Non-Local Spatial Redundancy Reduction for Bottom-Up Saliency Estimation☆

Jinjian Wu[a], Fei Qi[a], Guangming Shi[a,1,*], Yongheng Lu[a]

[a]*School of Electronic Engineering, Xidian University, Xi'an, Shaanxi, 710071, P. R. China*

Abstract

In this paper we present a redundancy reduction based approach for computational bottom-up visual saliency estimation. In contrast to conventional methods, our approach determines the saliency by filtering out redundant contents instead of measuring their significance. To analyze the redundancy of self-repeating spatial structures, we propose a non-local self-similarity based procedure. The result redundancy coefficient is used to compensate the Shannon entropy, which is based on statistics of pixel intensities, to generate the bottom-up saliency map of the visual input. Experimental results on three publicly available databases demonstrate that the proposed model is highly consistent with the subjective visual attention.

Keywords: Redundancy Reduction, Image Structure, Self-Similarity, Bottom-Up Visual Saliency, Visual Attention, Non-Local, Entropy, Human Visual System

1. Introduction

The human visual system (HVS) has a remarkable ability to analyze complex visual inputs in real-time, which locates regions of interest very quickly [1]. Finding interesting objects is a critical task in many image and video applications, such as region-of-interest based image compression [2], object recognition [3], image retrieval [4], image composition from sketch [5], advertisement design [6], image and video content adaptation [7], and quality evaluation [8, 9].

☆This work is supported in part by National Natural Science Foundation of China under grant NO. 60805012, 61033004, 61070138, and 61100155.

*Corresponding author.

Email address: gmshi@xidian.edu.cn (Guangming Shi)

[1]Tel.: +86-29-88201402.

Researchers attempt to build computational models imitating the ability of the HVS to improve vision related intelligent systems.

The rapid process during which the HVS scans the whole scene and guides eyes to focus on the most informative areas is called visual attention [1]. There exist two distinct mechanisms governing this procedure [10, 11], which are the bottom-up stimulus driven and the top-down goal driven mechanisms, respectively. The two mechanisms jointly determine the distribution of attention [12]. Bottom-up saliency estimation is the first step for image understanding and analysis, which is involuntarily response of environmental stimulus [1, 13]. In this paper, rather than building a saliency model including both the bottom-up and top-down mechanisms, we provide a model for pure bottom-up visual saliency estimation from the perspective of redundancy reduction.

1.1. Related works

Current research on the computational visual attention tries to model bottom-up and top-down mechanisms. The bottom-up based computational model imitates the function of the preattention, which is involuntary and pure data-driven, to generate a *saliency map* showing the conspicuousness of each position. The top-down mechanism determines the final response of the HVS [14] and directs eye fixation [15] according to voluntary affections. The existing top-down based computational models focus mainly on assessing contributions of each feature to fuse outputs of bottom-up based computational models [16–19].

Researches in neuropsychology show that the bottom-up saliency of a given location is determined by how distinct it is from its surroundings [10, 20, 21]. Furthermore, the bottom-up attention is driven by visual features of images, such as color [22], contrast in luminance [23, 24], sizes of objects [25], distributional statistics [26], contrast in histogram [27], and discriminant analysis [28, 29]. Based on these results, many computational models have been proposed to estimate the bottom-up visual saliency [1, 10, 13, 30–33]. In summary, bottom-up saliency is estimated with following steps: a) select a set of adequate features, b) evaluate the distinction over each feature, and c) fuse all channels of distinctions into the final saliency map.

Most existing models try to select some "good" features, on which objects are the most distinct against surroundings, for saliency estimation. In [34], Privitera and Stark evaluated the per-

formances of ten contrast based models on a single feature by comparing the generated saliency map with the eye fixation density map [34]. Koch and Ullman suggested that a set of elemental features affect the visual attention jointly rather than only one feature [20]. Taking a set of features into account, Itti et al. proposed a famous model for bottom-up visual saliency estimation [13]. In this work, Gaussian pyramids are created over color, orientation, and intensity, respectively. The "center-surround" differences are computed between levels in these pyramids and then combined as a uniform saliency map. Following this architecture, Le Meur et al. improved the normalization procedure by providing a coherent one [1]. Furthermore, Gao et al. compute the center-surround differences on these features employing a classification procedure [35]. In [36], Achanta et al. adopted the difference of Gaussian approach to compute the center-surround differences for saliency estimation. In the recent, according to the global contrast on color histogram, Cheng et al. [27] introduced an effective and efficient saliency detection model.

However, studies on primary visual cortex suggest that saliency map might be irrelevant to certain visual features [37] which means that saliency is untuned to specific features. Intuitively, any place distinct from its surrounding, with respect to any feature, is salient. In [18], Judd et al. collected a large set of low, mid and high-level features, such as intensity, orientation, color, horizon, people, face and car. Then a linear support vector machine is employed to train a saliency model, within which these distinct features are highlighted with large weights. However, it is imposible to exhaust all of the potential features for saliency detection. So a measurement of distinction is required to estimate the saliency under any potential features. In [38], a spectral residual procedure in spatial domain is introduced to estimate the distinction. According to the mechanism that visual attention focuses on informative places, we suggest that the quantity of information is an appropriate measurement for bottom-up saliency estimation.

Estimating bottom-up visual saliency based on self-information emerges in recent years [39, 40]. In [39], independent component analysis is firstly applied on image blocks to reduce the correlation among pixels. Then the likelihood of each pixel is estimated globally for self-information computation. In [40], with principle component analysis based dimensionality reduction, the likelihood of the local region for each pixel is computed via kernel density estimation. Then, the visual saliency map is generated according to Shannon self-information.

Using self-information is less robust than entropy because the latter is an expectation of the former. In the meanwhile, component analysis breaks down, at least in some degree, the spatial correlations among pixels, which is one of the dominant factors affecting visual saliency as we will show in this paper.

1.2. Our approach

In this paper, we focus on estimating bottom-up visual saliency based on redundancy reduction. In natural images, similar regions, which jointly represent some self-repeating structures, contain abundant redundancy. The redundant place carries little information and is simple to be understood. So, these redundant regions can be processed instantly (by brain) and attract little attention. While the informative regions are much complex and will attract most of attention for further processing. In summary, when scanning an input scene, the HVS can quickly process redundant places, and focus attention on the rest regions which contain abundant information. Therefore, we attempt to compute the redundancy of each input part, and filter out the redundancy to pop-out the saliency regions.

Traditionally, the saliency is computed based on information measurement. Shannon entropy is an effective metric for information measurement, which is based on the statistic probabilities of events. For simplicity, the intensity distribution of pixels are adopted for entropy computation [26, 34]. This procedure only considers the probabilities [41] and ignores the correlations among pixels, so the result is redundant. Places with identical statistic probabilities may have different correlated pixels and present different structures. For example, places with regular arrangements are highly correlated and present self-similar structures, which are very redundant to the HVS. So we need to consider the dependence among pixels, and remove the redundant information for saliency estimation. To this end, PCA/ICA procedures are adopted to eliminate the correlations among pixels for information computation [39, 40]. These methods isolate the central blocks from their surroundings and only consider the correlations of pixels in a small local patch. Therefore, how to effectively remove the correlations and discard the redundancy is the key issue for image saliency estimation.

