Computer Vision and Image Understanding 118 (2014) 40–49

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Tag-Saliency: Combining bottom-up and top-down information for saliency detection

Guokang Zhu [a,b], Qi Wang [a], Yuan Yuan [a,*]

[a] Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, PR China
[b] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, PR China

ARTICLE INFO

Article history:
Received 14 September 2012
Accepted 4 July 2013
Available online 30 August 2013

Keywords:
Computer vision
Saliency detection
Visual attention
Image tagging
Visual media
Semantic

ABSTRACT

In the real world, people often have a habit tending to pay more attention to some things usually noteworthy, while ignore others. This phenomenon is associated with the top-down attention. Modeling this kind of attention has recently raised many interests in computer vision due to a wide range of practical applications. Majority of the existing models are based on eye-tracking or object detection. However, these methods may not apply to practical situations, because the eye movement data cannot be always recorded or there may be inscrutable objects to be handled in large-scale data sets. This paper proposes a Tag-Saliency model based on hierarchical image over-segmentation and auto-tagging, which can efficiently extract semantic information from large scale visual media data. Experimental results on a very challenging data set show that, the proposed Tag-Saliency model has the ability to locate the truly salient regions in a greater probability than other competitors.

1. Introduction

Visual attention is an important mechanism of the human visual system. It helps access the enormous amount of complex visual information from the world effectively through rapidly selecting the most prominent or highly relevant subjects. In computer vision, this mechanism is modeled as saliency detection, which can provide the computational identification of scene regions that are more attractive to human observers than their surroundings. Based on the detected results, a higher and more complex processing can focus only on the salient regions when there is large-scale visual media data to be handled.

It is believed that visual attention is driven by two independent factors: (1) a bottom-up component, which is a task-independent component purely based on the low-level information, and (2) a top-down component, which is based on high-level information and guides attention through the volitionally controlled mechanisms. In recent years, many bottom-up saliency detection methods have been designed because they can provide a lot of useful information without prior knowledge about the scene. This kind of methods has already achieved a laudable performance in practical multimedia applications. For example, they have been successfully used for object detection and recognition [1,2], image quality assessment [3,4], video summarization [5], image/video compression and resizing [6,7], adaptive content delivery [8], and image segmentation [9,10]. In the meantime, in contrast to the focused interest in modeling the bottom-up guided attention, few studies have attempted to explore top-down factors.

Evidence from visual cognition researches indicates that the low-level factors dominate the early visual stage. But late on, it is mainly the high-level factors that direct eye gazing and changing [11,12]. For instance, when looking at images in the first row of Fig. 1, people are usually attracted by some specific regions standing out from the rest of the scene with respect to color, intensity, or orientation during the first few hundreds of milliseconds. Then immediately they are able to allocate attention spotlight to the cars, texts, pedestrians, or other objects with special concepts or meanings, while are easy to ignore the things usually treated as background. Thus, top-down factors should also be taken into account separately in visual saliency detection, and it is reasonable to believe that efficient and effective utilization of high-level information can help improve current saliency detection performance.

Most studies on top-down attention are still at the descriptive and qualitative level. Few completely implemented computational models are available, and most of them are based on eye tracker or a series of specific object detectors. However, in practical situations, there may be no eye movement data or too many objects to be handled in large-scale data sets. These methods therefore will be greatly limited by the harsh conditions of use and the high computational complexities.

In fact, there are many other techniques in computer vision besides eye-tracking and object detection, which can help automatic extraction of high-level information. Inspired by the works of

* Corresponding author. Tel.: +86 29 88889302.
E-mail address: yuany@opt.ac.cn (Y. Yuan).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.07.011

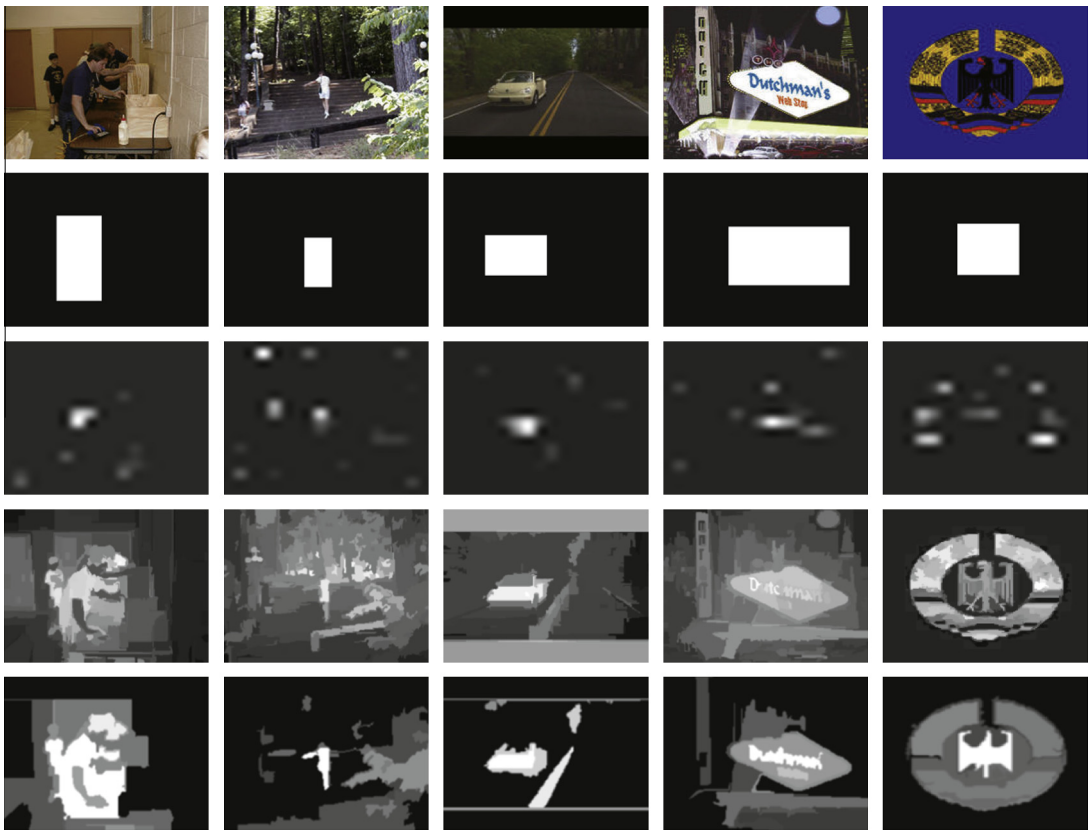

Fig. 1. Saliency detection. From top to bottom, each row respectively represents the original images, the ground truths, the saliency maps calculated by IT [13], RC [14], and the proposed model.

image auto-tagging [15–17] and tag completion [18], which can predict for images the relevant keywords from a vocabulary, this paper proposes a Tag-Saliency model capable of tackling high-level information and convenient to use. The proposed model takes advantage of the following two aspects:

First, the image auto-tagging technique is introduced into saliency detection task. Compared with the previous top-down saliency methods, the proposed method can extract high-level semantic information from large scale visual media data through only one unified image tagging model. This image tagging model can be obtained from the suitable training sets, without the need for a large number of specific object detectors.

