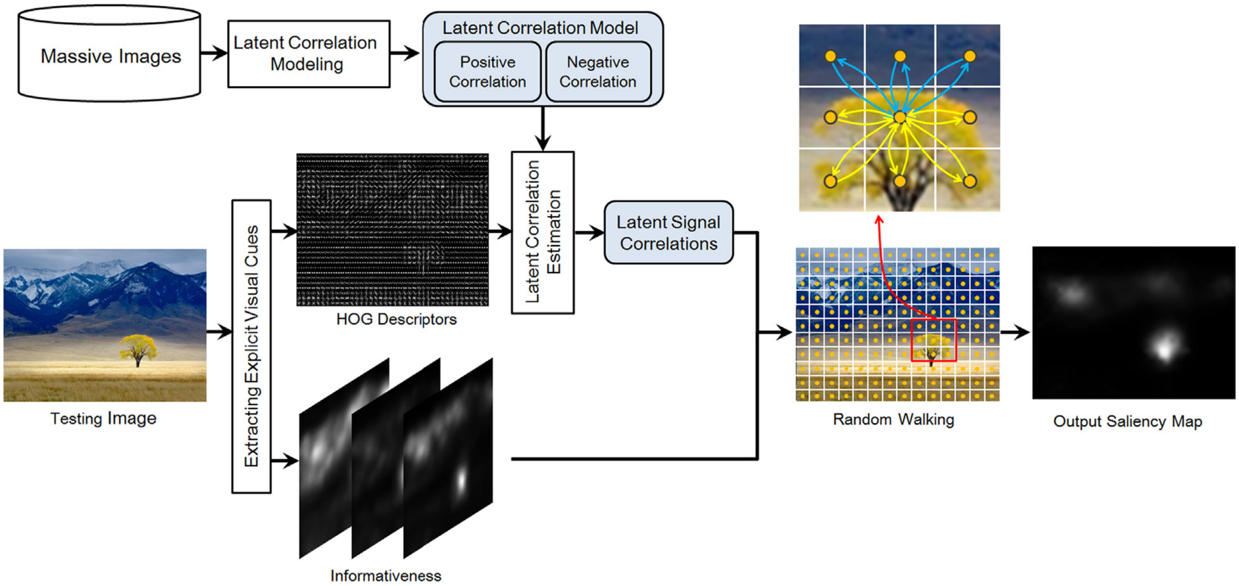

Fig. 2. The system framework of our approach. We first model the latent correlations between various types of visual signals using massive images. These learned latent correlations, which indicate inherently correlated relations with positive correlation coefficients or suggest mutually exclusive relations by negative ones, are then used to calculate visual saliency in a random walking framework along with explicit cues.

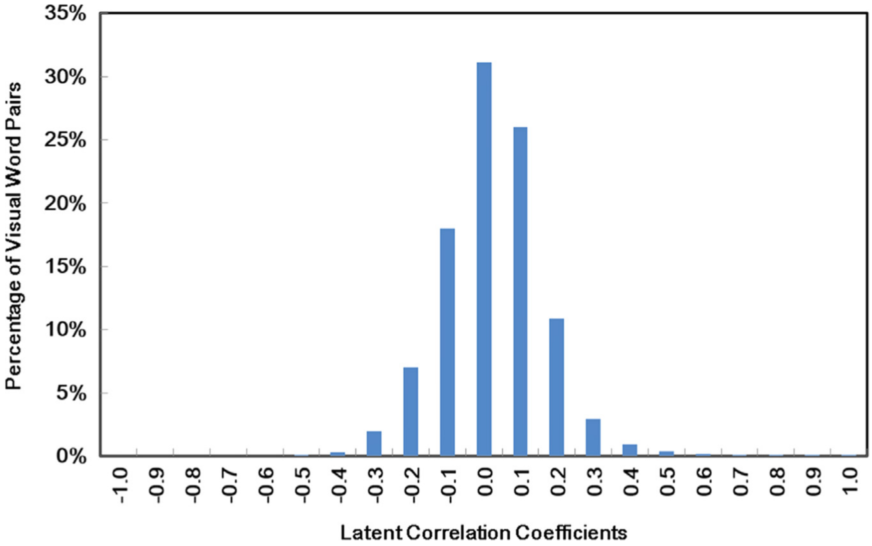

Fig. 3. The histogram of the latent correlation coefficients between all the 1000 visual words. Among all these visual words, most of them are independent or only weakly correlated, while some visual words have demonstrated strong positive or negative correlations.