Non-local means algorithm is an effective way to deal with the spatial redundancy of images.

In [42], with the non-local procedure, the spatial redundancy is fully used and the central pixel is restored according to the correlations with its surrounding pixels.

From the perspective of non-local means, redundant regions are highly similar with some other regions in their non-local neighborhoods. So we can directly measure the redundancy of each region by computing the structural similarity with its surrounding. In this paper, a non-local self-similarity procedure is introduced to thoroughly analyze the structural redundancy in an image. With the computation of the structural similarity between the central pixel and its neighboring pixels in a non-local region, the self-similarity coefficient of a pixel is acquired, which is called the redundancy coefficient. Then, we amend the normal entropy with the redundancy coefficient to discard the redundant information. Finally, integrating with color and scale spaces, we extend our procedure from scalar images to color images, and create a novel redundancy reduction based saliency estimation model for natural images. We test the proposed model on three publicly available databases and achieve results which are highly consistent with the subjective visual attention.

The rest part of this paper is organized as follows. In Section 2, we analyze the spatial redundancy in images, and create a novel redundancy reduction based bottom-up saliency estimation model. Then experimental results on three public databases and conceptual test images are illustrated in Section 3. Finally, conclusions are drawn in Section 4.

2. Redundancy Reduction Based Saliency Estimation

In this section, we firstly analyze the structural redundancy based on self-similarity. Then, we construct a computational procedure to measure the saliency of each pixel in scalar images with redundancy reduction. Finally, by taking color and scale information into account, a novel redundancy reduction based saliency estimation model is created for natural images.

2.1. Structural Redundancy and Self-Similarity

As pixels provide information jointly, a critical issue for computing the information that each pixel possesses is to quantitatively evaluate the redundancy among pixels. In this subsection, we discuss the computation of structural redundancy based on self-similarity within a non-local region.

(a) (b)

Figure 1: Example of structural redundancy and self-similarity. (a) The original image. (b) The redundancy and self-similarity in different parts of the image.

The concept of self-similarity originates from research on fractals [43]. Several methods have been proposed for image analysis [44, 45] and points of interests detection [46]. Here, for the purpose of redundancy evaluation, *image self-similarity* is considered as how well a region in an image can be approximated by other regions [47].

Intuitively, a region with regular or similar structure is more redundant than that with irregular or varying contents because one part can be inferred easily from the other parts in the former case. For example, as shown in Fig. 1, the three patches A, B, and C are with three representative structures. Patch B locates in a lawn, where the structure is highly self-similar. The information in patch B is so redundant that even though this patch is covered, it is quite easy for the HVS to reconstruct the content with the help of its surrounding contents. The structure of patch A is less self-similar than that of patch B, but it is also highly correlated with its surrounding and represents some redundant information, such as trees, mountains, and the sky. Therefore, the HVS can restore the rough content of patch B according to the structures of its surrounding. Furthermore, patch C possesses a unique horse object in the image. This patch is informative and presents quite different structure compared with its surrounding. We can hardly recover this patch since correct logical deduction is very hard to make according to its surrounding. Therefore, structural self-similarity is an effective measurement on redundancy.

According to discussions above, we introduce a quantitative self-similarity measurement employing the non-local means filtering kernel [42]. Suppose $f(x)$ be a scalar image, and symbols $F(x)$ and $F(y)$ denote the vectors formed by concatenating all columns in *local regions* $\Omega(x)$ and $\Omega(y)$, respectively. The *similarity* between the two regions is measured by the following kernel

$$S(x,y) = \tau \exp\left(-\frac{\left\|F(x) - F(y)\right\|_2^2}{2\sigma_x^2}\right), \tag{1}$$

where σ_x denotes a parameter related to region $\Omega(x)$, and τ the normalizing coefficient which is used to normalize the summation of $S(x,y)$ in the non-local region to be 1 (namely, $\sum_{y\in\Omega(x)} S(x,y) = 1$). The similarity kernel $S(x,y)$ measures the proportion of information when representing the region $\Omega(x)$ by the region $\Omega(y)$. According to Eq. (1), we can get

1. when the two regions are same, the distance between them is $\left\|F(x) - F(y)\right\| = 0$, and the similarity is $S(x,y) = \tau$,
2. when the two regions are completely different, the distance between them is $\left\|F(x) - F(y)\right\| > 3\sigma_x$, and the similarity is $S(x,y) \approx 0$,
3. in other cases, $3\sigma_x \geq \left\|F(x) - F(y)\right\| \geq 0$, the similarity is $0 < S(x,y) \leq \tau$.

According to the research on the HVS, the self-similarity of a pixel depends upon pixels near it more than farther ones. As shown in Fig. 2, to evaluate the redundancy of pixel x, we consider the similarity between its local region $\Omega(x)$ and other local regions with reference pixel in its *surrounding region* $\mathcal{R}$. Local regions outside the surrounding region $\mathcal{R}$ are omitted for the convenience of computation. Given the surrounding region $\mathcal{R}$, the *self-similarity* for location x is given by

$$\varrho(x) = \sum_{\substack{y\in\mathcal{R} \\ y\neq x}} \phi(\|y - x\|) S(x,y), \tag{2}$$

where $\phi(\cdot)$ is a radial basis function which weights the contribution of position y according to its distance to the reference position x. We take $\varrho(x)$ as the redundancy coefficient of pixel x in this paper.

2.2. Information Measurement for Saliency Estimation

According to Shannon information theory, image information is always measured based on the intensity distribution of pixels [26, 34]. For a pixel x, the information is measured based on

