NATAS: Neural Activity Trace Aware Saliency

Guokang Zhu, Qi Wang, and Yuan Yuan, *Senior Member, IEEE*

***Abstract*—Saliency detection has raised much interest in computer vision recently. Many visual saliency models have been developed for individual images, video clips, and image pairs. However, image sequence, one most general occasion in the real world, is not explored yet. A general image sequence is different from video clips whose temporal continuity is maintained and image pairs where common objects exist. It might contain some similar low-level properties while completely distinct contents. Traditional saliency detection methods will fail on these general sequences. Based on this consideration, this paper investigates the shortcomings of the classical saliency detection methods, which significantly limit their advantages: 1) inability to capture the natural connections among sequential images, 2) over-reliance on motion cues, and 3) restriction to image pairs/videos with common objects. In order to address these problems, we propose a framework that performs the following contributions: 1) construct an image data set as benchmark through a rigorously designed behavioral experiment, 2) propose a neural activity trace aware saliency model to capture the general connections among images, and 3) design a novel measure to handle the low-level clues contained among sequential images. Experimental results demonstrate that the proposed saliency model is associated with a tremendous advancement compared with traditional methods when dealing with the general image sequence.**

***Index Terms*—Computer vision, global contrast, machine learning, neural activity trace, preactivation, saliency detection, visual attention.**

I. Introduction

VISION is the most important component of the human sensory system, which can provide an intuitive way for us to understand the world. In the human visual system, there is an effective attention selection mechanism. This mechanism can drive the observers to allocate the limited perceptual processing resources to the most important visual subsets [1], [2]. For instance, when we open the window and look out, we will unconsciously see the cars, pedestrians, or other objects, while we will easily ignore the things usually treated as background. Visual attention selection is considered as a matter of course for human, but inspires a challenging task in computer vision—saliency detection.

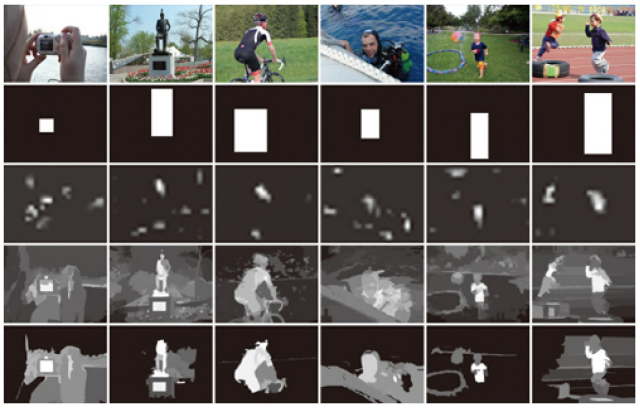


Fig. 1. Saliency detection. First row: original image sequence. Second row: ground truth labels of the corresponding images. Remaining rows: saliency maps calculated by IT [11], RC [12], and the proposed model, respectively.

Saliency detection basically aims at providing the computational identification of scene regions, which are more notable to a human observer than their surroundings. Typical examples of saliency detection are demonstrated in Fig. 1. Reliable saliency maps can provide a lot of useful information for further processing without prior knowledge about the scene. For example, it has been used for content-aware image scaling [3], [4], image segmentation [5], [6], object recognition [7], [8], and smart video presentation [9], [10]. Therefore, modeling visual saliency is considered as an important component in computer vision, and shows an increasing interest in both theory and practice recently.

Typical occasions of saliency detection are performed on individual images [13]–[15], video clips [16]–[18], and image pairs [19]. Besides these situations, there is another one that still has not been realized, i.e., an image sequence that might be displayed successively or observed one by one with a very short time interval. In this case, the processing is different from traditional models, because the information for saliency calculation includes not only color, texture, motion, and gradient, but also the connections among images. According to a psychophysical evidence [20], when physical stimulus excite a recipient within very short time intervals, the effects of the previous stimulus can influence the responses aroused by the subsequent stimulus through the biochemical traces of neural activities. There are many typical examples of this case in the real world, such as browsing photo albums, looking over Google image search results, and viewing products in online

Manuscript received December 4, 2012; revised July 15, 2013; accepted August 6, 2013. Date of publication December 11, 2013; date of current version June 12, 2014. This work is supported in part by the State Key Program of National Natural Science of China under Grant 61232010, in part by the National Natural Science Foundation of China under Grant 61172143 and Grant 61105012, and in part by the Natural Science Foundation Research Project of Shaanxi Province under Grant 2012JM8024. This paper was recommended by Associate Editor F. Hoffmann.

The authors are 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 710119, China (e-mail: zhuguokang@opt.ac.cn; crabwq@opt.ac.cn; yuany@opt.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2013.2279002

Fig. 2. Example of the occasion focused in this paper, where the images in the sequence will be observed one by one.

stores. Fig. 2 demonstrates an intuitive example. For this new occasion, the key point is how to capture the effect of the previous stimulus.[1] The problem addressed in this paper is the modeling of visual saliency on a sequence of images, considering the neural activity traces among images.

A. Related Work

Classical saliency detection methods tackle visual attention as a bottom-up process without prior knowledge, and choose to employ a low-level approach to calculate local or global contrasts of image regions with respect to their surroundings. According to the occasion they targeted, these methods can be roughly classified into three groups: separate-scene saliency, dynamic saliency, and cosaliency.

Works that belong to the first group evaluate saliency of image regions by calculating the contrasts with respect to local neighborhoods or the entire images. They limit their focus only on each individual image, instead of other reference ones. Many of these methods are related with the biologically plausible model of human visual system introduced in [21]. For example, inspired by this model, Itti *et al.* [11] first extracted multiscale features using a difference of Gaussians (DoG) approach, and then, defined saliency by combing and normalizing these features through a center-surround difference scheme. Walther *et al.* [22] modified the method in [11] with a hierarchical recognition system. Han *et al.* [5] identified attention seeds mainly by the method in [11], and then extended these seeds to regions by a Markov random field (MRF) model. Harel *et al.* [23] designed a graph-based method, which can, first, form activation maps by normalizing the feature maps of [11] as well as other importance maps, and then, combine them to highlight conspicuous parts.

Besides the above-mentioned methods, many other local contrast-based saliency detection methods have been proposed for separate scenes, which are related strongly with purely local analysis of image neighborhoods. Achanta *et al.* [24] calculated local contrast based on a sliding local window, by which the Euclidean distance between the average feature vectors of the inner subregion and its outer neighborhoods is calculated as the saliency value for each location. Gao *et al.* [25] measured contrast on the histograms of a series of DoG and Gabor filter responses, and determined saliency of a location as the Kullback–Leibler (KL) divergence between this location and its surrounding region. Hou and Zhang [26] introduced a spectral residual model relying on frequency domain processing. They used the difference between the log Fourier spectrum of an image and its locally averaged version to find the innovation locations. Zhang *et al.* [27] evaluated saliency using Shannon's self-information and pointwise mutual information under a Bayesian framework. Seo and Milanfar [28] measured saliency based on local steering kernels (LSK), which can capture the gradient contrast between the examined location and its surrounding region.