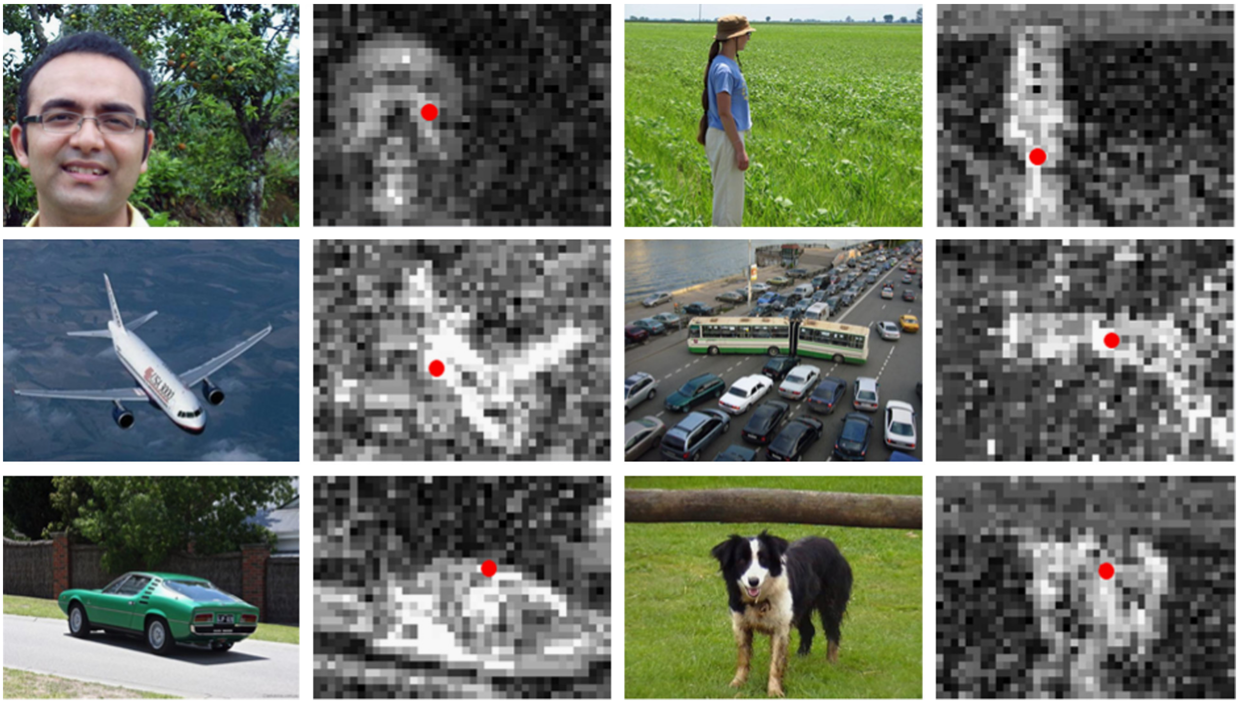

Fig. 4. Examples of the latent correlations between visual words. For each image, we show the latent correlations between the image patch marked with red and all the other patches. All the latent correlation coefficients are normalized to [0, 255] for illustration purpose.

$(\varphi(\mathcal{W}_i,\mathcal{W}_j)\approx 0)$ or negatively correlated $(\varphi(\mathcal{W}_i,\mathcal{W}_j)<0)$. As shown in Fig. 3, most visual words are independent or only weakly correlated, while some of them show strong positive or negative correlations. Moreover, 99.9% correlation coefficients fall in the range of $[-1,1]$ and 93% fall in $[-0.2,0.2]$. For the sake of simplification, we normalize $\varphi(\mathcal{W}_i,\mathcal{W}_j)$ into the range of $[-5,\ 5]$ if $i\neq j$ so that most correlation coefficients (about 93%) we may encounter fall in $[-1,1]$.

4. Saliency estimation with latent correlations

In this section, we will introduce how to estimate visual saliency via random walk guided by informativeness and latent signal correlations. The system framework of our approach is shown in Fig. 2. As illustrated in this framework, we will describe how to construct a graph with explicit visual cues and latent signal correlation, followed by details of image saliency calculation via random walking.

4.1. Extracting explicit and latent saliency cues

Before estimating image saliency, we have to first extract a set of explicit and latent saliency cues. Since the latent correlations are modeled on down-sampled images, we also resize each testing image to have a maximum side length of no more than 320 pixels. From each resized testing image, we extract K 8×8 image patches, denoted as $\{\{p_i\}_{i=1}^{K}$. For a patch p_i, we calculate a HOG descriptor h_i and find the nearest visual word $\mathcal{W}_{i^*}$ using the ℓ-2 distance. As a consequence, the latent correlation coefficient between image patches p_i and p_j can be derived by $\varphi(\mathcal{W}_{i^*}, \mathcal{W}_{j^*})$. Some representative examples of such latent correlation can be found in Fig. 4. From these samples, we conclude that

1. Patches in the same object often have positive correlations since they are likely to co-occur at the same time.
2. Targets and distractors are often negatively correlated since they rarely appear simultaneously.

Beyond the latent correlations, explicit cues play an important role in visual saliency estimation as well. In this study, we extract the wavelet energy $b_L(p_i)$, $b_a(p_i)$ and $b_b(p_i)$ as in [37] to evaluate the informativeness of the image patch p_i in three channels of CIE Lab color space. When a scene is being viewed, highly informative patches are often attended with high priority. Thus we can safely assume that image patches with high informativeness are likely to be salient. However, distractors, which can be also visually irregular, may have large wavelet energy, leading to large informativeness. Therefore, it is necessary to combine informativeness with latent correlations to separate targets from distractors so as to pop-out only the most salient locations.

4.2. Constructing graph with latent correlations

After extracting the explicit and latent saliency cues, three graphs are constructed by using the same latent correlations and different patch informativeness obtained from three color channels. For the sake of simplicity, we take only the luminance channel as an example to show how to build a directed graph for inferring saliency through random walking while the other two color channels can be processed in the same way. For the sake of simplification, we use $b_i = b_L(p_i)$ and $\varphi_{ij} = \varphi(\mathcal{W}_{i^*}, \mathcal{W}_{j^*})$ for short. Note that b_i is normalized into $[0, 1]$ for the purpose of computational efficiency.