Second, a hierarchical over-segmentation and global contrast based paradigm is proposed, which can integrate the high-level and low-level information for effective saliency estimation. In this paradigm, each image will be pyramid decomposed into a sequence of hierarchical regions, where regions at one segmentation level may be partitioned into several more finer spatial subregions at the next level. Based on the hierarchical over-segmented regions, both low-level and high-level information are extracted for global contrast analyzing. The main advantage of hierarchical over-segmentation is the ability to provide multiple information for regional tagging, which can yield the excellent tagging accuracy. The global contrast analysis is employed here. This is mainly because the experimental experiences which indicate that global contrast based models are often connected with outstanding performances in practice [14,19,20].

The rest of this paper is organized as follows. Section 2 reviews some relevant mainstream works. Section 3 details the proposed model. Section 4 presents the extensive experiments to verify the effectiveness of the proposed model. Section 5 gives a quantitative discussion to analyze all the factors that influence the performance of our model, and the conclusion follows in Section 6.

2. Related work

Classical bottom-up saliency detection methods choose to utilize low-level information to calculate contrasts of image regions with respect to their surroundings. According to the range of comparative reference regions they used, these methods can be roughly classified into local contrast based and global contrast based.

Local contrast based methods determine saliency of each image region by calculating the low-level contrast between the examined region and its local neighborhoods. Many early studies are related with the biologically inspired visual model introduced by Koch and Ullman [21]. The ground-breaking implementation of this model is the work of Itti et al. [13]. They employ a Difference of Gaussian (DoG) approach to extract multi-scale low-level information from images, and utilize this information to define saliency by calculating center-surround differences. Later on, in the work of Walther et al. [22], the method of [13] is modified with a hierarchical recognition system to recognize salient objects. Similarly, Han et al. [9] modify the method of [13] with Markov Random Field (MRF) based region growing to produce salient regions.

Besides the aforementioned approaches, there are many other methods which are not clearly related with a biologically inspired visual model, but strongly related with pure local contrast analysis. For instance, Gao et al. [23] detect saliency by estimating the Kullback–Leibler (KL) divergence of a series of DoG and Gabor filter responses to calculate the local contrast between the examined location and its surrounding local region. Harel et al. [24] design a graph based method. They firstly form activation maps by combining some excellent feature maps, and then use graph algorithms and a measure of contrast to achieve conspicuous parts. Hou and Zhang [25] introduce a model in frequency domain, which defines saliency of a location based on the difference between the

log-spectrum feature and its surrounding local average. Achanta et al. [26] calculate saliency by computing center-surround contrast of the average feature vectors between the inner and outer subregions of a sliding square window. Zhang et al. [27] utilize a Bayesian framework to evaluate saliency as the Shannon self-information of pointwise visual features. Seo and Milanfar [28] measure saliency by using Local Steering Kernels (LSK) to build a "self-resemblance" map, which actually captures the local gradient contrasts.

Recently, global contrast based methods have attracted many interests and announced promising results. These methods take into account the global statistics over the whole image. For example, Zhai and Shah [29] propose to utilize the global motion contrast by calculating the keypoint correspondences and geometric transformations between consecutive images, as well as the pixel-level global color contrast in each individual image. Shao et al. [30–32] deem that the salient regions have the globally unpredictable characteristics, and construct their saliency model based on Shannon entropy analysis and scale auto-selection. Achanta et al. [33] highlight salient regions by computing the global contrast between each original color and the average color of the entire filtered image. Goferman et al. [34] measure saliency based on patch-level global contrast, which takes into account the k most similar patches of each examined patch in the whole image. Bruce and Tsotsos [35] detect saliency by calculating a probability density function to find the maximum information component over the complete scene. Cheng et al. [14] propose a region-level saliency extraction method, which is based on analyzing the color contrasts over all the over-segmented image region. Wang et al. [36] detect saliency by extracting the anomaly region relative to a large web image dictionary through k-nearest-neighbor (kNN) retrieval. Liu et al. [19] define a global feature of color spatial distribution in their works to extract prominent colors, which can be further enhanced by the local center-surround histogram and multi-scale contrast through Conditional Random Field (CRF) learning. Li et al. [37] introduce a Co-Saliency model, which can detect the common foreground object from image pairs. They firstly employ a linear combination of the saliency maps from some traditional saliency detection methods (e.g., [13,33,25]) to generate single-image saliency maps. Then, a co-multilayer graph is constructed based on the over-segmented region to estimate multi-image saliency through the global similarity computation. Wang et al. [38] incorporate near-infrared cues into contrast analyzing. More recently, Perazzi et al. [20] introduce a refreshing work. They reconsider some previous excellent global contrast based models [14,19], and use a series of Gaussian filters to integrate them in an unified way. Lang et al. [39] detect salient positions by seeking the consistently sparse elements from the entire image.

The aforementioned bottom-up models utilize only the low-level information, such as color, intensity, orientation, texture, depth, shadow, and motion. Thus, success of such approaches will be limited to free-viewing and early visual dominated tasks [13,35]. When compared with the actual performance of human in the daily lives, there exists a large gap due to neglecting the role of top-down factors [40].

As for the top-down attention components, there are many evidences which indicate that objects, such as humans, faces, cars and texts, are better eye fixation predictors than low-level information [41–43]. However, only a few completely implemented computational models have been proposed to utilize such high-level information [44–46]. Most of these methods are based on eye movement tracking or object detection. For instance, Navalpakkam and Itti [47] introduce the conceptual guidelines to model the task driven visual attention. Peters and Itti [44] use gist to predict view fixation, and learn their prediction model from instances where people are looked in scenes with different gists under particular tasks. Judd et al. [45] indicate that high-level information such as humans, faces, and texts will attract more gazes, because these cues can convey more information in a scene than other low-level cues. All these existing top-down saliency detection methods have a common disadvantage that the practicability and applicability will be greatly limited by the harsh conditions of use, as well as the high computational complexity.