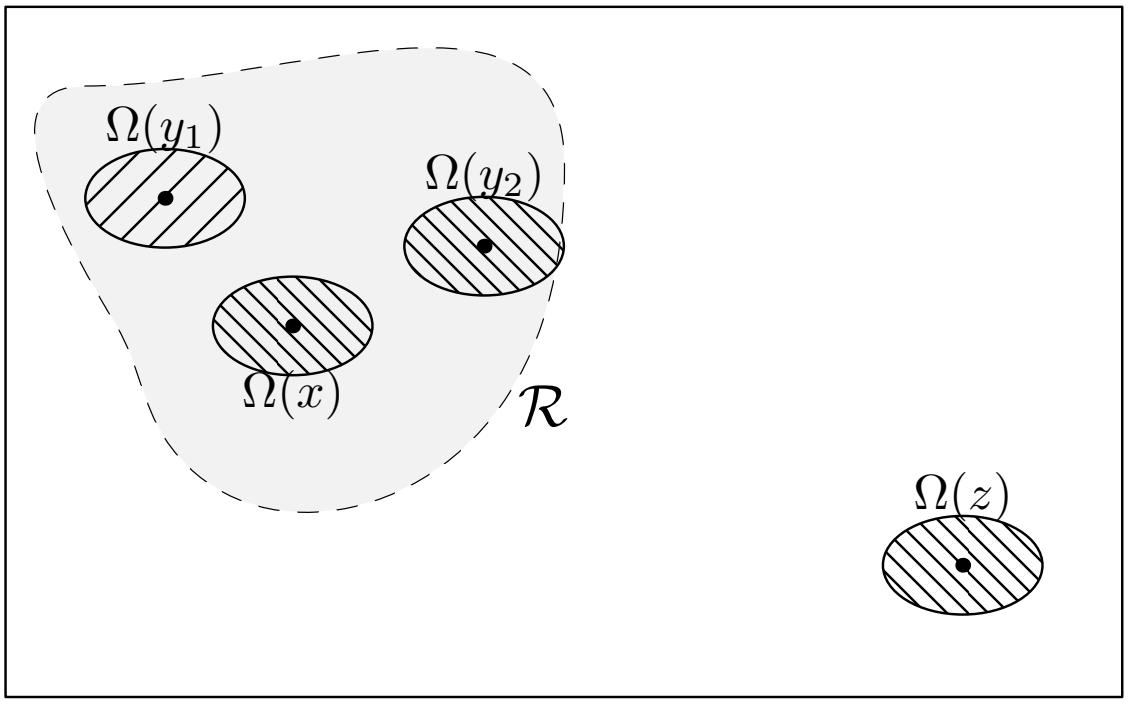


Figure 2: Illustration of image self-similarity computation. The shaded region $\mathcal{R}$ represents the surrounding region, and the cross-hatched regions $\Omega(\cdot)$ are local regions.

Shannon entropy of its local region $\Omega(x)$, i.e., $H(x) = H(\Omega(x))$. To compute this entropy, we map the region $\Omega(x)$ to a histogram with K bins where $p_b(x)$ denotes the probability of pixels taking an intensity within the bth bin. The entropy of the pixel x is given by

$$H(x) = \sum_{b=1}^{K} -p_b(x) \log p_b(x). \tag{3}$$

In addition, we need to consider the structural redundancy for image information measurement. Since Eq. (3) is based on the histogram of intensity, it is sensitive to luminance change. As Fig. 1 shows, all of the three patches A, B, and C possess luminance changes and they will acquire large information values according to Eq. (3). However, based on the analysis in the previous subsection, patch B is very redundant to the HVS. This patch is located at a region with self-repeating glassing texture and shares a tiny part of glass information. Patch C contains a unique object and it represents a very different kind of structure from its surrounding. Therefore, patch C shares little information with its surrounding and possesses a large quantity of information. In summary, we need to remove the structural redundancy from normal entropy Eq. (3) to measure the informativeness or saliency of each pixel in an image. With the redundancy coefficient Eq. (2) and the normal entropy Eq. (3), the redundancy reduction based saliency is estimated as

$$\hat{H}(x) = (1 - \varrho(x))H(x). \tag{4}$$

A rough saliency map for a scalar image can be obtained directly with Eq. (4). As $I(x)$ in Eq. (4) takes only one color channel and one scale into account, it is not applicable for general color images which are multichannel and contain objects with different sizes. In the subsection, we extend above formulation to general color images for estimating the bottom-up visual saliency map.

2.3. Computational Saliency Model

In respect that the quantity of information of a pixel is an appropriate measurement of how distinct the pixel is, we propose a novel bottom-up saliency estimation model based on redundancy reduction.

In order to make the saliency estimation procedure applicable for color images, the information of all color channels must be considered. One of the tricks is converting color images to a grayscale one, but it loses color information which has strong affection on visual saliency [48].

If the channels are independent, we can process all channels separately and sum them up to produce the final saliency estimation. Here we choose the opponent color space whose channels have been proved to be independent [49]. Given an image in the RGB space, it can be transformed to Weijer's [49] opponent color space by

$$\begin{aligned} o_1 &= \frac{\beta R - \alpha G}{\sqrt{\alpha^2 + \beta^2}} \\ o_2 &= \frac{\alpha\gamma R + \beta\gamma G - (\alpha^2 + \beta^2)B}{\sqrt{(\alpha^2 + \beta^2 + \gamma^2)(\alpha^2 + \beta^2)}}, \\ o_3 &= \frac{\alpha R + \beta G + \gamma B}{\sqrt{\alpha^2 + \beta^2 + \gamma^2}} \end{aligned} \tag{5}$$

where o_1, o_2 and o_3 are the three independent channels, R, G, and B are the red, green, and blue components, respectively, and (α, β, γ) denotes the illuminant which generally takes a value $(1, 1, 1)$.

As the sizes of objects in an image are arbitrary, we employ the multiscale framework to handle this issue. For simplicity, the sizes of the local and surrounding regions are fixed while the image is resized to several levels of scales with the pyramid technique. According to the multiscale theory, a Gaussian pyramid with L scale levels is created. On each level, the saliency is computed according to the resized image. As shown in Fig. 3, images (c) to (g) are the saliency maps on five scales of image (a). With the reduction of size, the fire extinguisher is pop-out as a whole. The saliency map of each level is adjusted to the original size for fusion which is discussed later this subsection.

The final saliency map is constructed by combining the scalar saliency maps under all channels and scales. As we focus on bottom-up visual saliency estimation, no prior information is available. So, all channels and scales are treated equally in the fusion. Let $\hat{H}_{lc}(x)$ denotes the scalar saliency map for channel c on the lth scale level, the visual saliency map is constructed with

$$\mathcal{S}(x) = \sum_{l=1}^{L} \sum_{c=1}^{C} w_{lc} \hat{H}_{lc}(x), \tag{6}$$