Recently, global contrast-based methods, which take into account global relations over the entire image, have announced promising results on separate scenes. For example, Achanta *et al.* [29] presented a frequency tuned algorithm, which can achieve globally more consistent saliency maps by computing the global contrast between each pixel color and the average color of the Gaussian-filtered image. Goferman *et al.* [30] measured saliency based on global dissimilarities of each image patch compared with the corresponding k most similar patches in the whole image. Bruce and Tsotsos [31] detected saliency based on the maximum information over the complete scene, and calculated the probability density function based on a Gaussian kernel density estimate (KDE) in a neural circuit. Cheng *et al.* [12] proposed a region-level saliency detection method based on the global contrasts between histograms of over-segmented regions. Wang *et al.* [32] tackled saliency as anomaly in an image relative to a large image dictionary through k-nearest-neighbor (kNN) retrieval. Liu *et al.* [16], [33] defined a global feature describing the color spatial distribution, and enhanced it with the local center-surround histogram and the multiscale contrast to extract prominent regions through conditional random field (CRF) learning. More recently, Perazzi *et al.* [13] reconsidered some previous excellent global contrast-based features [12], [16] in a unified way using Gaussian filters.

Dynamic Saliency detection methods are generally also constructed on the basis of local or global analysis of contrast, but have the additional capacity to utilize motion cues. Zhai and Shah [17] proposed to utilize motion contrast based on the interest point correspondences and the geometric transformations between consecutive images. Rahtu *et al.* [18] proposed a saliency measure method based on a statistical framework, which utilizes local feature contrast in illumination, color, and motion information in a sliding window. Liu *et al.* [16] extended their separate-scene saliency detection method to detect a salient object from video clips by introducing two dynamic salient features, i.e., the motion salient feature and the appearance coherence feature. The motion salient feature is defined based on the motion fields obtained by the SIFT flow technique [34], while the second feature models the appearance coherence of the salient objects between two successive frames.

[1]The stimulus in this paper are the images observed by subjects.

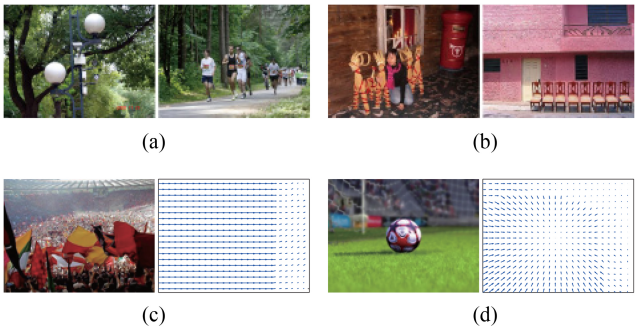


Fig. 3. Negative cases for exciting methods. (a) Two images displayed with a very short time gap. (b) Image pair without foreground objects in common. (c) Motion map with the same speed for the whole scene caused by lens shifting. (d) Motion map with the radiative increased speed for the scene caused by lens zooming.

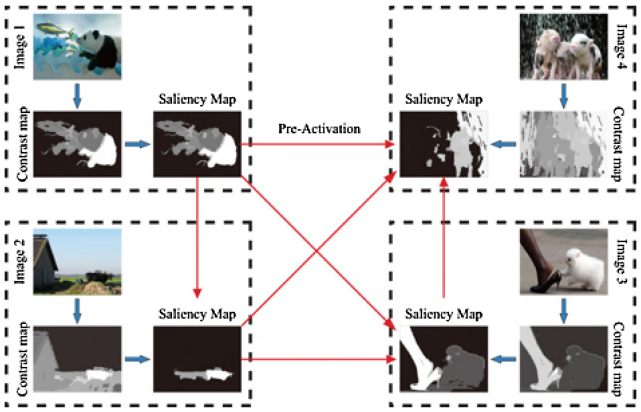


Fig. 4. Summary of NATAS, which can capture the preactivation caused by the biochemical traces during image sequence displaying.

As for image pairs with some foreground objects in common, a cosaliency model [19] has been proposed. This model is constructed as a linear combination of the saliency maps from the separate-scene saliency feature [11], [26], and [29] and the pair-image saliency feature, which is generated based on comultilayer graph construction and normalized Simrank [35] similarity computation.

B. Limitations of Existing Methods

Though various saliency detection methods have been presented in the past few years, and a laudable performance for predicting human attentional spotlight has been achieved in separate scenes, video clips, and image pairs, there are still several limitations for extending these methods to successively displayed image sequences.

Separate-scene saliency detection methods are designed for the separate scenes. These methods consider the information obtained only in the current image, while ignoring the connections among the examined image and the previously displayed ones. However, it is proved that the previous stimulus can influence the response aroused by the subsequent stimulus [20]. Take Fig. 3(a) for example. If these two images are successively displayed with a very short time interval, the observer's attention in the second image is more likely to be allocated on the men in the white shirt, which is consistent with the foreground white lampshades of the first scene according to its appearance and location.

Dynamic saliency detection methods are suitable for the circumstance where the relative velocity can be observed easily between the salient object and the background, i.e., the salient object is with the motion magnitude significantly larger or smaller than the background. However, in the occasion focused in this paper, there is typically neither a common salient object nor a similar background in the adjacent two images. Thus, no motion field can be obtained. Besides, as shown in Fig. 3(c) and (d), even when two adjacent frames present the same scene, there are still many negative cases, where there is no significantly relative motion existing between the salient object and background.

Cosaliency model is based on an assumption that the common objects in image pairs are more likely to capture observer's attention. This model, therefore, works well only when an image pair has some common foreground objects or at least very similar foreground objects. But in most cases, as shown in Fig. 3(b), the strict condition is hardly satisfied. The most common circumstance for saliency detection is that scenes are with some similar low-level properties, e.g., the presence of some local similarities in color, texture, or shape, instead of some objects in common.

C. Overview

To deal with the new occasion: 1) an image data set containing sequences of images and ground truth labels is constructed to serve as a benchmark platform in this paper, and 2) a neural activity trace aware saliency (NATAS) model is developed, which can capture the effect of neural activity trace during image sequence displaying.

In order to construct a reliable data set, the collected images are divided into blocks. The images within a block possess connections between them, which makes the research toward image sequence possible. Each image is then labeled by 20 subjects through a rigorously designed behavioral experiment. As some methods simultaneously consider the salient regions over scales, this data set also has the characteristic that it contains images with both large and small salient regions.