The constructed graph is denoted as $G = \langle V, E \rangle$, where $V = \{v_i\}_{i=1}^{K}$ is the set of nodes (*i.e.*, image patches) and $E = \{e_{ij}\} i \neq j$ is the set of edges. Each node v_i can be characterized with the informativeness b_i and $K-1$ latent correlations $\{\varphi_{ij}\}_{j \neq i}^{K}$ while each directed edge e_{ij} is assigned a non-negative weight w_{ij}. Instead of defining the weights of edges with a uniform strategy, we first divide all the K image patches into three categories, including

1. *Probable distractors* $\mathbb{D}$: Considering the center bias prior, distractors often distribute around image borders. We randomly choose $0.2 \times K$ image patches with informativeness lower than 0.3 from the border area and treat them as probable distractors while the border area is defined as the 70% image regions around image borders.
2. *Probable targets* $\mathbb{T}$: Most targets have the capability to pop-out from simple background with simple features. Thus we select the patches with informativeness higher than $\epsilon = 0.8$ as probable targets.
3. *Other patches* $\mathbb{O}$: All the patches that are not included in $\mathbb{D}$ and $\mathbb{T}$ are grouped into this category.

Note that such a division of image patches represents only an initial guess for probable targets and distractors. As the initial guess is often rough and inaccurate, we refer to a random walking process to determine which patches can become real targets and distractors and which cannot. In the random walking, we propose to revise the weights of edges between these patches according to their latent correlations instead of heuristically adjusting the saliency of the probable targets and distractors. To estimate the weight w_{ij}, we have to consider which kinds of patches the edge actually links. Intuitively, a patch that has strong latent correlation with probable distractor will also be a probable distractor, we weight the edge for a patch $p_i \in \mathbb{D}$ as

$$w_{ij} = d_{ij} \cdot \max(b_j - \varphi_{ij}, 0), \quad \forall i \neq j, \ p_i \in \mathbb{D}, \tag{3}$$

where d_{ij} is a Gaussian distance weight that can be computed as

$$d_{ij} = \exp\left(-\frac{(x_i - x_j)^2 + (y_i - y_j)^2}{2\sigma^2} \right). \tag{4}$$

Here (x_i, y_i) and (x_j, y_j) are the coordinates of image patches p_i and p_j, respectively. σ is empirically set to 6% of the average image width and height. With this weight, a patch that is positively correlated with probable distractors will be less visited (*i.e.*, less salient), even though it has high informativeness. On the contrary, a patch that is negatively correlated with probable distractors will be visited for more times under the guidance of informativeness and latent signal correlations (*i.e.*, more salient).

Similarly, a patch that has strong latent correlation with probable target will also be a probable target. Thus we weight the edge for a patch $p_i \in \mathbb{T}$ as

$$w_{ij} = d_{ij} \cdot \max(b_j + \varphi_{ij}, 0), \quad \forall i \neq j, p_i \in \mathbb{T}. \tag{5}$$

In this manner, a patch that is positively correlated with probable targets will be visited more frequently even though its informativeness is low. On the contrary, a patch that is negatively correlated with probable targets will be less visited, making it being suppressed in the random walking process.

For the rest image patches, the edge weight only depends on the informativeness of the destination node:

$$w_{ij} = d_{ij} \cdot b_j, \quad \forall i \neq j, p_i \in \mathbb{O}. \tag{6}$$

From these definitions, we can see that our approach differs from existing graph-based saliency models mainly from two aspects. First, we divide all image patches into three categories while the edges for nodes in different categories are weighted with different strategies. Second, the latent correlations are incorporated into graph construction. In this manner, prior knowledge from human observations and massive image statistics are reflected in the graph structure, which can be used to guide the random walking process so that only the most salient locations can pop-out.

4.3. *Estimating saliency via random walk*

Based on the graph with fully-connected nodes and asymmetric edges, a Markov random walking process can be conducted to derive visual saliency. In the random walking process, the transition probability from node p_i to node p_j, denoted as H_{ij}, is defined according to the non-negative edge weights

$$H_{ij} = \frac{w_{ij}}{\sum_{j \neq i}^{K} w_{ij}}. \tag{7}$$

In the random walking process, targets can be gradually highlighted and distractors become suppressed while the equilibrium distribution of the random walk is computed to produce a visual saliency map. Note that we conduct the random walking process twice to balance the computational complexity and performance.

By conducting the random walking on the three graphs built from the informativeness of all the three Lab color channels, we can obtain three saliency maps. For a patch p_i, we denote its saliency values from three saliency maps as $S_L(p_i)$, $S_a(p_i)$ and $S_b(p_i)$. Thus the final saliency map can be obtained by linearly fusing the three saliency values:

$$S(p_i) = S_L(p_i) + S_a(p_i) + S_b(p_i). \tag{8}$$

Finally, we normalize the estimated saliency map into the dynamic range of [0,255] with the min–max normalization. Different from many other approaches, Gaussian smoothing is NOT used as a post-processing step in our approach to ensure that only the most salient locations can pop-out in the estimated saliency map.

4.4. *Comparison with related work*

Our approach is closely related to some existing approaches such as [41,42], which also used the latent signal correlations. Although in these approaches the correlations between patches or superpixels are all learned from the concurrence attributes, our approach differ remarkably from [41,42] in how to use the learned correlations.

In our approach, the learned correlations play different roles in linking different types of patches, while the patch types are determined by both informativeness and locations. During the graph-based random walking, the same degrees of latent correlations can either enhance or weaken a graph edge, depending on which types of graph nodes it actually links. In this manner, a patch positively correlated with probable distractors (*i.e.*, randomly selected non-informative patches in border area) will be inhibited, while a patch positively correlated with probable targets (*i.e.*, highly informative patches) will be enhanced. On the contrary, the approach in [42] only focuses on *enhancing* the patches that are positively correlated with probable targets (*i.e.*, patches popped-out in the bottom-up competition). It may have difficulties in performing the suppression operation due to its Bayesian formulation. Moreover, in our approach the learned correlations are only used to modulate the weights of graph edges, and salient locations are detected by considering both informativeness and such correlations. This is different from [41] which detects salient targets by mining the cohesive sub-graph from the graph formed only by the affinity scores of *superpixels in video*.