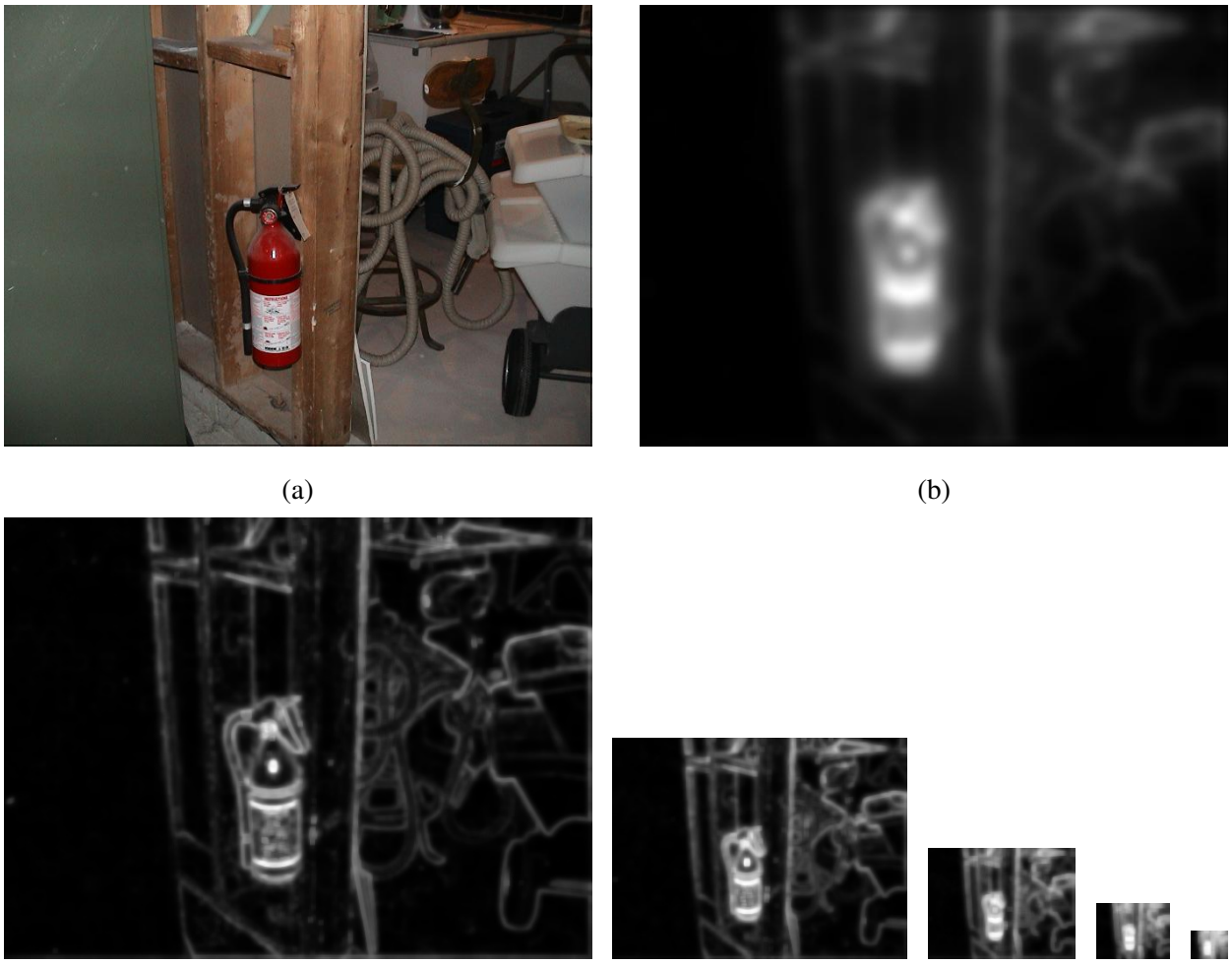


Figure 3: Multiscale processing for saliency estimation. (a) Original image. (b) The overall saliency map. (c)-(g) Intermediate saliency maps for each scale.

where L is the number of pyramid levels, C is the number of image channels, and w_{lc} is the normalizing coefficients for each channel and scale. In this paper, we set $w_{lc} = 1/\max_x \hat{H}_{lc}(x)$ so that the normalized scalar saliency maps $w_{lc}\hat{H}_{lc}(x)$ are with values in the range from 0 to 1.

3. Experiments

In this section, experiments are demonstrated to evaluate the performance of the proposed redundancy reduction based bottom-up visual saliency estimation model on general images. Firstly, we illustrate the surpass performance of the proposed model over the simple entropy model. Then, we compare the proposed model with the classical Itti et al.'s IT model [13] and three state-of-the-art models (i.e., the AIM model [39], Judd et al.'s LP model [18], and Cheng et al.'s HC model [27]) on three public databases (i.e., DB1 provided by Bruce and Tsotsos [50], DB2 by Achanta et al. [36], and DB3 by Judd et al. [18]). The implementation codes of the four comparisonal models are available on the authors' homepages (IT [1], AIM [2], LP [3], and HC [4]). Finally, some concept images are chosen to illustrate the effectiveness of the proposed bottom-up model with the pure effect of low level factors.

In the proposed model, the performance varies according to several parameters. In all experiments given in this section, we use a same set of parameters. The local regions Ω are 7×7 blocks, the surrounding regions $\mathcal{R}$ are 21×21 rectangles, and the decay factor σ_x is set to 20. The weighting function $\phi(\cdot)$ uses a Gaussian kernel with a fixed variance 7. The number of pyramid levels is given by $L = 1 + \lfloor \log_2 \frac{\min\{H,W\}}{21} \rfloor$ where H and W are the height and width of the input image, respectively.

3.1. Redundancy Reduction

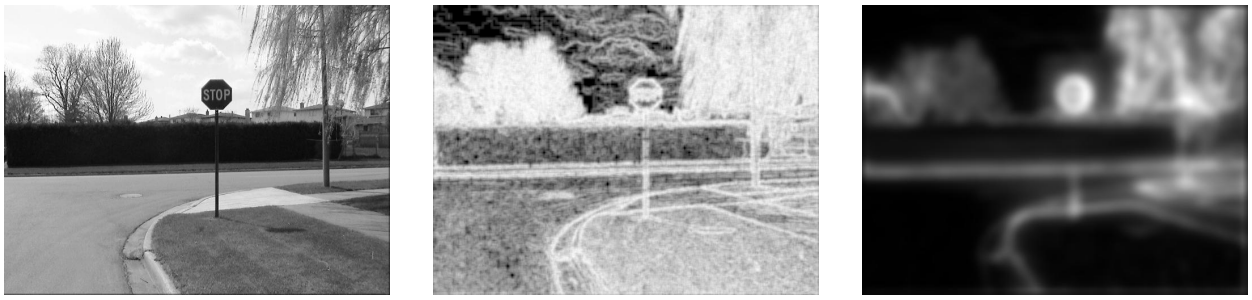
As we stated, pixels provide information jointly, and redundancy reduction is the most important issue in bottom-up saliency estimation. Here we show the effectiveness of the proposed

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

[2] http://www-sop.inria.fr/members/Neil.Bruce/#SOURCECODE

[3] http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html

[4] http://cg.cs.tsinghua.edu.cn/people/ cmm/Saliency/Index.htm

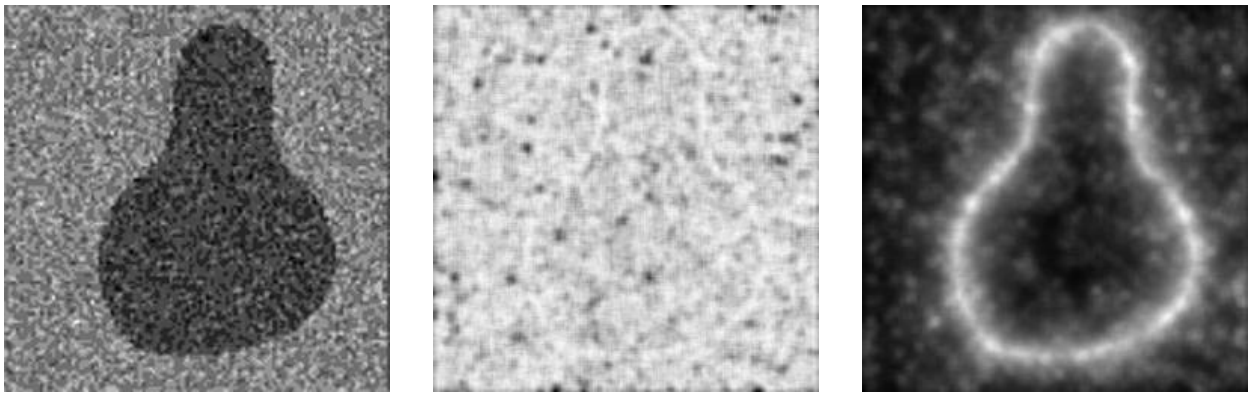


(a) (b) (c)