The second contribution of this paper is the NATAS model suitable for the new occasion. Fig. 4 shows the flowchart. This model is motivated by the need for overcoming the limitations of exiting methods and takes advantage of two components.

1) The global analysis of contrast is employed in NATAS based on color and visual complexity cues available in each examined image. Analyzing the global contrast has both theoretical and practical bases. The psychological discovery of Chen *et al.* [36] and Zhou *et al.* [37] reveals that the global-first topological perception completely dominates early vision. Experimental experiences also indicate that the global contrast-based models can often connect with consistently outstanding performance in practice [12], [13], [29], [38].
2) The effect of the biochemical trace of neural activity is considered in NATAS and defined as preactivation. The existence of such effect is proved in our

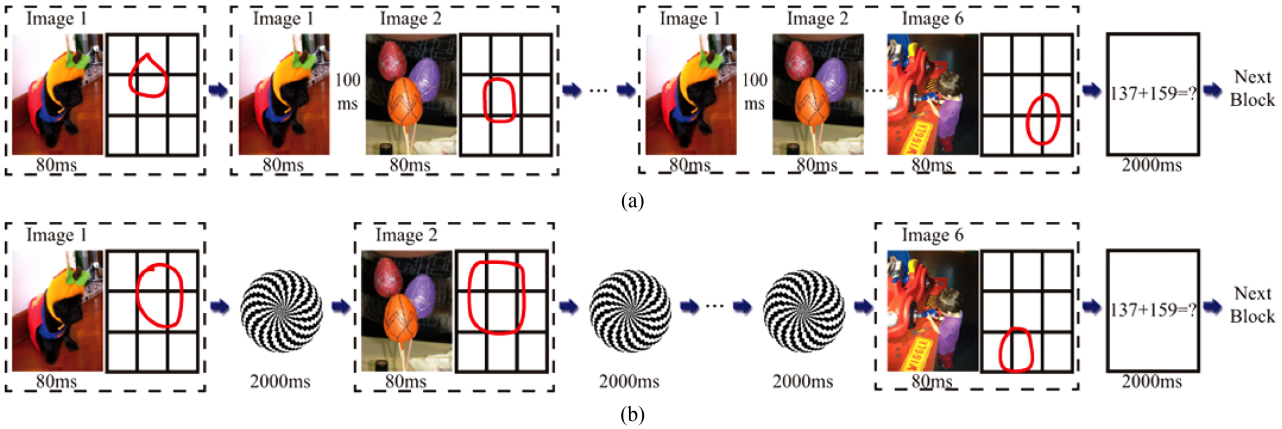


Fig. 5. Flow chart for one block of the behavioral experiment. (a) Treatment group. (b) Control group.

behavioral experiment, and the treatment of preactivation is the most essential difference between NATAS and the existing works.

The rest of this paper is organized as follows. Section II details the behavioral experiment designed to construct the data set. Section III introduces the proposed mode. Section IV presents the extensive experiments to prove the effectiveness of the proposed model. Section V analyzes what is the key factor for the proposed model, and the conclusion follows in Section VI.

II. Neural Activity Trace in Attention Allocation

Although there is an intuition that the foci in previous images will effect the attention distributions in the subsequent images, implementing this effect has not been addressed in the literature. Is this phenomenon true only for one person or consistently the same for anyone else? It still remains to be answered. In this section, we will detail a rigorously designed behavioral experiment to clarify such a phenomenon in the context of saliency detection. At the same time, a new data set is constructed as a benchmark for further research.

A. Behavioral Experiment

The previous observations of images will influence the judgement of later concentration. This is also true for saliency detection and we call the effect as neural activity trace in this paper. To justify this point, a comparative behavioral experiment is conducted to prove the rightness. Detailed introduction is presented below.

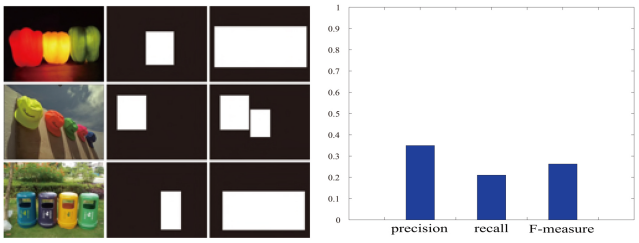
Firstly, a data set of 240 images is constructed. These images have a wide variety of contents and are divided into 40 blocks. Each block contains six images, with similar color or shape for salient objects. This makes intuitive connections among them. Secondly, 40 subjects are invited in the behavioral experiment. These 40 participants are randomly divided into two groups. Half of them are assigned to treatment group, which aims to construct the saliency masks for the sequential occasion. The remaining half are assigned to control group, who are set to collect the saliency masks for individual images.[2] The main difference between these two groups is the ways for presenting images in the next step. After that, all the 40 participants are asked to sit in front of a 19″ LCD screen and keep a distance of three times the screen width. Images are then presented to them and they are, respectively, requested to label the salient objects with a rough drawing.

Fig. 5 shows the flow chart for the two experimental groups. In both cases, images are presented block by block. But the time intervals between adjacent presentations and the number of images presented at one time are different. For the treatment group, suppose I_e is the examined image requiring to be labeled in a block. It is displayed for a short time duration of 80 ms and then the salient area is immediately labeled in a drawing board by the participant. However, before displaying I_e, each of the previous labeled images in the same block will rapidly flash for 80 ms again, with a 100 ms gap between different image presentations. Once I_e has been labeled, the participant has to press the space button of the keyboard to view the next image in this block. If this block finishes, the next block starts. Between two block presentations, a simple mathematical problem will be displayed for 2000 ms. The participant is asked to figure out the raised question within the time interval. Through this arrangement, the participant's memory trace is flushed out and the next block will not be influenced by the previous block.

As for the control group, the general procedure is similar, except that its presentation procedure is image by image. Every time, only the current image is displayed and no previous ones appear. The interval between adjacent presentations is 2000 ms and during this time interval, there is an interfering image displaying on the screen to ensure no neural activity traces on the next image.

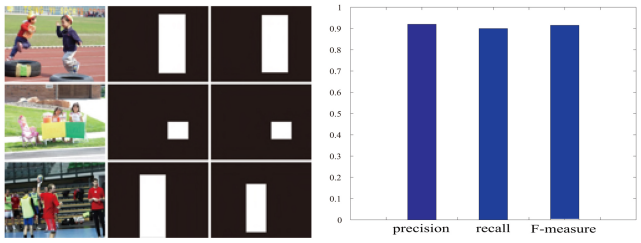
The time intervals (80 ms, 100 ms, and 2000 ms) set above are not random. Instead, they are determined according to the psychological principles. Evidences in psychology reveal that

[2]The concept of treatment group and control group is borrowed from psychophysical and medical terms. A control group is treated as a baseline measure. The control group has identical experimental items that you are examining, with the exception that it does not receive the treatment or the experimental manipulation that the treatment group receives. Please refer to http://en.wikipedia.org/wiki/Treatment_and_control_groups and http://www.ncsu.edu/labwrite/il/controltreatmentgr.htm.