5. Experiments

In this section, we conduct several experiments to validate the effectiveness of our proposed approach. We will first introduce our experimental settings. After that, we will show the performance of our approach on eye fixation benchmarks to demonstrate its ability to capture the most attractive locations. At last, some extensive experiments are performed to provide a detail analysis of our approach.

5.1. *Experimental settings*

There exist various image benchmarks for visual saliency estimation. In this study, we use three benchmarks that are most frequently used in the literature:

- *MIT1003*: This benchmark was proposed by Judd et al. [25] which consists of 1003 images. For each image, the eye tracking data were recorded from 15 subjects in free-viewing conditions. Images in this benchmark often contain rich targets and distractors, making this benchmark very challenging.
- *Toronto*: This popular benchmark was first presented in [27]. It contains 120 images of indoor and outdoor scenes. Eye fixations were recorded from 20 subjects when each image was presented to each subject for 3 s.
- *ImgSal*: This benchmark was first proposed in [18]. It contains 235 color images, including 50 images with large salient regions, 80 images with intermediate salient regions, 60 images with small salient regions, 15 images with cluttered background, 15 images with repeating distractors and 15 images with both large and small salient regions. We use this benchmark to analyze the performances of various saliency models in processing salient regions at various scales.

On these three benchmarks, we compare our approach with 13 state-of-the-art approaches. According to the visual cues used in these approaches, we can roughly categorize them into two groups:

- *Bottom-up group*: This group consists of eight bottom-up approaches, including CS [1], GB [2], SR [13], RND [6], CA [26], LG [30], HFT [18] and BMS [10]. These approaches only utilize the explicit visual cues in visual saliency estimation. Note that the winner-take-all competition is not used in CS [1], and the face detectors are not used in CA [26];

- *Statistical group*: This group consists of five top-down approaches that statistically learn the visual dictionaries for visual saliency estimation, including AIM [27], SER [28], SUN [29], ICL [32] and SP [42]. Note that the bottom-up saliency maps used in SP are generated by CS.

In the experiments, the saliency map from each saliency model is resized to have the same resolution of the input image and normalized into [0,255]. To measure the performance of these saliency models, we adopt two different evaluation metrics. Toward this end, we first adopt the area under the *ROC* curve [43], which is a classic evaluation metric in visual saliency estimation. As a consequence, the visual saliency model is treated as a binary classifier by using all probable thresholds in $\{0, 1, \ldots, 255\}$ to calculate *ROC*. On each threshold, the estimated saliency maps are split into foreground and background regions to calculate the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN):

$$\begin{aligned}
&\mathrm{TP} = \#(\textit{foreground \& fixated}),\\
&\mathrm{TN} = \#(\textit{background \& non}-\textit{fixated}),\\
&\mathrm{FN} = \#(\textit{background \& fixated}),\\
&\mathrm{FP} = \#(\textit{foreground \& non}-\textit{fixated}).
\end{aligned} \tag{9}$$

where fixated locations are selected as all fixations received by each image. To avoid the center bias effect in the evaluation, the same number of non-fixated locations is chosen from the unattended area according to the overall fixation distribution on each benchmark. Note that