3. Proposed Tag-Saliency model

This section specifies a general Tag-Saliency model, which aims at estimating the probability (i.e., between 0 and 1) of each over-segmented region being salient, according to the global contrast of both low-level and high-level information in the scene. An overview of this model is presented in Fig. 2. In this model, the saliency value $S(r_i)$ assigned to a region r_i, is determined by two independent components:

$$S(r_i) = U(r_i) \cdot \exp[\sigma_v^2 \cdot V(r_i)], \tag{1}$$

where the first component $U(r_i)$ denotes the global contrast of low-level information calculated between r_i and other regions from the same segmentation level. The second part $V(r_i)$ denotes the global contrast of high-level semantic information extracted by an image

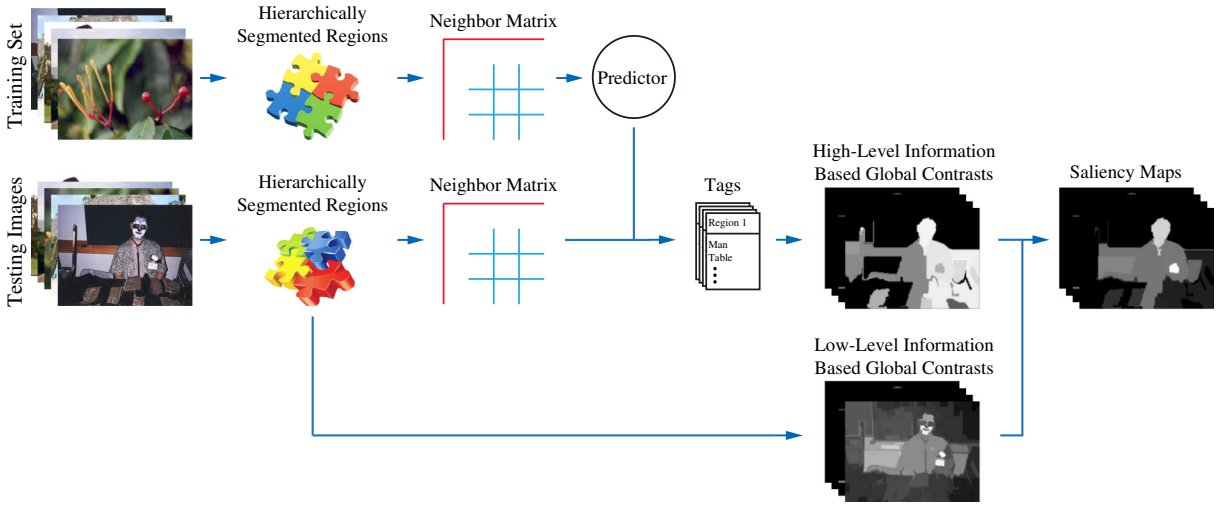

Fig. 2. Summary of the Tag-Saliency model, which utilizes both the low-level and high-level information to estimate saliency.

auto-tagging technique [15]. It should be noted that, as the second component is experimentally observed with much larger range of values than the first, in this formulation, an exponential function is employed to normalize $V(r_i)$ with a parameter σ_v, where σ_v controls the decay of the exponential, and is experimentally fixed to *0.5* in all our experiments. In the final step, the produced saliency map will be linear projected to the range $[0,1]$.

3.1. Hierarchical over-segmentation

For the image pre-processing, an excellent over-segmentation method [48] is employed to hierarchically segment the image into regions. The selected segmentation method is reported with highly efficiency, and can capture the perceptually important regions. An initial over-segmentation is performed by partitioning an image into multiple regions, and the hierarchical segmentation is implemented by repeatedly partitioning regions at one segmentation level into several more finer spatial subregions at the next level. The final saliency map is integrated at the target segmentation level ($l = 1$) that contains about 15 regions.

3.2. Low-level information based global contrast

In practice, there is an intuition that some regions in a real-world image with distinctive color or visual complexity are more likely to attract the interests from observers at their first glance. Therefore in this subsection, the global contrasts of regions are analyzed directly based on these two visual cues.

Visual Complexity Contrast. For the visual complexity, there is a popular measurement derived from information theory—entropy, which is associated with the uncertainty in information [30,28], and has been used as a simple and effective quantitative indicator [49]. Based on this measurement, the contrast of visual complexity $D_e(r_i,r_j)$ between region r_i and r_j of the examined segmentation level can be simply defined as

$$D_e(r_i,r_j) = [H(r_i) - H(r_j)]^2, \tag{2}$$

where $H(r_i)$ represents the entropy of image region r_i, which is estimated by

$$H(r_i) = \sum_{p=1}^{n_{c,i}} f(c_{p,i}) \cdot \log_2 f(c_{p,i}), \tag{3}$$

where $c_{p,i}$ is the pth color of region r_i, $n_{c,i}$ is the number of colors contained in r_i, and $f(c_{p,i})$ is the probability of $c_{p,i}$ in this region.

Color Contrast. Color is usually considered as one of the most important cues in computer vision. Generally, the distributions of colors in image regions are independent of each other with different variances, and are truly not normal most of the time. In this case, the independent t-test is the appropriate choice to analyze the difference between two sets of data. Accordingly, the Student t-value is employed to provide an accurate measure for the color difference between two regions, rather than directly calculating the Euclidean distances of colors in all the positions [14]. More specifically, the color contrast between region r_i and r_j is defined as

$$D_c(r_i,r_j) = \|\mu_{c,i} - \mu_{c,j}\|^2 \cdot \left(\frac{\sigma_{c,i}^2}{n_i} + \frac{\sigma_{c,j}^2}{n_j}\right)^{-\frac{1}{2}}, \tag{4}$$

where $u_{c,i}$ and $u_{c,j}$ represent the average color of region r_i and r_j respectively, $\sigma_{c,i}^2$ and $\sigma_{c,j}^2$ are the variances, and n_i and n_j are the total numbers of pixels in the corresponding regions.

Global Contrast Integration. After individually introducing the measurements for the differences between regions based on two visual cues, it is necessary to integrate them in an unified way. In this paper, the final measurement of contrast is defined as

$$D_r(r_i,r_j) = D_c(r_i,r_j) \cdot \exp[\sigma_e^2 \cdot D_e(r_i,r_j)]. \tag{5}$$

For the similar reason as defining Eq. (1) that $D_e(r_i,r_j)$ are found to be associated with a larger variation than $D_c(r_i,r_j)$, an exponential function is also employed here to normalize $D_e(r_i,r_j)$ with a parameter σ_e. σ_e is designed as the scaling factor for the exponential, and is experimentally fixed to $1/\sqrt{6}$ in all our experiments.

When the measurement of the contrast between two local regions is defined, the global contrast of a region to the entire scene can be determined by measuring its contrast to all other regions in the whole image. Furthermore, as recommended in [14,20], a spatial weighting term is introduced to incorporate spatial information into the definition of global contrast. This term can increase the effects of regions closer to the examined region. Specifically, for any region r_i, the spatially weighted global contrast is defined as

$$U(r_i) = \sum_{j \neq i} w_{ij} \cdot D_r(r_i,r_j) \cdot \phi_j, \tag{6}$$

$$w_{ij} = \frac{1}{Z_i} \cdot \exp[-\sigma_s^2 \cdot D_s(r_i,r_j)]. \tag{7}$$

Here $\phi_j = n_j$ is used to emphasize contributions of larger regions. $D_s(r_i,r_j)$ is the spatial distance between r_i and r_j. σ_s is employed to control the strength of spatial weighting w_{ij}. Larger values of σ_s will reduce the effect of w_{ij}, so that the farther regions can contribute more affections to the global contrast of r_i. σ_s^2 is set to *0.4* in all our experiments. Z_i is the normalization factor ensuing $\sum_{j \neq i} w_{ij} = 1$.