(a) (b)

Fig. 6. Comparison for the saliency masks collected in different groups. (a) Sample images (left) with saliency masks from treatment group (middle) and control group (right). (b) Averaged precision, recall, and F-measure bars.



(a) (b)

Fig. 7. Comparison for the saliency masks collected in different half of the treatment group. (a) Sample images with two kinds of saliency mask. (b) Averaged precision, recall, and F-measure bars.

the human visual system can achieve the bottom-up attention within a very short duration (80 ms) [39], and the visual sensory memory (or iconic memory) can last only a brief amount of time (250–1000 ms) [40]. Therefore, the time intervals set for the treatment group can ensure that the neural activity traces, which are aroused by the previous images of the same block, do not completely subside. For the control group, the 2000 ms interval between neighboring images is long enough to guarantee that the adjoining images do not influence each other.

It should be noted that each block in our experiment consists of six images. This setting is based on a psychophysical discover, which reveals that working memory typically holds three or four items at once [41]. Besides, as the masks drawn by participants are rough irregular shapes, a postprocessing program is applied to them after the completion of the behavioral experiment. In this postprocessing, each salient region labeled by a participant is reshaped by a rectangular box, which covers the entire marked region with the minimum area. Since each image is labeled by multiple participants, to get a consistent ground truth, the common area covered by more than half number of rectangles of each image in treatment group is treated as the ultimate saliency mask for sequential occasion. The same procedure is implemented in the control group to obtain the saliency masks for individual images.

B. Comparative Analysis

There are visible differences between the saliency masks obtained from two experimental groups, one of which follows traditional labeling paradigm that treats images individually and the other considers the previous influence on current example. These differences can be seen from Fig. 6(a), where the treatment group and control group generate two kinds of saliency masks. Before further comparative analysis for these two groups, quantitative measures should be introduced first. Three indexes of precision, recall, and F-measure are employed in this paper. These indexes have achieved great popularity in saliency detection [42] and other information retrieval community. To be specific, given an image with pixels $X = \{x_i\}$ and a reference binary mask (ground truth) $G = \{g_i \in [0, 1]\}$, for any other mask $L = \{l_i \in [0, 1]\}$ to be evaluated, these three indexes are defined as

$$precision = \sum_i g_i l_i \Big/ \sum_i l_i \tag{1}$$

$$recall = \sum_i g_i l_i \Big/ \sum_i g_i \tag{2}$$

$$F_\beta = \frac{precision \times recall}{(1-\beta) \times recall + \beta \times precision} \tag{3}$$

where β is set to 0.5 according to [43].

With these three metrics, quantitative analysis is conducted to evaluate the similarity between the labeled saliency results of the two groups. The saliency maps from control group are treated as references (ground truths). The results from treatment group are to be evaluated. According to the above introduced measures, averaged precision, recall and F-measure can be obtained. It is clear from Fig. 6(b) that the employed three metrics are all at a low value, which indicate their results differ greatly (if their results are identical with each other, the values should be one.). These further tell us that saliency detection based on image sequence does not equal to saliency detection of its component individual image.

C. Consistency Analysis

Another question that should be addressed is the consistency among different participants' responses. This can ensure the obtained statistics are reliable and reasonable. To evaluate the consistency, the treatment group is split into two independent halves. Then, we identify how well the saliency masks labeled by the first half of the participants can match those obtained from the second half. Some visual comparisons are presented in Fig. 7(a) for qualitative evaluation. Besides, the average values of precision, recall, and F-measure are presented in Fig. 7(b) to provide the quantitative results.

The comparison results indicate that the two groups of saliency masks are highly consistent in appearances, and the average values of precision, recall, and F-measure are significantly high. All these results together can prove that the collected saliency masks are highly consistent. This suggests that, using these saliency masks as the ground truth, is appropriate and the obtained statistics are reliable and reasonable. The same analysis is conducted on the control group and there follows a similar result.

D. Summary

In previous subsections, we have constructed a benchmark dataset containing blocks of image sequences, and conducted a behavioral experiment to prove that detecting saliency on

image sequence does not equal to treating the images individually. This is because there are neural activity traces among the sequential images. If this connection is cut off by treating the images separately, the obtained result will be different from what it should be. Traditional methods are not suitable for this new occasion because they do not consider the neural activity traces. To get a more reasonable saliency detection result, the problem of image sequence is explored in this paper. Detailed discussion will be presented in the next section.

III. NATAS Model

This section details the NATAS model specifically designed for tackling image sequences. This model detects saliency at region level, i.e., estimates the probability of each over-segmented region being salient, according to: 1) the global contrast in the scene, as well as 2) the preactivation caused by those salient regions in the previous images.

A. Preprocessing

For the preprocessing, an excellent over-segmentation method [44] is employed to segment the image into distinct regions. Since the Gestalt movement in psychology, it is widely recognized that perceptual grouping is fundamental to the process of human visual perception [19], [45]. Some experimental research on early visual process [36], [39] also proved that the first information that can be distinguished by the biological vision systems is the rough shape of object. Therefore, our saliency detection is conducted on the region-level instead of the pixel-level. It is reasonable to believe that this preprocessing is consistent with the characteristics of human visual system.

B. Saliency Assignment

As mentioned before, in the new saliency detection occasion, the attention allocated to a region is dependent on the global contrast in the scene, as well as the similarities between this region and those salient regions in previous images. Therefore, in NATAS, the saliency value $S(r_{i,k})$ assigned to region $r_{i,k}$, the ith segmented area in the kth image of a sequence, is defined as

$$S(r_{i,k}) = U(r_{i,k}) \cdot \exp[\sigma_v^2 \cdot V(r_{i,k})] \cdot \frac{1}{Z_k^{(s)}} \tag{4}$$

where the first component $U(r_{i,k})$ denotes the global contrast between region $r_{i,k}$ and others from the same scene, while the second part $V(r_{i,k})$ denotes the preactivation degree between $r_{i,k}$ and those salient regions in the previous images detected before. $Z_k^{(s)}$ is the normalization factor, which linearly projects the saliency values to the range [0,1].