Figure 4: The effectiveness of the proposed model with spatial redundancy reduction. (a) The original image. (b) The entropy map without redundancy reduction. (c) The proposed saliency map with redundancy reduction.

saliency estimation approach on spatial redundancy reduction for images. We build saliency maps for some simple grayscale images with conceptually salient objects based on the general entropy with redundancy and the proposed procedure, respectively. As depicted by Fig. 4, (a) is the original grayscale image, (b) is the entropy map with redundancy, and (c) is the proposed saliency map with redundancy reduction. According to the HVS, the background parts in Fig. 4 (a), such as the road, the grassland, and the bush, provide little information. In the entropy map Fig. 4 (b), the response is very sensitive to the small changes of intensity in these background which contradicts that of the HVS. With the proposed approach, the response is insensitive to such changes and the object regions are effectively highlighted, as shown by Fig. 4 (c).

In the meanwhile, the proposed approach is very robust to noise, as shown by Fig. 5. The original image (a) contains a cucurbit polluted by heavy noise, however the shape of the cucurbit can be deduced out by human perception easily while challenging for machines. As given in Fig. 5 (b), the entropy map without redundancy reduction is very similar to an image of random noise, which means this model completely fails to detect the salient cucurbit. With redundancy reduction, the proposed model successfully locates the salient parts of the image, as shown by Fig. 5 (c).



(a) (b) (c)

Figure 5: The robustness of proposed model with spatial redundancy reduction. (a) The original image with heavy noise. (b) The entropy map without redundancy reduction. (c) The proposed saliency map with redundancy reduction.

3.2. Saliency Estimation on General Images

We apply the proposed model on three public databases, DB1 [50], DB2 [36], and DB3 [18]. Then we compare the proposed model with the classical IT [13] and three state-of-the-art models (i.e., AIM [39], LP [18], and HC [27]). To make a fair comparison among the four bottom-up based computational models (except LP, since it is a bottom-up and top-down combining model, and all information channels are combined with learning weights), we equally add up all information channels and no special combination procedure is adopted.

3.2.1. Saliency Objects Detection

Some results of saliency estimation on DB1 [50] are illustrated in Fig. 6. In this figure, the rows are the original images, the ground truth images (eye fixation density maps), the saliency maps produced by IT [13], AIM [39], HC [27], and the proposed model, respectively, from top to bottom.

As shown in the first original image in Fig. 6, this scene is with one salient object and simple background. All these tested approaches produce very good saliency maps that accurately highlight the doorknob and are highly coincided with the corresponding eye fixation density map. The backgrounds of the second and third original images are much complex, as they are composed of

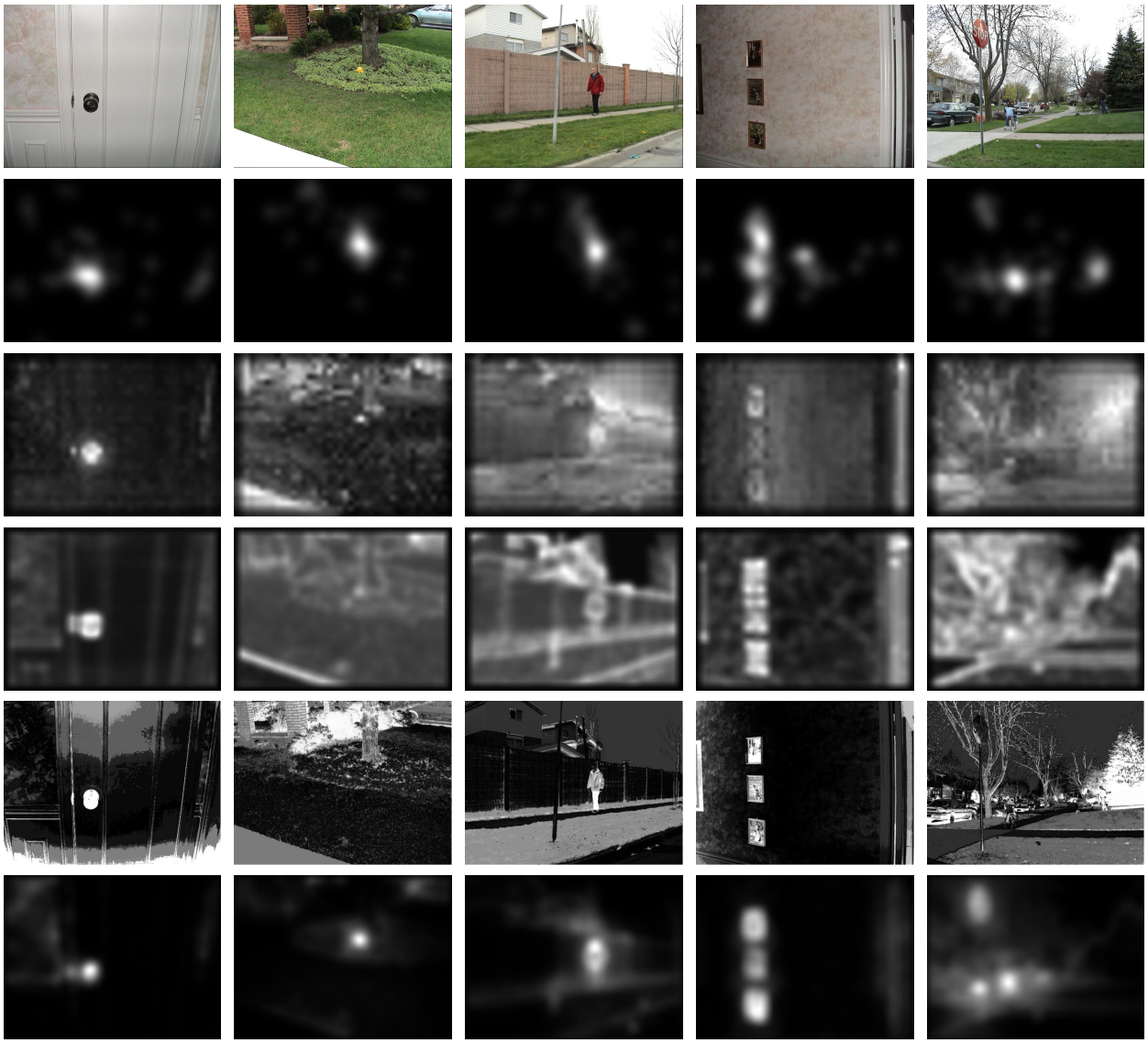


Figure 6: Results of saliency estimation on natural images. From the top to the bottom rows, they are the original images, the ground truth, the saliency maps produced by IT [13], AIM [39], HC [27], and the proposed model, respectively.