3.3. High-level information based global contrast

The most essential difference between the Tag-Saliency model and the existing works is the treatment for high-level information. In most of the previous works, high-level information is extracted by a series of established object detectors. Obviously, when there are generally hundreds of objects contained in the testing image set, the previous methods cannot effectively make use of the high-level information. Compared with these methods, the proposed Tag-Saliency model can fully utilize the semantic information contained in images and simultaneously process hundreds of images through the induction of image auto-tagging technique [15].

In the proposed model, the procedure of high-level information extraction is composed of five steps: (1) Describe each segmented region through color and texture descriptor. (2) Compute the Euclidean distance matrix. (3) Use the obtained region distance matrix to construct neighborhood matrix based on Simrank [50] calculation, which is a link-based similarity measure proposed in data ming works. The neighborhood range k is set to 200 according to [15]. (4) Automatically tag each over-segmented region through the prediction model obtained from training image set. (5) Obtain the semantic information of regions in the target segmentation level. This processing is implemented by predicting tag information for all regions in all segmentation levels firstly, and then adding the tag information of the lower level regions to the target segmentation level regions. For the high-level information extraction, the closest to our method are the works of [51,49]. In [51] the objects are represented by the "Visual language" modeled visual words, and in [49] the objects in the images are segmented and tagged by users. Differently, in this paper the semantic representation relies on natural words rather than visual words, and is conducted automatically.

When the tagging information of regions is available, the probability of a region being a component of foreground can be determined by referring to a keyword indicator. More specifically, the

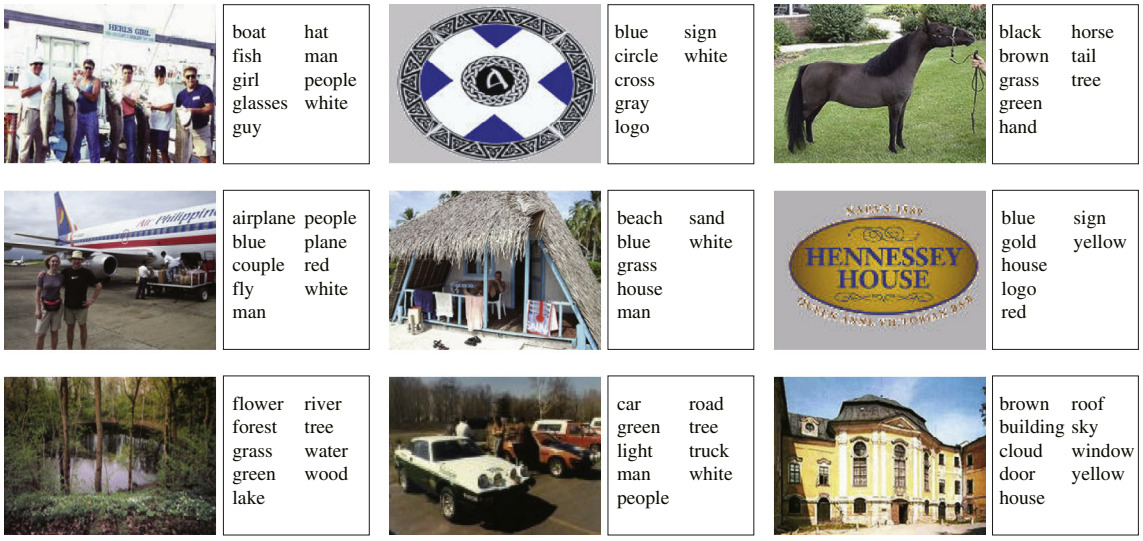

Fig. 3. Example images with the ground truth tags from the ESP-Game data set.

probability $T(r_i)$ of a region r_i being a component of foreground is defined as:

$$T(r_i) = \frac{Fg(r_i) - Bg(r_i)}{Neu(r_i)}, \tag{8}$$

where $Fg(r_i)$ denotes the maximum tagging value among the keywords that are always related with foreground (e.g., man, face, cat and flower), $Bg(r_i)$ denotes the maximum tagging value among the keywords that are always related with background (e.g., grass, lake, and sky), and $Neu(r_i)$ denotes the maximum tagging value among the keywords that are neutral (i.e., ambiguous to be foreground or background, such as blue, play, and music).

Once $T(r_i)$ has been obtained, there is a direct way to define the relation between this style of high-level information and saliency through global contrast analysis, that is

$$V(r_i) = \sum_{j \neq i} [T(r_i) - T(r_j)]^2, \tag{9}$$

where r_j is a over-segmented region of the same segmentation level with r_i. As $V(r_i)$ is depended on the semantic-level information rather than the direct visual perception, there is no spatial weighting term introduced to incorporate spatial information into the definition of this kind of global contrast.

4. Results

4.1. Image data set

In order to evaluate the performance of the proposed Tag-Saliency model, a very challenging public data set ESP-Game [52] is employed. This data set has been widely used in image auto-tagging and the keyword based image retrieval works. Each of the collected image is associated with an average of 5 keywords to describe the essential semantic information of the scene. All the labeled keywords are generated from an online ESP game. In this game, two participants cannot communicate in the game and will gain points if they use the same words to describe the image. There are a total of 268 unique words used in this data set. Some example images are presented in Fig. 3. Note that the ESP-Game is not constructed for saliency detection, therefore there are no ground truth saliency masks. In our experiments, a subset of *2500* images randomly sampled from the *60,000* images is employed, and 20 participants are invited to label the saliency masks. For each image, the common area covered by more than half of the salient regions labeled by the participants is treated as the ultimate ground truth salient region. In the employed subset, *2000* images are used for training, and the rest 500 images are used as testing set.

4.2. Evaluation measure

In all the experiments for quantitative analysis, the performance of the proposed method is evaluated by measuring its precision and recall rate. Precision measures the rate of correctly assigned salient regions to the whole detected region, while recall measures the percentage of positive detected salient regions in relation to the ground truth.

High recall can normally be achieved at the expense of the reduction in precision, and vice versa. Therefore, it is necessary to evaluate these two measures together. In this paper, a statistical precision-recall curve is employed to capture the trade-off between the accuracy and sensitivity. This cave can be sketched by varying the threshold used to generate the binary saliency maps. Here the employed thresholds are 21 fixed value, i.e., $[0{:}0.05{:}1] \times 255$. In addition to precision-recall curve, a weighted harmonic mean measure of precision and recall—F-measure [53], is also taken to provide a single index. To be specific, given the image with pixels $X = \{x_i\}$ and binary ground truth $G = \{g_i\}$, for any detected binary saliency mask $L = \{l_i\}$, these three indexes are defined as:

$$precision = \sum_i g_i l_i \Big/ \sum_i l_i, \tag{10}$$

$$recall = \sum_i g_i l_i \Big/ \sum_i g_i, \tag{11}$$

$$F_\beta = \frac{precision \times recall}{(1-\beta) \times recall + \beta \times precision}, \tag{12}$$

where β is set to 0.5 according to [53].