In image sequences, the preactivation is observed of higher significance and with more discriminative power. Therefore, in this formulation, an exponential function is employed to emphasize $V(r_{i,k})$ with a parameter σ_v. σ_v is a scaling factor that controls the strength of $V(r_{i,k})$, and is experimentally fixed to two. Besides, it should be noted that, since there is no neural activity traces for the first image in a sequence, $\exp[\sigma_v^2 \cdot V(r_{i,k})]$ are fixed to one for all the corresponding regions.

C. Global Visual Contrast

It is widely believed that the visual system preferentially responses to the high contrast stimulus [12], [46]. In this paper, two visual cues, color and visual complexity, are used for global contrast analyzing.

Color-based contrast: In existing region-based saliency detection methods [12], [13], [19], the color contrast between two regions is measured by directly calculating: 1) the Euclidean distances of colors in all positions, or 2) the chi-square distance of color histograms. Assume there are N pixels and M regions in an image, and Z dimensions in color histogram. The first operation will take $O(N^2)$ time, which is computationally too expensive for a common web image. The second operation will take $O(N+Z\times M^2)$ time, which greatly reduces the computation if Z is not a large number. However, a small Z has the disadvantage that the discriminative ability of color is severely reduced. Therefore, a more efficient but statistically convincing measurement will be beneficial. Since the distributions of colors are independent in image regions, and are truly not normal most of the time [47], [48], the Student t-value is employed in this paper. More specifically, the color contrast between region r_i and r_j is defined as

$$D_c(r_i, r_j) = \|\mu_i - \mu_j\|^2 \cdot \left(\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}\right)^{-\frac{1}{2}} \tag{5}$$

where μ_i and μ_j denote the average colors of region r_i and r_j, respectively, σ_i^2 and σ_j^2 are their variances, and n_i and n_j are the total numbers of pixels in the corresponding regions. This measurement takes $O(N + M^2)$ time. As an image will be segmented to only around 20 regions in our paper, the computational efficiency is improved to approximately $O(N)$.

Visual complexity-based contrast: Visual complexity is another important cue that could play an important role in guiding human attention [31]. For this visual cue, a measure derived from information theory is employed in our paper, i,e, entropy. Based on entropy estimation for color distribution, the contrast of visual complexity $D_e(r_i, r_j)$ between region r_i and r_j can be simply defined as

$$D_e(r_i, r_j) = [H(r_i) - H(r_j)]^2 \tag{6}$$

where $H(r_i)$ represents the entropy of region r_i, which can be easily estimated by

$$H(r_i) = \sum_{p=1}^{n_{c,i}} f(c_{p,i}) \cdot \log_2 f(c_{p,i}) \tag{7}$$

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

Contrast integration: As a necessary step, a final measure for the contrast between two regions should be defined based on the employed visual cues. We assume that the measures for the contrast of color and visual complexity are independent; hence, they could be efficiently integrated as

$$D_r(r_i, r_j) = D_c(r_i, r_j) \cdot \exp[\sigma_e^2 \cdot D_e(r_i, r_j)]. \tag{8}$$

As $D_e(r_i, r_j)$ is found to be associated with a larger range of values 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, where σ_e is experimentally fixed to $1/\sqrt{6}$.

Spatially weighted global contrast: After defining the contrasts between local regions, the global contrast of each region can be obtained by calculating its contrast to all other regions in the whole image. In addition, as stated in [12] and [13] that spatial relationship is also an important factor in saliency detection, therefore, a spatial weighting term is also introduced here. More specifically, the spatially weighted global contrast for region r_i is defined as

$$U(r_i) = \sum_{j \neq i} w_{ij}^{(u)} \cdot D_r(r_i, r_j) \cdot \phi_j^{(u)} \tag{9}$$

$$w_{ij}^{(u)} = \frac{1}{Z_i^{(u)}} \cdot \exp[-\sigma_s^2 \cdot D_s(r_i, r_j)]. \tag{10}$$

Here, $\phi_j^{(u)} = n_j$ is used to emphasize the contribution of larger region. This is a common setting for the global contrast-based saliency detection [12], [13], which meets the global-first topological perception rule [36], [37]. $w_{ij}^{(u)}$ is the spatial weighting term which can increase the contributions of regions closer to r_i. $D_s(r_i, r_j)$ is the spatial distance between the centroids of r_i and r_j. σ_s is employed to control the strength of $w_{ij}^{(u)}$ and is set to $\sqrt{0.4}$ in all our experiments. $Z_i^{(u)}$ is the normalization factor ensuing $\sum_{j \neq i} w_{ij}^{(u)} = 1$.

D. Preactivation

The most essential difference between NATAS and the existing works is the treatment for preactivation. The proposed model takes into account the unevenly distributed preactivations in the target scene, while the existing works have not yet considered this factor. Obviously, in the case that all the images in a sequence are presented only for a very short duration, only the salient regions in the previous images are illuminated by the attentional spotlight, and will effect the observers' attention in the subsequent images. Besides, the biochemical traces of neural activities will naturally subside over time. In this paper, all these characteristics are well integrated in the definition of preactivations

$$V(r_{i,k}) = \sum_{q=\max(0,k-L)}^{k-1} \sum_{j} w_{ij}^{(v)} \cdot D_r(r_{i,k}, r_{j,q})^{-1} \cdot \phi_{ij}^{(v)} \tag{11}$$

$$w_{ij}^{(v)} = \frac{1}{Z_i^{(v)}} \cdot \exp[-\sigma_v^2 \cdot (k-q)^2 \cdot D_s(r_{i,k}, r_{j,q})] \tag{12}$$

$$\phi_{ij}^{(v)} = S(r_{j,q}) \cdot \frac{\min(n_i, n_j)}{\max(n_i, n_j)} \tag{13}$$

where $r_{i,k}$ denotes the ith region in the target image I_k, and $r_{j,q}$ denotes the jth region in the previous image I_q. L is the maximum number of previous images taken into account. The appreciative value of L is empirically determined as two through a quantitatively comparative analysis, which is described in detail in Section V-B. $w_{ij}^{(v)}$ is the time-spatial weighting term, which is introduced to incorporate the time cue $(k-q)$ and the spatial information $D_s(r_{i,k}, r_{j,q})$ into the definition of $V(r_{i,k})$, and is normalized by $Z_i^{(v)}$. $\phi_{ij}^{(v)}$ is used to emphasize the contributions of the regions that are with the areas more similar to $r_{i,k}$, and are with the larger saliency values detected before.

IV. Results

This section firstly specifies the quantitative indicators popular in saliency detection literatures, and then evaluates the performance of the proposed NATAS model on the constructed data set.

A. Evaluation Measure

In saliency detection works, precision and recall are two major quality indicators for the detected results. Precision measures the proportion of positive detected salient regions in the entire detected regions, while recall measures the percentage of correctly assigned salient regions in relation to the truly salient regions according to the ground truth.