several regions, such as grass, bush, wall and building. It is easy for the HVS to find out the salient objects in the two images since all of these backgrounds are with self-similar textures and are less informative than the saliency objects. While it is challenging for the computational models. As can be seen, the IT, AIM and HC models are seriously disturbed by the complex backgrounds and highlight some background regions. Since the background regions of the two images are with self-similar textures, according to the analysis in Section 2, these regions are very redundant and each pixel in them provides very little information. With redundancy reduction, the proposed model effectively estimates the saliency of the two images and returns saliency maps which are highly identical to the eye fixation density maps.

Furthermore, the last two original images contain multiple saliency objects. Since the background of the fourth image is very simple, all of the models can accurately highlight the multiple saliency regions. However, the content of the last original image is very complex, which consists of people, road sign, cars, building, trees and grass. From the eye fixation density map we can see that the human attention is mainly focused on the people, the road sign and the cars. The IT model is totally failed in this image, which highlights most of the place especially the trees located at the right side. The AIM model plays no better than the IT model, which also highlights almost all of the image. The HC model mainly highlights the trees located at the right side which is not salient. Since the trees, the grass and the ground have highly self-similar structures, the proposed model can effectively filter out these backgrounds and accurately highlight the child, the car and the road sign. The computational result on the last original image from the proposed model is highly coincided with the eye fixation density map.

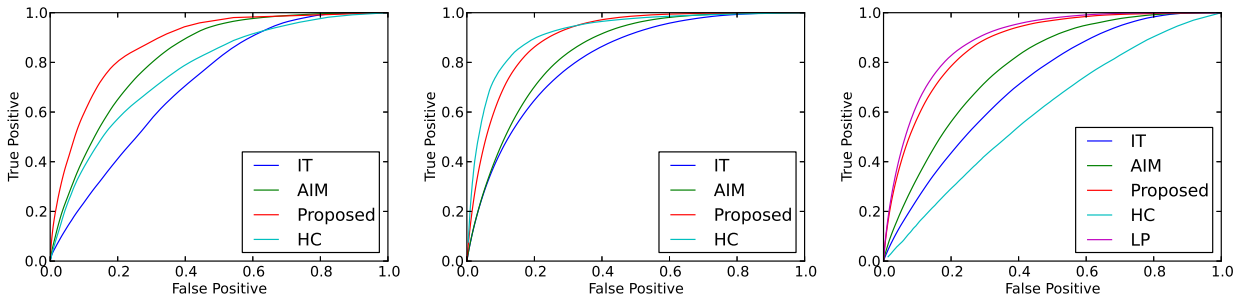
Therefore, with the non-local self-similar procedure, the redundancy from the image can be effectively removed. The proposed model can accurately find out saliency objects from both simple and complex backgrounds.

3.2.2. Overall Performance

In order to make a comprehensive analysis, we verify the proposed model on three publicly available databases. These databases consist of a variety of indoor and outdoor scenes. The characteristics of the three databases are summarized in Table 1.

Table 1: Three publicly available databases for saliency estimation

Character	DB1 [39]	DB2 [36]	DB3 [18]
Image Number	120	1000	1003
Data Achieve	Eye Track	Human Marked	Eye Track
Ground Truth	Gray Map	Binary Mask	Gray Map



(a) (b) (c)

Figure 7: The ROC curves of performance for these saliency models on the three public databases, (a) DB1 [39], (b) DB2 [36], and (c) DB3 [18].

We compare the proposed model with four saliency estimation models and adopt the receiver operating characteristic (ROC) metric to assess their performances. The ROC metric measures the area under the ROC curve [18]. To calculate this measurement, the saliency map is treated as a binary classifier, where a pixel with a greater saliency value than a threshold is classified as fixation and the rest of the pixels as nonfixated pixels. By varying the threshold, the ROC curve is acquired. The larger the area under the curve is, the better the saliency estimation method performs.

The ROC curves of these saliency models on the three public databases are shown in Fig. 7, and their corresponding ROC areas ($\mathcal{A}$) are listed in Table 2. As can be seen, the proposed model (with $\mathcal{A}$=0.876) performs better than the other three bottom-up based computational models (IT with $\mathcal{A}$=0.713, AIM with $\mathcal{A}$=0.823, and HC with $\mathcal{A}$=0.767) on DB1. On DB2, as shown in Fig. 7 (b), the proposed model (with $\mathcal{A}$=0.903) outperforms IT model (with $\mathcal{A}$=0.814) and AIM model (with $\mathcal{A}$=0.840), and approximates to HC (with $\mathcal{A}$=0.919, the state-of-the-art performance on this database). And on DB3, the proposed model (with $\mathcal{A}$=0.874) also outperforms

Table 2: The ROC areas of the saliency models on the three public databases

Algorithm	DB1 [39]	DB2 [36]	DB3 [18]
IT [13]	0.713	0.814	0.714
AIM [39]	0.823	0.840	0.780
HC [27]	0.767	**0.919**	0.600
LP [18]	–	–	**0.890**
Proposed	**0.876**	0.903	0.874

the three bottom-up based computational models (IT with $\mathcal{A}$=0.714, AIM with $\mathcal{A}$=0.780, and HC with $\mathcal{A}$=0.600). In LP, a large sets of low-level features (such as intensity, orientation and color), mid-level features (such as horizon) and high-level features (such as people, face and car) are extracted and combined with a top-down learning procedure [18]. As a result, LP returns highly consistent saliency maps with the ground truth (with $\mathcal{A}$=0.890). Though the proposed model is purely bottom-up based, its performance on DB3 is approximate to LP. In summary, the proposed model is comparable with the state-of-the-art models and is highly consistent with the subjective visual attention.

3.3. Visual Saliency of Concept Images

To further validate the effectiveness of the proposed model, we demonstrate our procedure on some concept images, whose saliency regions are merely determined by low level features.