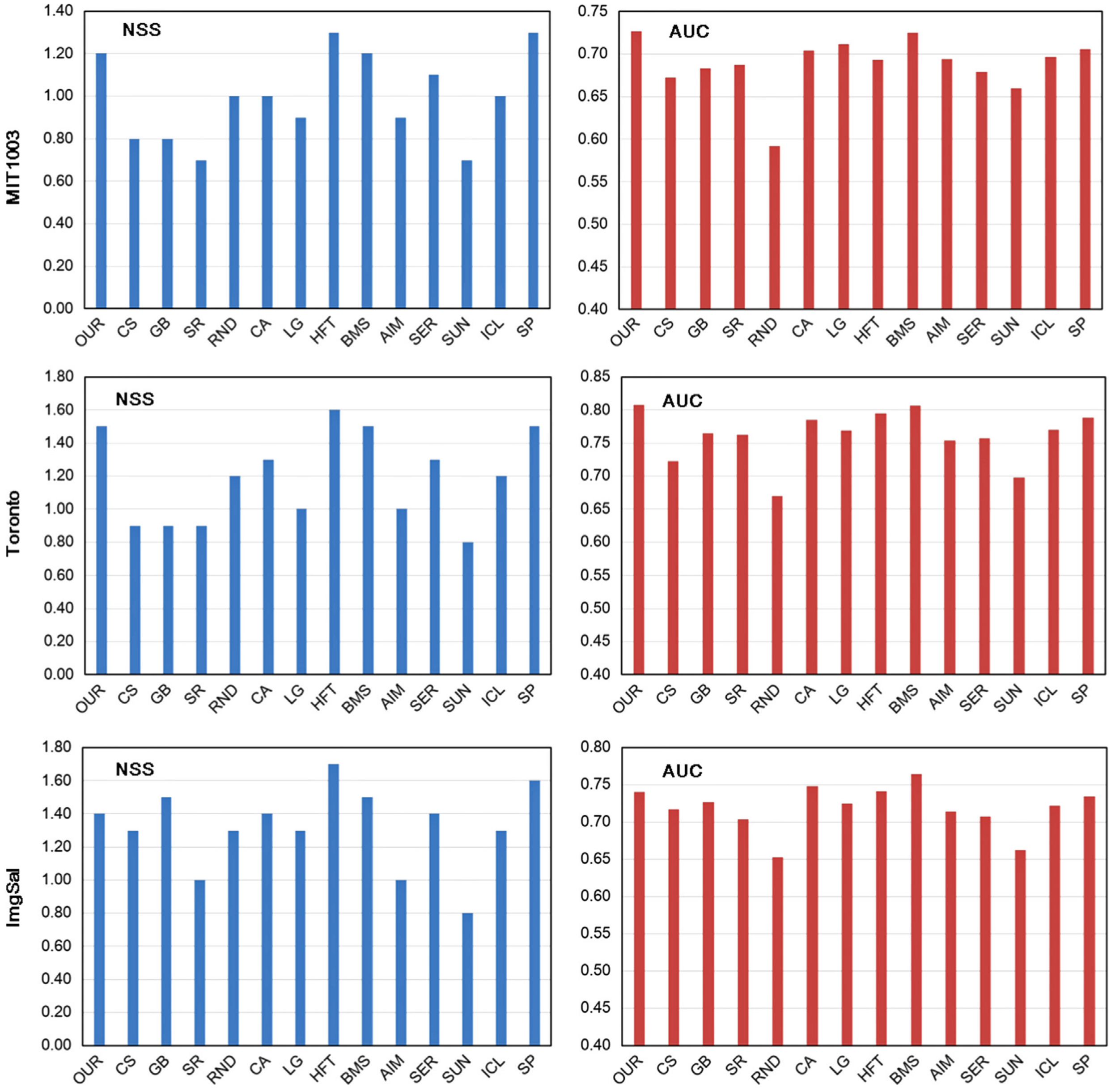

Fig. 5. Performance comparisons on *MIT1003*, *Toronto* and *ImgSal*.

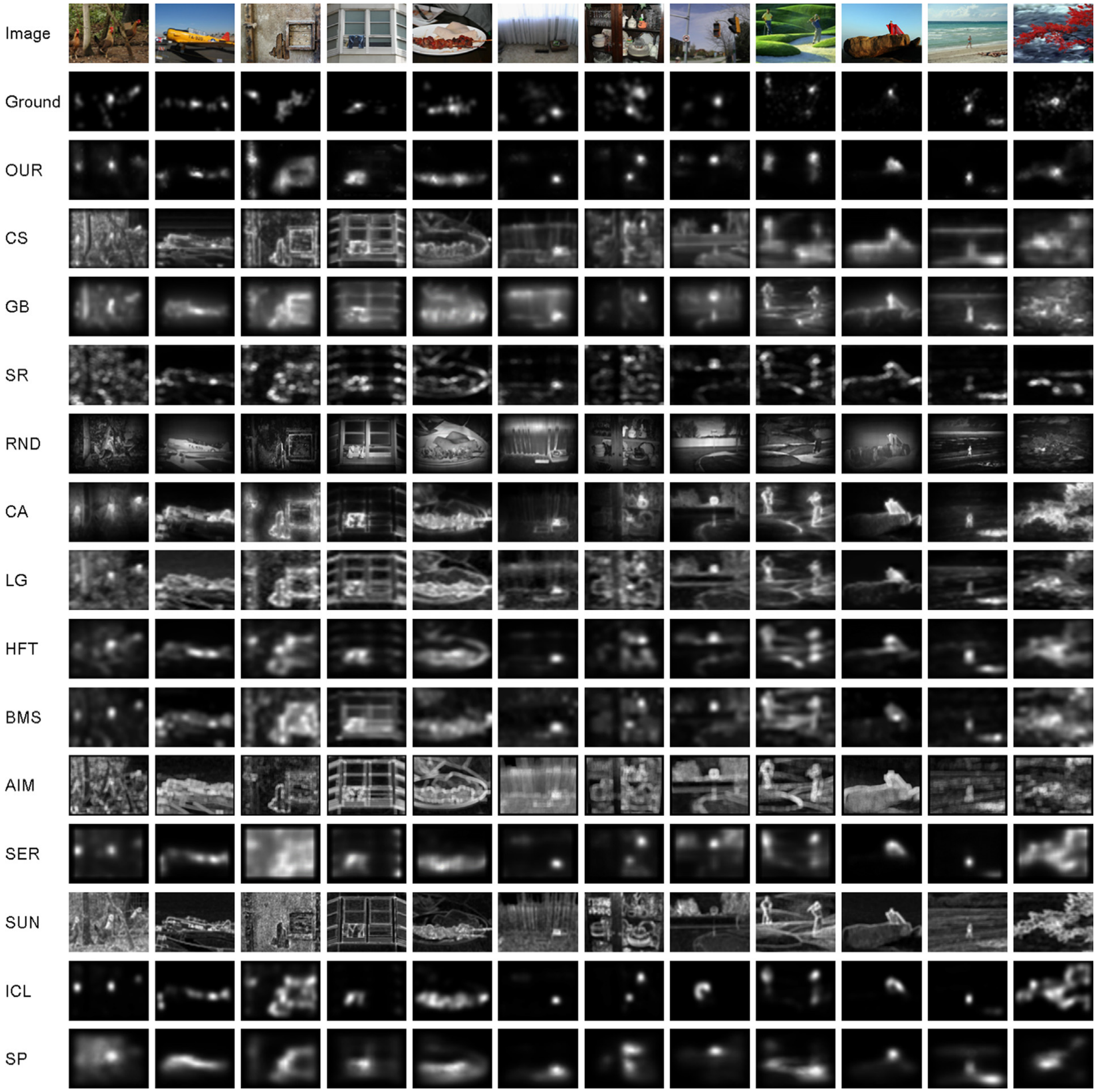

Fig. 6. The representative examples of our approach and the other 13 approaches on the *MIT1003*, *Toronto* and *ImgSal* benchmark.