Precision and recall are often mutually antagonistic, i.e., an algorithm emphasizing on high recall often tends to select larger salient regions with the sacrificing of precision rate, and vice versa. Therefore, these two measures are often considered together. In practice, the precision–recall curve is popular in capturing the tradeoff between precision and recall. Besides, a single indicator, F-measure [43], which is a harmonic mean for precision and recall, is also important in saliency detection. The definitions of precision, recall, and F-measure are already shown in Section II-B. In order to calculate the statistical curve and F-measure, we binarize the produced saliency maps under different thresholds. The thresholds used in all our experiments are 21 fixed value, i.e., $[0 : 0.05 : 1] \times 255$.

B. Performance

In order to verify the effectiveness of NATAS, 14 state-of-the-art saliency detection methods are selected as the baselines according to four principles: recency, prevalence, variety, and relevance. These competitive methods include: contrast determination filter-based saliency (AC [24]), local information maximization-based saliency (AIM [31]), context-aware saliency (CA [30]), frequency-tuned saliency (FT [29]), graph-based visual saliency (GB [23]), histogram-based saliency (HC [12]), extended contrast sensitivity function-based saliency (IM [49]), center-surround difference-based saliency (IT [11]), spatiotemporal saliency (LC [17]), region-level saliency (RC [12]), Bayesian surprise-based saliency (SEG [18]), local self-resemblance-based saliency (SeR [28]), spectral residual-based saliency (SR [26]), and natural statistics-based saliency (SUN [27]).

The code for NATAS is implemented in MATLAB. The implementation of IT used here is a SaliencyToolbox.[3] This implementation is more compact while contains the core functionality for the original code in [11], and is approved in Itti's project webpage.[4] As for LC, we have not found the authors' implementation, and the code used here is implemented by Cheng *et al.* [12].[5] For the other 11 methods,

[3] http://www.saliencytoolbox.net/

[4] http://ilab.usc.edu/toolkit/downloads.shtml

[5] http://cg.cs.tsinghua.edu.cn/people/~cmm/

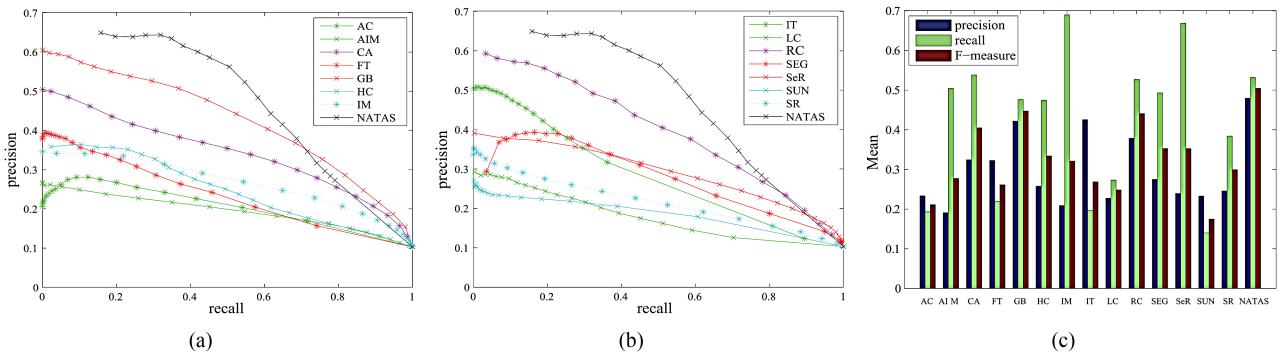


Fig. 8. Quantitative comparison between NATAS and the competitive methods representing the state-of-the-art. (a) and (b) Precision-recall curves. (c) Averaged precision, recall, and F-measure bars.

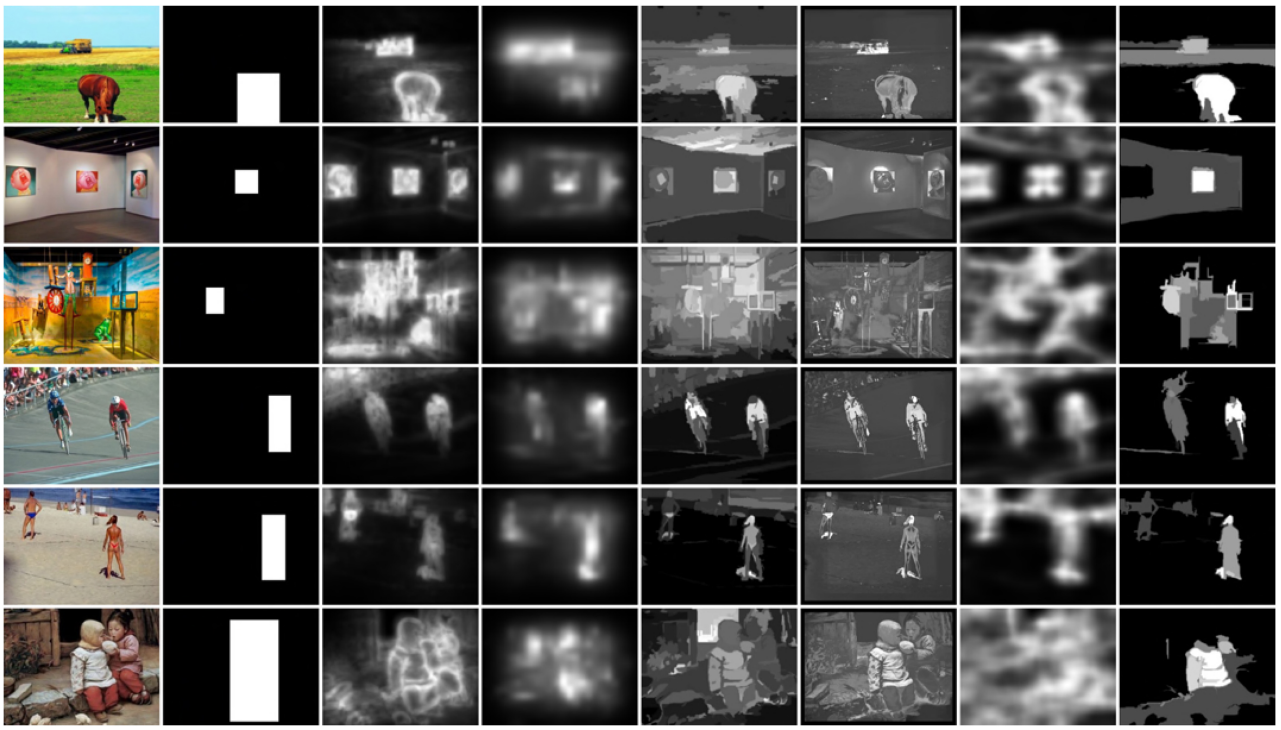


Fig. 9. Visual comparison of saliency maps. First column: original image sequence in a block. Second column: ground truth labels of the corresponding images. Remaining columns: saliency maps calculated by CA [30], GB [23], RC [12], SEG [18], SeR [28], and the saliency maps calculated by NATAS, respectively.