Fig. 8 shows three concept images with salience objects due to single factor. The left column shows an image with some colored points, and they are differ in intensity. The light point is unique and informative. While the other points are homogeneous, they are much more redundant than the lighted one. Therefore, the lighted colored point is with the highest saliency on the estimated map. The middle column shows salience objects under different shapes. Since the sign "-" is distinct among signs "+", the sign "-" is less redundant than that "+" in this image, and this is also well located in its corresponding saliency map. The right column shows a salient case caused by the change of object orientation. There are three orientations of the objects, two orientations are with multiple objects, and the third orientation is with a single object. As the unique object contains

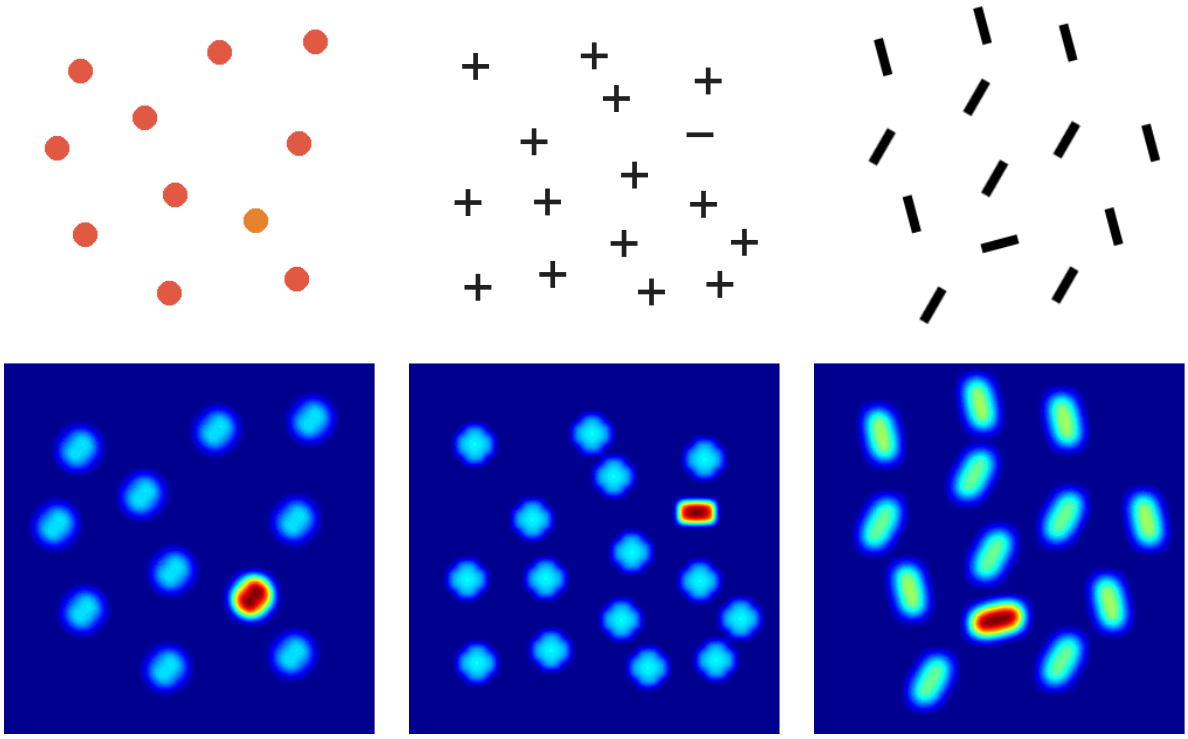


Figure 8: Visual saliency estimation on single factor. From left to right, the factors considered are color, shape, and orientation, respectively. The top row is the original images and the bottom row is their corresponding saliency maps.

much information, it is the most salient as shown by its saliency map. Other objects are with nearly same saliency, that is because the first two orientations are with 6 objects.

This experiment (Fig. 8) indicates that the proposed algorithm is adaptive to the saliency caused by different features. Since our approach does NOT have a special feature extraction step, the adaptive capability originates from our redundancy reduction based image information metric.

Besides these decoupled factor cases, we further test our approach with coupled factors. As Fig. 9 shows, the objects are with two colors and two orientations, and these coupled factors affect the saliency of the image. In addition, in the original image, ignore the color, the two orientations have the same number of objects. The saliency map performs well, in which the green object is with the highest saliency, following are the red objects with the orientation same to the green one, and the lowest ones are other red objects. The output is in accord with the logic that multiple

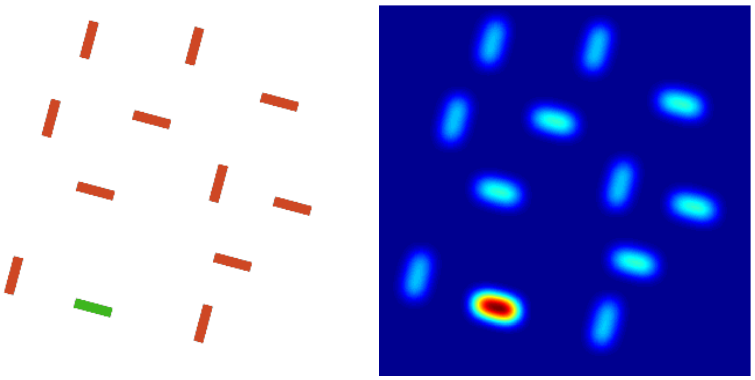


Figure 9: Visual saliency estimation with two independent factors, which are color and orientation. The left is the original image, and the right is saliency maps.

instances with same configuration are redundant, and they provide little additional information than only one instance.

Fig. 10 shows a complex case whose saliency is affected by some coupled factors. It is challenging for most existing bottom-up saliency models. Objects in the original images are with different color, size, orientation, and even an mirrored object. Applying the proposed approach, all these factors are successfully detected. As the saliency map shows, the most salient object is the one with different orientation, and the objects with saliency on size, color, and mirrored, also pop out. Though the orange object is distinct in color, this one has the same shape to most of the objects in the image. Since these objects provide shape information jointly, the redundant coefficient is large and each one share a very little part of shape information. Therefore, the orange one is not so salient. Meanwhile, the differences in orientation and size bring new information, and these objects gain high saliency values. As the mirrored object share the same color and three horizontal lines to most objects, it is hard to be detected at the early stage of the visual perception. Furthermore, most existing bottom-up saliency algorithms fail in this case.

4. Conclusions and Future Works

In this paper, we propose a redundancy reduction based visual saliency estimation model. The model focuses on the reduction of spatial structural redundancy in images. Modeling redundancy coefficient is the foundation of the proposed model. For images, we introduce the spatial structural

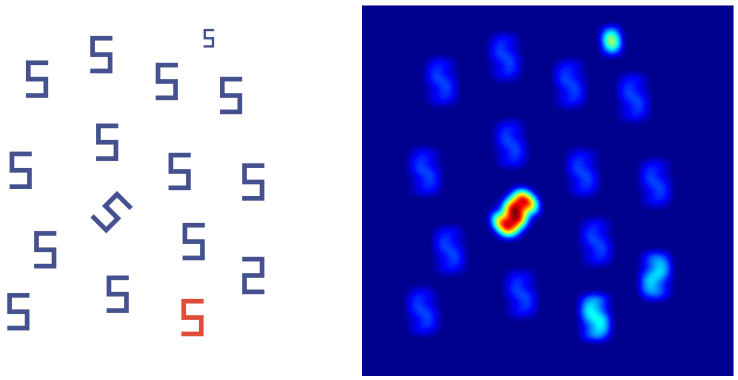


Figure 10: Visual saliency estimation on multiple factors, which are color, size, orientation, and even a mirrored object. (Left) The original image. (Right) The saliency map.

self-similarity as an approximation to the redundancy coefficient for pixels. Following that, we obtain the informativeness of pixels for scalar images at a particular scale. Taking color and scale spaces into account, we construct the novel model for visual saliency estimation. Experiments on the three publicly available databases show that the proposed model is comparable with the state-of-the-art saliency models.