4.3. Performance

The results of the proposed method are compared with 14 state-of-the-art saliency detection methods. They are respectively the information maximization saliency (AIM [35]), adaptive whitening saliency (AWS [54,55]), context-aware saliency (CA [34]), frequency-tuned saliency (FT [33]), histogram based saliency (HC [14]), non-parametric low-level saliency (IM [56]), visual attention measurement (IT [13]), spatiotemporal saliency (LC [29]),

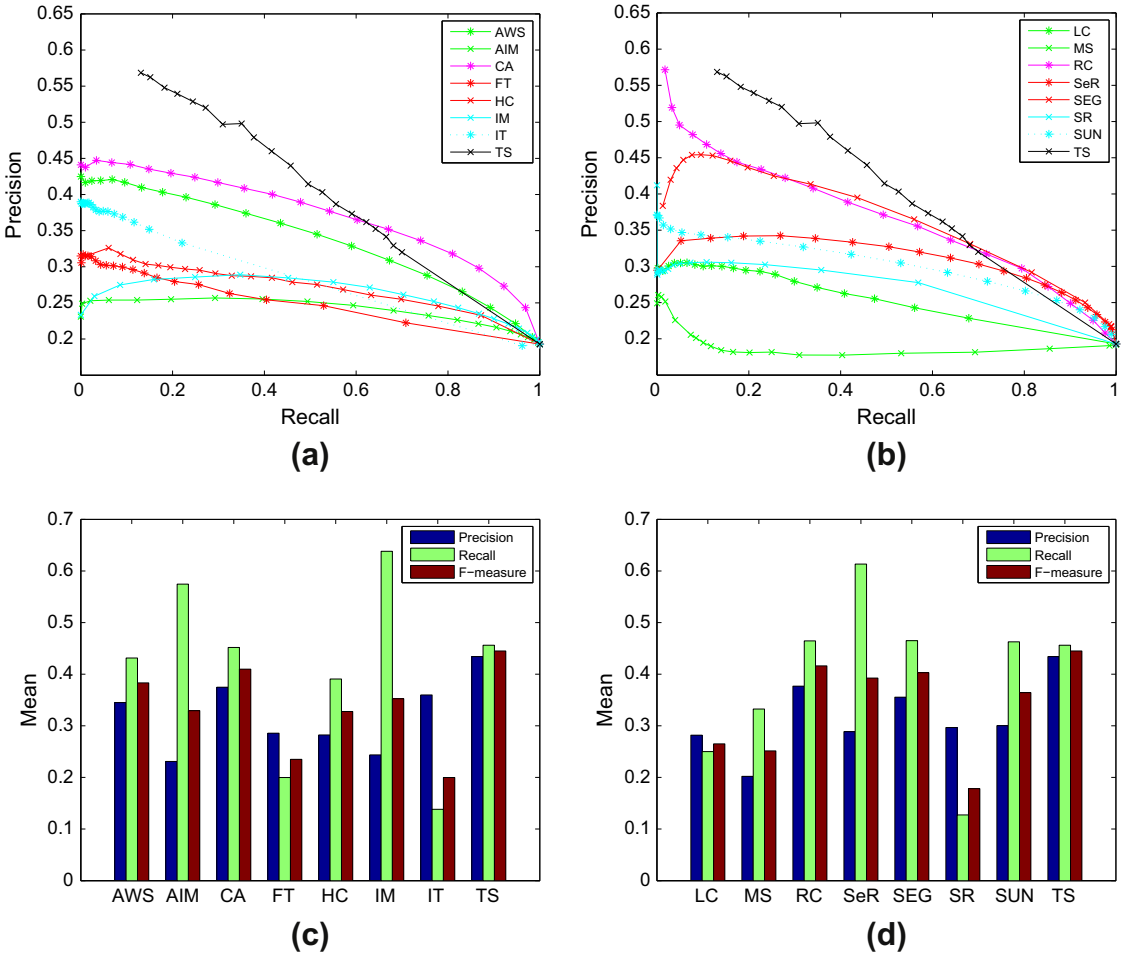

Fig. 4. Quantitative comparison between the Tag-Saliency (TS) model and state-of-the-art. (a) and (b) The precision-recall curves. (c) and (d) The averaged precision, recall, and F-measure bars.

multi-spectral saliency (MS [38]), region-based saliency (RC [14]), saliency segmentation (SEG [57]), self-resemblance saliency (SeR [28]), spectral residual saliency (SR [25]), and natural statistics saliency (SUN [27]).

Following [33,14], the methods selected here are based on 4 principles: prevalence (AIM, IT, and SR have been cited over 200 times), recency (AWS, CA, HC, IM, MS, RC, and SEG are proposed during the last two years), variety (FT, HC, LC, RC are global contrast based; SeR and SUN are local contrast based; IT and AWS are biologically inspired), and relevance (FT, RC). The codes for LC and SR are from Cheng et al. [14].[1] The implementation for IT is from a publicly available SaliencyToolbox[2] recommended in Itti's project webpage.[3] For other 11 selected methods, the codes are directly downloaded from the corresponding authors' homepages.

Fig. 4 illustrates the results. The precision-recall curves in Fig. 4(a) and (b) show that the proposed method clearly dominates AWS, AIM, FT, HC, IM, IT, LC, MS, and SR. These curves are sufficient enough to prove that the proposed method can locate salient regions with much more accuracy than these 9 competitors. Besides, the proposed method also outperforms CA, RC, SeR, SEG, and SUN most of the time, except with the disadvantage of lower precision rates at extremely high recall rates. However, in practice, the unilateral emphasis on the extremely high recall rate cannot lead to satisfying results. A moderation between the emphasis of precision and recall rate must be more appropriate [53]. In order to provide more discriminative clues for CA, RC, SeR, SEG, and SUN, the F-measure should be taken into account. As shown in Fig. 4(c) and (d), the proposed method dominates others in F-measure indicator.

Several visual comparison are also presented in Fig. 5 for qualitative evaluation. Notice that only the saliency maps of the top 7 of the 14 aforementioned methods are presented here, i.e., AWS, CA, IM, RC, SeR, SEG, and SUN. As can be seen in Fig. 5, the competitive 7 methods tend to highlight more non-salient locations, or produce morphological changed or internally incongruous salient regions in the maps, while the proposed method is prone to generate much more accurate and consistent results.

5. Discussion

As discussed in the previous Section 4, there is a significant advancement when employing the proposed measurement of global contrast based on both low-level and high-level information. But there are still some uncertainties worthy of further consideration. What kind of information considered in the proposed method plays an essential role in the significant advancement? What are the key factors that affect the high-level information extraction? This section will address these uncertainties, and give the quantitatively comparative analysis.

In order to further validate the effectiveness of the utilization of both low-level and high-level information, the complete

[1] http://cg.cs.tsinghua.edu.cn/people/ ~ cmm/.
[2] http://www.saliencytoolbox.net/.
[3] http://ilab.usc.edu/toolkit/downloads.shtml.