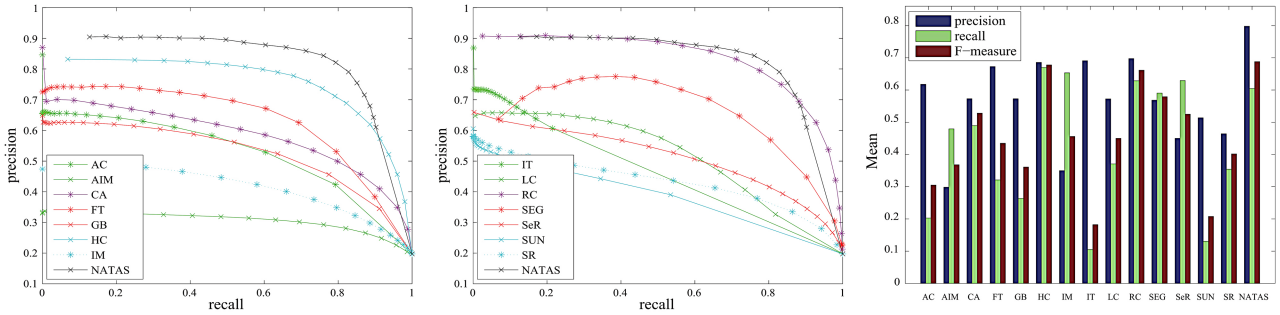
we used the authors' implementations downloaded from their homepages.

From Fig. 8(a) and (b), the precision-recall curves show that NATAS clearly outperforms AC, AIM, FT, HC, IM, IT, LC, RC, SEG, SeR, SR, and SUN. Compared with these 12 methods, the proposed model can yield saliency maps with higher accuracy at the same recall rate, or detect more truly salient regions at the same precision rate. Besides, we can observe that the proposed method dominates CA and GB most of the time until a high recall rate is reached. However, the unilateral emphasis on high recall rate is not very meaningful in practical applications. The more appropriate choice should emphasize on both precision and recall rate in compromise [43]. In this case, the averaged precision rate, recall rate, and F-measure value in Fig. 8(c) can provide more discriminative clues. This figure indicates that, NATAS has the obvious advantages compared with others.

Fig. 9 presents some visual comparison for the saliency detection results of the top five of the 14 aforementioned competitive methods and NATAS. As can be seen, in the saliency maps produced by the proposed method: 1) the salient area can be distinguished with the background much easier than others, and 2) the detected salient regions are much more consistent with the ground truths than the other competitors. All these comparison results are sufficient to demonstrate that, the proposed method is indeed a more suitable choice for the detection of visually saliency in image sequences.

V. Discussion

As presented in Section IV, there is a tremendous advancement when taking the global contrast and the preactivation into account. In this section, we will evaluate the two components independently.



(a) (b) (c)

Fig. 10. Quantitative comparison between the restricted NATAS and the state-of-the-art methods on the data set constructed by Achanta *et al.* [29]. (a) and (b) Precision–recall curves. (c) Averaged precision, recall, and F-measure bars.

A. Global Contrast Measurement

In order to further validate the effectiveness of the proposed global contrast, the NATAS model is evaluated on separate scenes without preactivations. A publicly available dataset containing 1000 images [24] is employed. This data set is reported to achieve great popularity in saliency detection [12], [13]. Each image in this data set contains one or several salient objects, and has a manually-labeled ground truth.

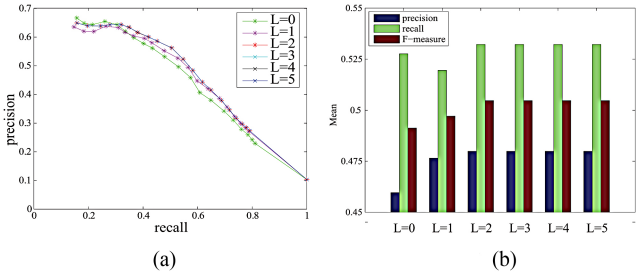
Fig. 10 demonstrates the results. As can be seen from the precision–recall curves [presented in Fig. 10(a) and (b)], it is manifest that the restricted NATAS, which is based only on the proposed measure of global contrast, can also outperform AC, AIM, CA, FT, GB, HC, IM, IT, LC, SEG, SeR, SR, and SUN in separate scenes. These curves in together indicate that, the proposed global contrast measurement can help NATAS to locate salient regions more accurate than these 13 exiting methods.

However, as can be seen in Fig. 10(b) that there are several crossovers between the curves of the restricted NATAS and RC. In such a case, the precision–recall curves cannot provide discriminative clues to support the comparative analysis for these two methods. However, as discussed in Section IV-B, the F-measure can help to provide more intuitive information. The corresponding results are demonstrated in Fig. 10(c). It is manifest that the restricted NATAS clearly dominates all the other methods on this indicator. As a summary, it is reasonable to believe that the proposed global contrast measure has made an important contribution in the advancement of NATAS.

B. Preactivation Consideration

In our experiment, the length of image sequences is fixed to six. However, does the proposed model have the ability to capture all the effects of the biochemical traces in an image sequence? Or, how long is the scope that NATAS can capture these effects caused by the previous images? This subsection will answer these questions by comparing the performances of NATAS under different settings of L.

As can be seen from Fig. 11, the worst performance is associated with the case of setting $L = 0$. In this case, no preactivation is considered. Then, with the increasing of L, which means that the effect of preactivation is gradually incorporated in the model, the performance will increasingly improve until L reaches two. After that, setting higher L cannot get a better performance. Instead, the computational cost is more expensive. These results indicate that the preactivation is indeed an important factor in the new saliency detection occasion. However, the proposed model is at most able to capture the preactivations aroused by the last two previous images.



(a) (b)

Fig. 11. Comparison of different settings of L. (a) Precision–recall curves. (b) Averaged precision, recall, and F-measure bars.