Though the proposed model succeed on the three publicly available databases, it has limitations in some aspects and needs to be improved. As the current model applies to still images only, we expect to extend it to dynamic sequences which requires a temporal or joint spatial-temporal saliency model. The similarity measurement is based on pixel-by-pixel block comparison, we plan to find a more versatile function for general textures. As the proposed approach is fit for implementation on massively parallel architectures, we expect to build a neural network model to explain the topology and working mechanism of the visual cortex.

[1] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, 2006.

[2] J.-S. Lee, F. D. Simone, and T. Ebrahimi, "Efficient video coding based on audio-visual focus of attention," *J. Visual Communication and Image Representation*, vol. 22, no. 8, pp. 704–711, 2011.

[3] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 37–44.

[4] L. Shao and M. Brady, "Invariant salient regions based image retrieval under viewpoint and illumination variations," *J. Visual Communication and Image Representation*, vol. 17, no. 6, pp. 1256–1272, 2006.

[5] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM Transactions on Graphics*, vol. 28, no. 5, pp. 124:1–10, 2009.

[6] L. Itti, "Models of bottom-up and top-down visual attention," Ph.D. dissertation, California Institute of Technology, Pasadena, California, Jan 2000.

[7] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin, "A shape-preserving approach to image resizing," *Computer Graphics Forum*, vol. 28, no. 7, pp. 1897–1906, 2009.

[8] J. You, A. Perkis, M. M. Hannuksela, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proc. 17th ACM Int'l Conf. Multimedia*, ser. MM '09. New York, NY, USA: ACM, 2009, pp. 561–564.

[9] J. You, J. Korhonen, A. Perkis, and T. Ebrahimi, "Balancing attended and global stimuli in perceived video quality assessment," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1269–1285, Dec. 2011.

[10] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[11] R. J. Peters, A. Iyer, C. Koch, and L. Itti, "Components of bottom-up gaze allocation in natural scenes," *J. Vision*, vol. 5, no. 8, pp. 692–692, 2005.

[12] A. Borji, M. N. Ahmadabadi, B. N. Araabi, and M. Hamidi, "Online learning of task-driven object-based visual attention control," *Image and Vision Computing*, vol. 28, no. 7, pp. 1130–1145, 2010.

[13] 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. 20, no. 11, pp. 1254–1259, 1998.

[14] S. Kastner and L. G. Ungerleider, "Mechanisms of visual attention in the human cortex," *Annu Rev Neurosci*, vol. 23, pp. 315–41, 2000.

[15] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition*, vol. 12, no. 6, pp. 1093–1123, 2005.

[16] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[17] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Advances in Neural Information Processing Systems*, 2004.

[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE 12th Int'l Conf. Computer Vision*. IEEE, Sep. 2009, pp. 2106–2113.

[19] S. Lee, K. Kim, J. Kim, M. Kim, and H. Yoo, "Familiarity based unified visual attention model for fast and robust object recognition," *Pattern Recognition*, vol. 43, no. 3, pp. 1116–1128, Mar. 2010.

[20] C. Koch and S. Ullman, "Shifts in selection in visual attention: Toward the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[21] G. Krieger, I. Rentschler, G. Hauske, K. Schill, and C. Zetzsche, "Object and scene analysis by saccadic eye-

movements: An investigation with higher-order statistics," *Spatial Vision*, vol. 13, no. 2–3, pp. 201–214, 2000.

[22] E. Niebur and C.Koch, *Computational architectures for attention*, T. A. Brain, Ed. MIT Press, 1997.

[23] A. Yarbus, *Eye Movements and Vision*. Plenum Press, 1967.

[24] P. Reinagel and A. Zador, "Natural scene statistics at the centre of gaze," *Network-Computation in Neural Systems*, vol. 10, no. 4, pp. 341–350, 1999.

[25] J. M. Findlay, "The visual stimulus for saccadic eye movements in human observers," *Perception*, vol. 9, no. 1, pp. 7–21, 1980.

[26] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[27] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2011, pp. 409–416.

[28] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, 2008.

[29] T. Avraham and M. Lindenbaum, "Esaliency (Extended saliency): Meaningful attention using stochastic image modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 693–708, 2010.

[30] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, no. 1, pp. 77 – 123, 2003.

[31] L. Itti, G. Rees, and J. K. Tsotsos, "Models of bottom-up attention and saliency," in *Neurobiology of Attention*. San Diego, CA: Elsevier, 2005, pp. 576–582.

[32] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*, 2007, pp. 545–552.

[33] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[34] C. M. Privitera and L. W. Stark, "Algorithms for defining visual region-of-interesting: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, 2000.

[35] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 989–1005, 2009.

[36] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1597–1604.

[37] Z. Li, "A saliency map in primary visual cortex," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 9–16, 2002.

[38] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2007, pp. 1–8.

[39] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vision*, vol. 9, no. 3, pp. 1–24, 2009.

[40] C. Liu, P. C. Yuen, and G. Qiu, "Object motion detection using information theoretic spatio-temporal saliency," *Pattern Recognition*, vol. 42, no. 11, pp. 2897–2906, 2009.

[41] S. Guiasu, "Weighted entropy," *Reports on Mathematical Physics*, vol. 2, no. 3, pp. 165–179, Sep. 1971.

[42] A. Buades, B. Coll, and J. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 60–65.

[43] J. Hutchinson, "Fractals and self-similarity," *Indiana Univ. Math. J*, vol. 30, no. 5, pp. 713–747, 1981.

[44] S. Alexander, E. Vrscay, and S. Tsurumi, "A simple, general model for the affine self-similarity of images," in *Image Analysis and Recognition*, 2008, pp. 192–203.

[45] O. Boiman and M. Irani, "Detecting irregularities in images and in video," in *Tenth IEEE Int'l Conf. Computer Vision*, vol. 1, Oct. 2005, pp. 462–469.

[46] J. Maver, "Self-similarity and points of interest," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1211–1226, 2010.

[47] C. BenAbdelkader, R. Cutler, and L. Davis, "Gait recognition using image self-similarity," *EURASIP J. Applied Signal Processing*, pp. 572–585, 2004.

[48] J. Van De Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 150–156, 2006.

[49] J. van de Weijer, T. Gevers, and J. Geusebroek, "Edge and corner detection by photometric quasi-invariants," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 625–630, 2005.

[50] N. D. B. Bruce. [Online]. Available: http://www-sop.inria.fr/members/Neil.Bruce/