we avoid to sample non-fixated pixels that are near to fixated pixels to avoid ambiguity. In this manner, fixated and non-fixated locations have similar center-biased distributions to avoid favoring models that simply emphasize only center regions. Consequently, the True Positive Rate (TPR) and False Positive Rate (FPR) can be calculated as

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}},\quad \mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{10}$$

Given the TPR and FPR scores at each threshold, an *ROC* curve can be generated using all the (TPR, FPR) pairs, and the area under this curve is used to quantize the performance of the visual saliency model, denoted as *AUC*. Note that our sampling strategy of non-fixated locations actually leads to a kind of shuffled *AUC*, which is different from traditional *AUC* but still ensures the fixation density map has an *AUC* of 1.0.[2]

Beyond *AUC*, we also use the Normalized Scanpath Saliency (*NSS*), which is another frequently used metric. *NSS* is defined as the mean response at fixations if the estimated saliency map *S* is normalized to have zero mean and unit standard deviation. Zero *NSS* means random prediction, and higher *NSS* implies better performance.

[2] More details and explanations for the sampling process of non-fixated locations and shuffled *AUC* computation can be found in [42].

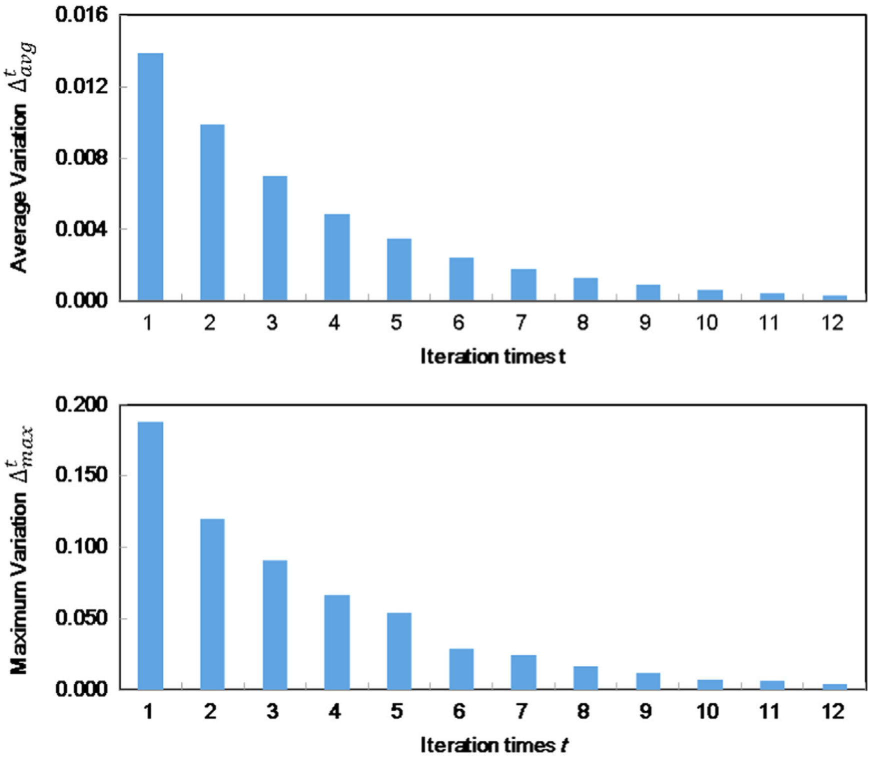

Fig. 7. The variation of the affinity scores when statistics is performed on different number of images. In the *t*th iteration, the average and maximum variations are computed according to the difference between the affinity scores derived from 2500 × 2^t images and 2500 × 2^{t-1} images. We can see that millions of images are used, both the average and the maximum variations are very low even if we double the number of images.

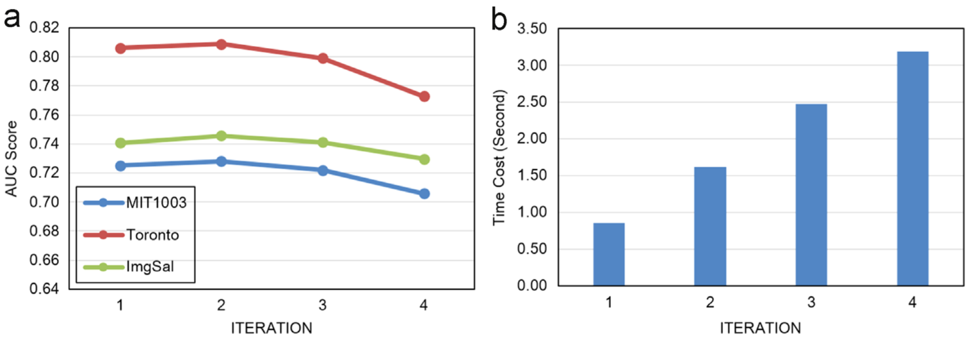

Fig. 8. The effect of random walking times on the *AUC* and the time cost. (a) The performance on three benchmarks; (b) the average time cost to process a 320 × 240 image.

Let $\mathbb{P}$ be the set of fixations, we compute *NSS* as

$$NSS = \frac{1}{|\mathbb{P}|}\sum_{p \in \mathbb{P}} \frac{S(p) - \mu_S}{\sigma_S}. \tag{11}$$

where μ_S and σ_S are the mean and standard deviation of all saliency values in the saliency map *S*, respectively. $|\mathbb{P}|$ is the number of fixations in $\mathbb{P}$.