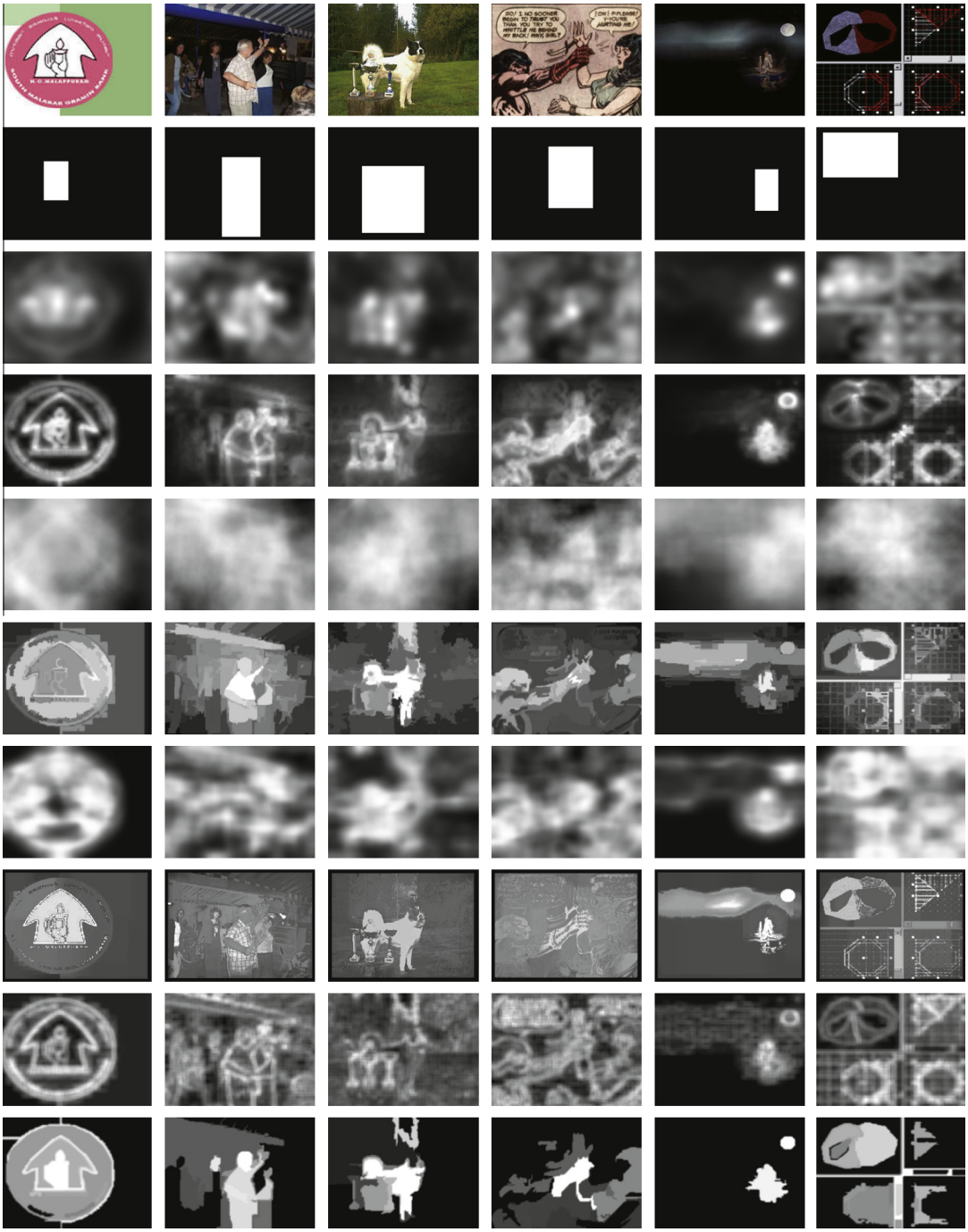

Fig. 5. Visual comparison of saliency maps. From top to bottom, each row respectively represents the original images, the ground truth masks, and the saliency maps calculated by AWS [54,55], CA [34], IM [56], RC [14], SeR [28], SEG [57], SUN [27], and the saliency maps calculated by the Tag-Saliency model.

Tag-Saliency model (TS) is compared with two restricted visions, i.e., based only on low-level (TS-L) or high-level (TS-H) information extraction. Fig. 6 demonstrates the results for all the 500 testing images. As can be seen from the corresponding precision-recall curves in Fig. 6(a), it is manifest that the complete Tag-Saliency model dominates the restricted vision TS-L. These curves indicate that, utilizing high-level information can help the Tag-Saliency model to locate salient regions more accurate than using only the low-level information. Besides, the TS outperforms TS-H most of the time. However, as can be seen in Fig. 6(a), when the tasks place more emphasis on achieving high recall rates, the precision-recall curves cannot provide discriminative clues for these two methods. In this case, as mentioned in Section 4.3, the F-measure should be also taken to provide more comparative information. As shown in Fig. 6(b), in F-measure indicator, the complete Tag-Saliency model clearly dominates others. Therefore, it is reasonable to believe that both the low-level and high-level information can play a significant role in saliency detection.

As for the aspect of high-level information extraction, there are two major factors to be discussed: (1) the number of hierarchical over-segmentation levels L, and (2) the neighborhood range k of regions used for tag prediction. There is an expectation that a greater number of segmentation level and neighborhood range will produce a better performance. However, it is necessary to make a trade-off between accuracy and computational complexity in order to derive the best strategy for practical saliency detection. The

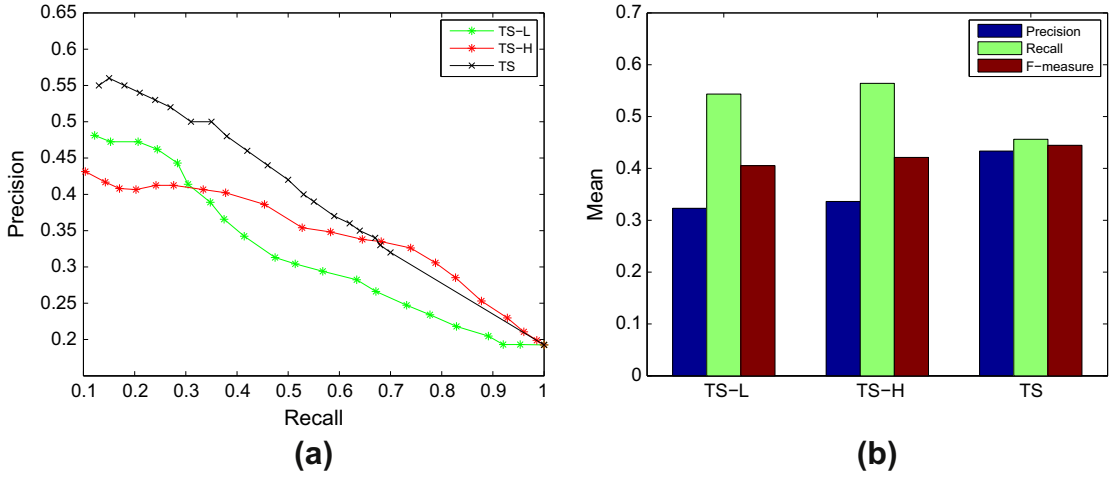

Fig. 6. Quantitative comparison between the complete Tag-Saliency model (TS) and two restricted visions (TS-L and TS-H). (a) The precision-recall curves. (b) The averaged precision, recall, and F-measure bars.