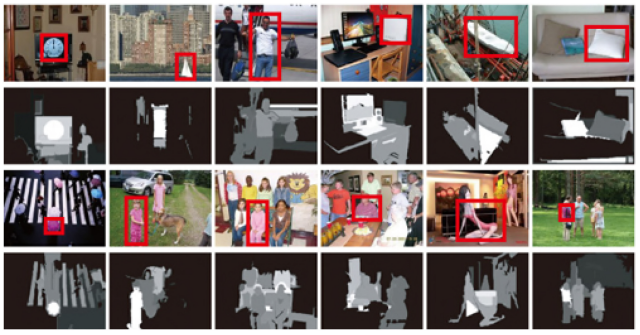


Fig. 12. Failure cases. The red rectangles illustrate the truly salient regions labeled by the users during image sequence displaying, and the gray scale images are the saliency maps detected by NATAS. For the first image sequence, the truly salient regions are mainly white, but the detected saliency maps fail in the last five images. Similarly, the truly salient regions in the second image sequence are mainly purple, but the detected saliency maps fail in the third–fifth images.

VI. Conclusion

There is a common sense that our attention will be influenced by what we observed shortly before. Some psychophysical researches [20], [41] can provide the theoretical basis for this intuition. However, detecting saliency in such an occasion

has not yet been addressed in computer vision. Directly employing traditional methods on successively displayed image sequences is not appropriate, because they do not consider the neural activity trace that actually exists in human vision system.

In this paper:

1) A data set for the new saliency detection occasion is constructed through a rigorously designed behavioral experiment. The images in the dataset are organized as blocks of image sequence. This makes the research toward neural activity trace possible.
2) A saliency model NATAS is designed specifically for the new occasion. This model can utilize the global contrast-based information, as well as the preactivation caused by previous images.

Experimental results on the constructed dataset show that, in the new saliency detection occasion, the proposed NATAS model can predict the salient regions with greater accuracy than the other 14 mainstream methods. Furthermore, a quantitatively analysis indicates that both the two components of the proposed model have made significant contributions in the advancement of NATAS.

As the performance on the constructed data set is still far from satisfying, there are many remaining questions to be investigated. For example, are there other cues in images that could be utilized to capture the effect of the neural activity traces? How can these cues be modeled? Besides, the failure cases of the proposed method shown in Fig. 12 illustrate that there is also a challenge associated with the tradeoff between the single-scene clues and the neural activity traces. These issues might be the key points in future work.

REFERENCES

[1] W. James, *The Principles of Psychology*. New York, NY, USA: Henry Holt, 1890, vol. 1.

[2] Y. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," *Vis. Res.*, vol. 46, no. 26, pp. 4333–4345, 2006.

[3] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.

[4] G. Zhu, Q. Wang, Y. Yuan, and P. Yan, "Learning saliency by MRF and differential threshold," *IEEE Trans. Cybern.*, to be published.

[5] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.

[6] C. Jung and C. Kim, "A unified spectral-domain approach for saliency detection and its application to automatic object segmentation," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1272–1283, Mar. 2012.

[7] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun.–Jul. 2004, pp. II-37–II-44.

[8] L. Yang, N. Zheng, J. Yang, M. Chen, and H. Chen, "A biased sampling strategy for object categorization," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1141–1148.

[9] M. Wang, R. Hong, X.-T. Yuan, S. Yan, and T.-S. Chua, "Movie2comics: Towards a lively video content presentation," *IEEE Trans. Multimedia*, vol. 14, nos. 3–2, pp. 858–870, Jun. 2012.

[10] R. Hong, M. Wang, X.-T. Yuan, M. Xu, J. Jiang, S. Yan, and T.-S. Chua, "Video accessibility enhancement for hearing-impaired users," *TOMCCAP*, vol. 7, no. Supplement, p. 24, 2011.

[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 11, pp. 1254–1259, Nov. 1998.

[12] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 409–416.

[13] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

[14] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 478–485.

[15] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, "Salient object detection for searched web images via global saliency," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3194–3201.

[16] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[17] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[18] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.

[19] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3374, Dec. 2011.

[20] J. Pearson and J. Brascamp, "Sensory memory for ambiguous vision," *Trends Cogn. Sci.*, vol. 12, no. 9, pp. 334–341, 2008.

[21] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 97–136, 1985.

[22] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition: A gentle way," in *Proc. Biol. Motiv. Comput. Vis.*, 2002, pp. 472–479.

[23] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[24] R. Achanta, F. J. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Proc. Comput. Vis. Syst.*, 2008, pp. 66–75.

[25] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 481–488.

[26] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[27] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 1–20, 2008.

[28] H. J. Seo and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 45–52.

[29] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[30] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2376–2383.

[31] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 155–162.

[32] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley, "Image saliency: From intrinsic to extrinsic context," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 417–424.

[33] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[34] C. Liu, J. Yuen, A. B. Torralba, J. Sivic, and W. T. Freeman, "Sift flow: Dense correspondence across different scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 28–42.

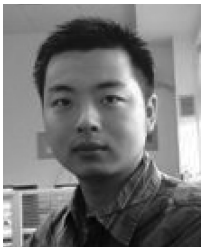
[35] G. Jeh and J. Widom, "Simrank: A measure of structural-context similarity," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 538–543.

[36] L. Chen, S. Zhang, and M. V. Srinivasan, "Global perception in small brains: Topological pattern recognition in honeybees," in *Proc. Nat. Acad. Sci.*, vol. 100. 2003, pp. 6884–6889.

[37] T. Zhou, J. Zhang, and L. Chen, "Neural correlation of 'global-first' topological perception: Anterior temporal lobe," *Brain Imag. Behav.*, vol. 2, no. 4, pp. 309–317, 2008.

[38] Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-spectral saliency detection," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 34–41, 2013.

[39] S. A. Hillyard and L. Anllo-Vento, "Event-related brain potentials in the study of visual selective attention," in *Proc. Nat. Acad. Sci.*, vol. 95. 1998, pp. 781–787.
[40] T. Pasternak and M. W. Greenlee, "Working memory in primate sensory systems," *Nature Rev. Neurosci.*, vol. 6, no. 2, pp. 97–107, Feb. 2005.
[41] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behav. Brain Sci.*, vol. 24, no. 1, pp. 87–185, Feb. 2001.
[42] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.
[43] D. R. Martin, C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.
[44] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 5, pp. 167–181, 2004.
[45] K. Koffka, *Principles of Gestalt Psychology*. London, U.K.: Routledge, 1955.
[46] J. Reynolds and R. Desimone, "Interacting roles of attention and visual salience in v4," *Neuron*, vol. 37, no. 5, pp. 853–863, 2003.
[47] P. Li, "An adaptive binning color model for mean shift tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 9, pp. 1293–1299, Sep. 2008.
[48] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 892–905, Aug. 2009.
[49] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 433–440.



Guokang Zhu is currently pursuing the Ph.D. degree at 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 current research interests include computer vision and machine learning.

Qi Wang received the B.E. degree in automation and the 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 current research interests include computer vision and pattern recognition.

Yuan Yuan (M'05–SM'09) is a full professor with the Chinese Academy of Sciences (CAS), Xi'an, 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.