5.2. *Performance comparison*

The objective of this experiment is to demonstrate the performance of our approach on locating targets and suppressing distractors. Experimental results on the three benchmarks are shown in Fig. 5. Note that a unified *AUC* in Fig. 5 is derived for each model over the whole dataset, while *NSS* is the average of *NSS* scores computed per image. Some representative examples are shown in Fig. 6.

From Fig. 5, we can see that on *Toronto* and *MIT1003* our approach ranks the first place twice in terms of *AUC*. For *NSS*, our approach ranks the second place and the third place on *Toronto* and *MIT1003*, respectively. The main reason is that our approach can utilize the prior knowledge on the latent signal correlations that are learned from massive images. Sometimes, targets and distractors can be very difficult to distinguish only using the explicit visual cues. In these cases, prior knowledge can help to pop-out true target and suppress true distractor. On *ImgSal*, the performance of our approach becomes somehow unsatisfactory in terms of *NSS*. This may due to the fact that our approach only pop-out the most salient locations in the saliency maps, which may have difficulties to process images with several large salient targets. As illustrated in Fig. 6, our estimated saliency maps are often much cleaner than the other 13 approaches. Generally, there are two reasons that can explain why our saliency maps are cleaner. First, the latent correlations between various patches are taken into account in visual saliency estimation. In this manner, the targets can be distinguished from distractors through the negative target–distractor correlations. Second, graph nodes of image patches are divided into three groups with respect to the location prior and informativeness while the edges of nodes in different groups are weighted differently under the guidance of informativeness and latent correlations. By performing random walking on the graph, targets will get enhanced and distractors can be suppressed effectively.

5.3. *Performance analysis*

Beyond the performance comparisons, we also conduct several small experiments to further demonstrate the effectiveness and efficiency of our approach.

5.3.1. *Robustness of the latent signal correlation model*

One probable concern in calculating affinity scores of visual words is whether ten million images are sufficient for inferring the prior knowledge. To address this concern, we conduct a small experiment. First, we generate $N_w = 1000$ visual words using 10,000 images newly crawled from the Internet. Using these visual words, we start from $M^0 = 2500$ images and iteratively estimate the latent correlation coefficients. In the tth iteration, we double the number of images (*i.e.*, $M^t = 2M^{t-1}$), and the correlation coefficients are denoted as $\{\varphi^t(\mathcal{W}_i, \mathcal{W}_j), i \neq j\}$. Consequently, the average and maximum variation of these coefficients in the t iterations can be calculated as

$$\Delta^t_{avg} = \frac{\sum_{i=1}^{N_w} \sum_{j \neq i}^{N_w} |\varphi^t(\mathcal{W}_i, \mathcal{W}_j) - \varphi^{t-1}(\mathcal{W}_i, \mathcal{W}_j)|}{N_w^2 - N_w},$$
$$\Delta^t_{max} = \max\{|\varphi^t(\mathcal{W}_i, \mathcal{W}_j) - \varphi^{t-1}(\mathcal{W}_i, \mathcal{W}_j)|, i \neq j\}. \qquad (12)$$

From Fig. 7, we find that the learned latent correlation coefficients may vary greatly at the first several iterations (*e.g.*, from 5000 images to 10,000 images). However, such variation will gradually reduce to zero when we bring in more images. In particular, we find that using five million new images in the final iteration will only slightly change such correlation coefficients. Therefore, we can safely assume that the learned latent correlation coefficients is statistically significant when ten million images are used.

5.3.2. *The effect of random walk iterations*

In our approach, a saliency map is calculated by iteratively performing the random walking process on graphs to highlight targets and suppress distractors. To obtain more less "noisy" saliency maps, the random walking process should be repeated several times, which may lead to high computational cost.

To this end, we conduct two experiments to find out the best number of random walking times to balance time cost and model performance. In the first experiment, saliency maps estimated with different iterations are evaluated by *AUC* on three benchmarks to check the effect of random walking times (as shown in Fig. 8(a)). In the second experiment, we record the average time cost to process a 320 × 240 image when the random walking is performed 1–4 times (as shown in Fig. 8(b)). All the experiments are conducted on a PC with 3.2 GHz CPU and 4G RAM using the Matlab implementation of our approach.

As shown in Fig. 8(a), the performance is already impressive by taking random walking process for once, which indicates that the proposed method can effectively pop-out targets and suppress distractors. With twice random walking, the *AUC* score slightly improves on all the three benchmarks (from 0.725 to 0.728 on *MIT1003*, 0.806 to 0.808 on *Toronto* and 0.741 to 0.746 on *ImgSal*). However, it will decrease in further random walking. The reason could be that more iterations are useless after that distractors get sufficiently suppressed.

On average, it costs about 0.85 s for calculating the saliency map of a 320 × 240 image by taking random walking process for once. From Fig. 8, two iterations of random walking are finally chosen in our approach to get a better performance while the time cost is still acceptable even with the Matlab implementation.

6. Conclusion

In this paper, we model the latent signal correlation by using the statistical prior derived from ten million reference images. With the latent signal correlations, targets can be distinguished from distractors as they often demonstrate negative correlations. In addition, the asymmetric and directed graphs are built by jointly considering both the explicit and latent saliency cues to effectively enhance targets and suppress distractors. Moreover, the random walking process is iteratively applied on graphs under the guidance of informativeness and latent correlations so as to pop-out only the most salient locations. Experimental results show that the latent correlation model is statistically significant, and the proposed approach achieves impressive performances on three public image benchmarks.