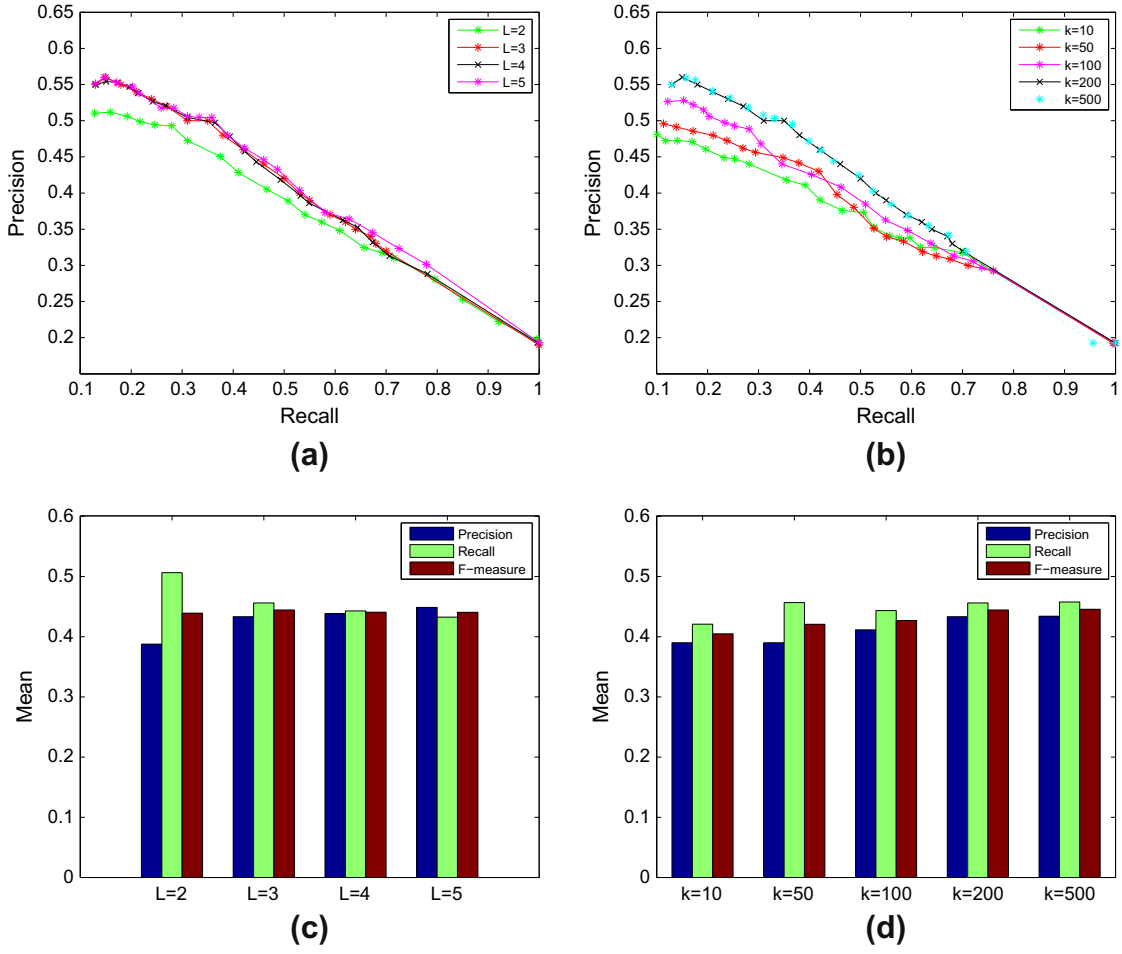

Fig. 7. Comparison of different settings of L and k. (a) and (b) The precision-recall curves. (c) and (d) The averaged precision, recall, and F-measure bars.

quantitative comparisons for different settings of the segmentation level are demonstrated in Fig. 7(a) and (c). It is manifest that setting $L = 3$ can already obtain a satisfactory result.

Then the comparison results under different settings of the neighborhood range k are presented in Fig. 7(b) and (d). As can be seen, the worst performances appear in the case that k is set to 10, and the performance will continue to improve until the value reaches 200. After that, higher k cannot visibly improve the performance, while the computational cost will be more intensive.

6. Conclusion

There is an intuition that the allocation of attention in natural scene viewing will be influenced by individuals' habits, i.e., people

always tend to pay more attention to some things, that are usually noteworthy in our daily life, and ignore others. Several visual cognition researches [11,12,47,41,43] can provide the theoretical basis for this intuition. However, detecting saliency in such a situation has not yet been well addressed in computer vision literatures. Traditional top-down saliency detection models utilize high-level information based on eye tracking or object detection techniques. But in practical situations, there may be no eye movement data available or too many objects to be handled in a large-scale data set. Therefore, these methods will be greatly limited by the harsh conditions of use and high conceptual complexities.

In this paper, a new Tag-Saliency model is designed specifically for the top-down attention prediction. This model can efficiently extract high-level semantic information from large scale visual media data by introducing the image auto-tagging technique, and can integrate the high-level and low-level information to measure the global contrasts of the hierarchical over-segmented regions. Experimental results on a very challenging data set show that, the proposed Tag-Saliency model has the ability to locate the truly salient regions in a greater probability than other competitors. A quantitative discussion is also presented in this paper, which indicates that the use of both kinds of information in the proposed model has played a significant role in the advancement of this model.

Acknowledgments

This work is supported by the State Key Program of National Natural Science of China (Grant No. 61232010), the National Natural Science Foundation of China (Grant No. 61172143 and 61105012), and the Natural Science Foundation Research Project of Shaanxi Province (Grant No. 2012JM8024).