In our future work, we will seek a better way to mine the latent correlations, *e.g.*, by incorporating images with annotations. Moreover, we will also try to build task-driven saliency models that incorporate such statistical priors, which may be useful for applications such as scene understanding.

Acknowledgements

This work was supported in part by grants from the National Basic Research Program of China under Contract 2015CB351806, the Chinese National Natural Science Foundation under Contract 61370113, 61390515 and 61325011, and the Fundamental Research Funds for the Central Universities.

References

[1] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[2] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Advances in Neural Information Processing Systems (NIPS), 2007, pp. 545–552.

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

[4] Q. Zhao, C. Koch, Learning visual saliency by combining feature maps in a nonlinear manner using adaboost, J. Vis. 12 (6) (2012) 1–15.

[5] D. Gao, V. Mahadevan, N. Vasconcelos, The discriminant center-surround hypothesis for bottom-up saliency, in: Advances in Neural Information Processing Systems (NIPS), 2009.

[6] T.N. Vikram, M. Tscherepanow, B. Wrede, A saliency map based on sampling an image into random rectangular regions of interest, Pattern Recognit. (2012) 3114–3124.

[7] D. Sen, M. Kankanhalli, Salience computation in images based on perceptual distinctness, Signal Process.: Image Commun. 32 (0) (2015) 129–147.

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

[9] N. Riche, M. Mancas, B. Gosselin, T. Dutoit, Rare: a new bottom-up saliency model, in: IEEE Conference on Image Processing (ICIP), 2012, pp. 641–644.

[10] J. Zhang, S. Sclaroff, Saliency detection: a boolean map approach, in: IEEE International Conference on Computer Vision (ICCV), 2013, pp. 153–160.

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

[12] S. Lu, C. Tan, J. Lim, Robust and efficient saliency modeling from image co-occurrence histograms, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 195–201.

[13] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.

[14] X. Hou, J. Harel, C. Koch, Image signature: highlighting sparse salient regions, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 194–201.

[15] C. Guo, Q. Ma, L. Zhang, Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.

[16] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, C.-W. Lin, Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum, IEEE Trans. Multimed. 14 (1) (2012) 187–198.

[17] B. Schauerte, R. Stiefelhagen, Quaternion-based spectral saliency detection for eye fixation prediction, in: European Conference on Computer Vision (ECCV), 2012, pp. 116–129.

[18] J. Li, M. Levine, X. An, X. Xu, H. He, Visual saliency based on scale-space analysis in the frequency domain, IEEE Trans. Pattern Anal. Mach. Intell. 35 (4) (2013) 996–1010.

[19] M.-M. Cheng, G.-X. Zhang, N. Mitra, X. Huang, S.-M. Hu, Global contrast based salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 409–416.

[20] F. Perazzi, P. Krahenbuhl, Y. Pritch, A. Hornung, Saliency filters: contrast based filtering for salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 733–740.

[21] Y. Wo, X. Chen, G. Han, A saliency detection model using aggregation degree of color and texture, Signal Process.: Image Commun. 30 (0) (2015) 121–136.

[22] W. Wang, D. Cai, X. Xu, A.W.-C. Liew, Visual saliency detection based on region descriptors and prior knowledge, Signal Process.: Image Commun. 29 (3) (2014) 424–433.

[23] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, S. Li, Salient object detection: a discriminative regional feature integration approach., in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 2083–2090.

[24] M. Cerf, J. Harel, W. Einhauser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, in: Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 2009.

[25] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: IEEE International Conference on Computer Vision (ICCV), 2009, pp. 2106–2113.

[26] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2376–2383.

[27] N.D. Bruce, J.K. Tsotsos, Saliency based on information maximization, in: Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 2005, pp. 155–162.

[28] W. Wang, Y. Wang, Q. Huang, W. Gao, Measuring visual saliency by site entropy rate, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2368–2375.

[29] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell, Sun: a Bayesian framework for saliency using natural statistics, J. Vis. 8 (7) (2008) 32.1–32.20.

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

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

[32] X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments, in: Advances in Neural Information Processing Systems (NIPS), 2009, pp. 681–688.

[33] D. Parikh, C. Zitnick, T. Chen, Determining patch saliency using low-level context, in: European Conference on Computer Vision (ECCV), vol. 2, 2008, pp. 446–459.

[34] J. Yang, M.-H. Yang, Top-down visual saliency via joint crf and dictionary learning, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2296–2303.

[35] W. Kienzle, F.A. Wichmann, B. Scholkopf, M.O. Franz, A nonparametric approach to bottom-up visual saliency, in: Advances in Neural Information Processing Systems (NIPS), 2007, pp. 689–696.

[36] Y. Lu, W. Zhang, C. Jin, X. Xue, Learning attention map from images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1067–1074.

[37] J. Li, Y. Tian, T. Huang, W. Gao, Probabilistic multi-task learning for visual saliency estimation in video, Int. J. Comput. Vis. 90 (2) (2010) 150–165.

[38] R. Peters, L. Itti, Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.

[39] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[40] G. Zhao, J. Yuan, Discovering thematic patterns in videos via cohesive sub-graph mining, in: IEEE International Conference on Data Mining, 2011, pp. 1260–1265.

[41] Y. Luo, G. Zhao, J. Yuan, Thematic saliency detection using spatial–temporal context, in: IEEE International Conference on Computer Vision Workshops (ICCVW), 2013, pp. 347–353.

[42] J. Li, Y. Tian, T. Huang, Visual saliency with statistical priors, Int. J. Comput. Vis. 107 (3) (2014) 239–253.

[43] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (8) (2006) 861–874.