References

[1] D. Walthera, U. Rutishausera, C. Kocha, P. Peronaa, Selective visual attention enables learning and recognition of multiple objects in cluttered scenes, Comput. Vis. Image Understand. 100 (2005) 41–63.
[2] Y. Yu, G.K.I. Mann, R.G. Gosine, An object-based visual attention model for robotic applications, IEEE Trans. Syst. Man Cybern. B: Cybern. 40 (2010) 1398–1412.
[3] A.K. Moorthy, A.C. Bovik, Visual importance pooling for image quality assessment, IEEE J. Select. Top. Signal Process. 3 (2009) 193–201.
[4] J. You, A. Perkis, M.M. Hannuksela, M. Gabbouj, Perceptual quality assessment based on visual attention analysis, in: ACM Int'l Conf. Multimedia, pp. 561–564.
[5] S. Marat, M. Guironnet, D. Pellerin, Video summarization using a visual attentional model, in: European Signal Processing Conf., pp. 1784–1788.
[6] Q. Wang, Y. Yuan, P. Yan, X. Li, Saliency detection by multiple-instance learning, IEEE Trans. Cybern. 43 (2013) 660–672.
[7] Q. Wang, Y. Yuan, P. Yan, Visual saliency by selective contrast, IEEE Trans. Circ. Syst. Video Tech. 27 (2013) 1150–1155.
[8] Y. Ma, H.-J. Zhang, Contrast-based image attention analysis by using fuzzy growing, in: ACM Int'l Conf. Multimedia, pp. 374–381.
[9] J. Han, K.N. Ngan, M. Li, H.-J. Zhang, Unsupervised extraction of visual attention objects in color images, IEEE Trans. Circ. Syst. Video Tech. 16 (2006) 141–145.
[10] C. Jung, C. Kim, A unified spectral-domain approach for saliency detection and its application to automatic object segmentation, IEEE Trans. Image Process. 21 (2012) 1272–1283.
[11] Y. Carmi, L. Itti, Visual causes versus correlates of attentional selection in dynamic scenes, Vision Res. 46 (2006) 4333–4345.
[12] B. Tatler, The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions, J. Vision 14 (2007) 1–17.
[13] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 21 (1998) 1254–1259.
[14] M. Cheng, G. Zhang, N.J. Mitra, X. Huang, S. Hu, Global contrast based salient region detection, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 409–416.
[15] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation, in: Int'l Conf. Computer Vision, pp. 309–316.
[16] Y. Yang, Y. Yang, Z. Huang, H.T. Shen, F. Nie, Tag localization with spatial correlations and joint group sparsity, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 881–888.
[17] Y. Yang, Z. Huang, H.T. Shen, X. Zhou, Mining multi-tag association for image tagging, World Wide Web 14 (2011) 133–156.
[18] L. Wu, R. Jin, A.K. Jain, Tag completion for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 716–727.
[19] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 353–367.
[20] F. Perazzi, P. Kráhenbühl, Y. Pritch, A. Hornung, Saliency filters: contrast based filtering for salient region detection, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1–8.
[21] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human Neurobiology 4 (1985) 97–136.
[22] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, C. Koch, Attentional selection for object recognition-a gentle way, in: Biologically Motivated Computer Vision, pp. 472–479.
[23] D. Gao, V. Mahadevan, N. Vasconcelos, On the plausibility of the discriminant center-surround hypothesis for visual saliency, J. Vision 8 (2008) 1–18.
[24] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Advances in Neural Information Processing Systems, pp. 545–552.
[25] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1–8.
[26] R. Achanta, F.J. Estrada, P. Wils, S. Süsstrunk, Salient region detection and segmentation, in: Computer Vision Systems, pp. 66–75.
[27] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell, Sun: a bayesian framework for saliency using natural statistics, J. Vision 8 (2008) 1–20.
[28] H.J. Seo, P. Milanfar, Nonparametric bottom-up saliency detection by self-resemblance, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 45–52.
[29] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in: ACM Int'l Conf. Multimedia, pp. 815–824.
[30] L. Shao, M. Brady, Specific object retrieval based on salient regions, Pattern Recogn. 39 (2006) 1932–1948.
[31] L. Shao, M. Brady, Invariant salient regions based image retrieval under viewpoint and illumination variations, J. Visual Commun. Image Represent. 17 (2006) 1256–1272.
[32] L. Shao, T. Kadir, M. Brady, Geometric and photometric invariant distinctive regions detection, Inform. Sci. 177 (2007) 1088–1122.
[33] R. Achanta, S. Hemami, F. Estrada, S. Süsstrunk, Frequency-tuned salient region detection, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1597–1604.
[34] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 2376–2383.
[35] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Advances in Neural Information Processing Systems, pp. 155–162.
[36] M. Wang, J. Konrad, P. Ishwar, K. Jing, H. Rowley, Image saliency: from intrinsic to extrinsic context, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 417–424.
[37] H. Li, K.N. Ngan, A co-saliency model of image pairs, IEEE Trans. Image Process. 20 (2011) 3365–3374.
[38] Q. Wang, P. Yana, Y. Yuana, X. Li, Multi-spectral saliency detection, Pattern Recogn. Lett. 34 (2013) 34–41.
[39] C. Lang, G. Liu, J. Yu, S. Yan, Saliency detection by multitask sparsity pursuit, IEEE Trans. Image Process. 21 (2012) 1327–1338.
[40] A. Borji, Boosting bottom-up and top-down visual features for saliency estimation, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1–8.
[41] W. Einhäuser, M. Spain, P. Perona, Objects predict fixations better than early saliency, J. Vision 8 (2008) 1–26.
[42] M. Cerf, J. Harel, W. Einhäuser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, in: Advances in Neural Information Processing Systems, pp. 241–248.
[43] L. Elazary, L. Itti, Interesting objects are visually salient, J. Vision 8 (2008) 1–15.
[44] R. Peters, L. Itti, Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition.
[45] T. Judd, K.A. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: Int'l Conf. Computer Vision, pp. 2106–2113.
[46] A. Borji, D.N. Sihite, L. Itti, Probabilistic learning of task-specific visual attention, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1–8.
[47] V. Navalpakkam, L. Itti, Modeling the influence of task on attention, Vis. Res. 45 (2005) 205–231.
[48] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vis. 59 (2004) 167–181.
[49] L. Wu, S.C.H. Hoi, N. Yu, Semantics-preserving bag-of-words models and applications, IEEE Trans. Image Process. 19 (2010) 1908–1920.
[50] G. Jeh, J. Widom, Simrank: A measure of structural-context similarity, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 538–543.
[51] L. Wu, Y. Hu, M. Li, N. Yu, X.-S. Hua, Scale-invariant visual language modeling for object categorization, IEEE Trans. Multimedia 11 (2009) 286–294.
[52] L. von Ahn, L. Dabbish, Labeling images with a computer game, in: ACM SIGCHI, pp. 319–326.
[53] D.R. Martin, C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color, and texture cues, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 530–549.

[54] A. Garcia-Diaz, V. Leborán, X. Fdez-Vidal, X. Pardo, On the relationship between optical variability visual saliency and eye fixations: a computational approach, J. Vision 12 (2012) 1–22.
[55] A. Garcia-Diaz, X. Fdez-Vidal, X. Pardo, R. Dosil, Saliency from hierarchical adaptation through decorrelation and variance normalization, Image Vis. Comput. 30 (2012) 51–64.
[56] N. Murray, M. Vanrell, X. Otazu, C.A. Parraga, Saliency estimation using a non-parametric low-level vision model, in: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 433–440.
[57] E. Rahtu, J. Kannala, M. Salo, J. Heikkilä, Segmenting salient objects from images and videos, in: European Conference on Computer Vision, pp. 1–14.

Guokang Zhu is currently working toward the Ph.D. degree in the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include computer vision and machine learning.

Qi Wang received the B.E. degree in automation and Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2005 and 2010 respectively. He is currently an associate professor with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include computer vision and pattern recognition.

Yuan Yuan is a full professor with the Chinese Academy of Sciences (CAS), China. She has published over 100 papers, including about 70 in reputable journals such as IEEE transactions and Pattern Recognition, as well as conferences papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.