Visual Saliency by Selective Contrast

Qi Wang, Yuan Yuan, *Senior Member, IEEE,* and Pingkun Yan, *Senior Member, IEEE*

Abstract—**Automatic detection of salient objects in visual media (e.g., videos and images) has been attracting much attention. The detected salient objects can be utilized for segmentation, recognition, and retrieval. However, the accuracy of saliency detection remains a challenge. The reason behind this challenge is mainly due to the lack of a well-defined model for interpreting saliency formulation. To tackle this problem, this letter proposes to detect salient objects based on selective contrast. Selective contrast intrinsically explores the most distinguishable component information in color, texture, and location. A large number of experiments are thereafter carried out upon a benchmark dataset, and the results are compared with those of 12 other popular state-of-the-art algorithms. In addition, the advantage of the proposed algorithm is also demonstrated in a retargeting application.**

Index Terms—**Attention, saliency, selective contrast, visual media.**

I. INTRODUCTION

THE HUMAN visual system interprets the world in a selective manner [1]. This phenomenon makes it easy to understand that when looking at an image or a video, only part of it attracts the observer's interest while other parts are paid less attention. Similar evidence concerning selective processing of visual information abound profoundly in other experiments [2]. In order to explain the phenomenon of visual attention, many efforts have been made by different disciplines [1], [3]–[6], such as neurobiology, computer vision, cognitive psychology, and psychophysics. However, the mechanism behind these observations remains a question.

In this letter, the focus is mainly from the viewpoint of computer vision, in which visual attention is approached by saliency detection. Basically, a computational or biological model is empirically established, and then a saliency map is computed accordingly. The saliency value in the map is normally normalized to the scope of [0, 1]. The greater value the pixel has, the higher possibility it is being salient. After detection, the extracted saliency maps can be used in a broad scope of applications in computer vision.

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: crabwq@opt.ac.cn; yuany@opt.ac.cn; pingkun.yan@opt.ac.cn).

Based on the efforts by previous work, great progress has been made in the last few years [6]. However, there are still limitations for existing algorithms [7]. First, models for interpreting saliency formulation do not necessarily conform with the human visual principle, which leads to an inconsistent saliency map. Second, the features involved for calculations are limited and the feature expression is inappropriate. Considering these two limitations of the previous efforts, in this letter, a saliency detection algorithm based on selective contrast (SC) is first proposed. This means that the most selective information from the visual media is chosen for the distinctive contrast calculation. The novel characteristics that distinguish the proposed algorithm from others are mainly twofold.

First, a selective model for saliency calculation is presented. According to physiological experiments, the neuron's receptive field (RF) is sensitive to attended stimulus when multiple stimuli appear simultaneously. This observation suggests a principle that cortical cells respond preferentially to the highest-contrast stimulus in their RF [26]. Based on this principle, the designed algorithm focuses on the choice of selective and distinctive components in visual media.

Second, in accordance with the selective model, a more compact feature description is utilized. There are three features involved in this letter—color, texture, and location.

The rest of this letter is organized as follows. In Section II, related work is reviewed. In Section III, the proposed saliency detection algorithm based on selective contrast is presented. Then, in Section IV, experiments on a dataset of 1000 images are conducted to prove the effectiveness of the proposed algorithm. Furthermore, evaluation in the context of image retargeting is shown in Section V to demonstrate its successful applications of the proposed algorithm. The conclusion is made in Section VI.

II. RELATED WORK

Existing algorithms for saliency detection are generally categorized as model based and computation based.

Model based algorithms are based on the imitation of a human visual structure of a selective attention mechanism. Koch and Ullman [2] introduced a concept of attention shift. Reference [1] is another typical example of this type that is inspired by the behavior and the neuronal architecture of the human visual system. Based on this work, Walther *et al.* [8] combined the saliency system of [1] with a hierarchical recognition system to recognize objects in images. Different from the above three models, Hou and Zhang [9] presented a method for saliency detection independent of low-level features and prior knowledge. They started from the principle
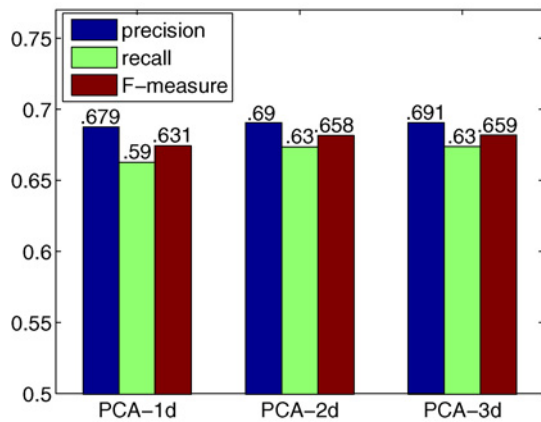
Fig. 1.   Averaged precision, recall, and F-measure in terms of different dimensions after PCA operation on RGB color space.

Fig. 2.   Precision–recall curves of saliency detection results using color and texture individually and together.

of natural image statistics and introduced a spectral residual model.

Different from model-based algorithms, the computation-based ones calculate saliency maps by contrast. Among those algorithms, most try to use different techniques to calculate contrast. For example, Frintrop *et al.* [10] captured the saliency by difference between central pixel and its surrounding area. A smart feature computation technique based on integral image is also involved to speed up the process. Han *et al.* [11] incorporated visual saliency with region growing. Harel *et al.* [12] proposed a bottom-up saliency model based on graph computation. Liu *et al.* [13] formulated the problem of saliency detection as an image segmentation task. Cheng *et al.* [6] proposed a regional saliency extraction algorithm simultaneously evaluating the global contrast and spatial coherence. Ke *et al.* [5] enhanced images and evaluated the results based on the detected saliency map. Ha *et al.* [4] tried to minimize video distortion in transmission by putting more weights on the saliency content in the media. Li and Ngan [3] employed saliency to segment video into focused foreground and uninteresting background. Judd *et al.* [14] treated saliency detection as a supervised learning problem. They constructed an eye-tracking dataset to train a support vector machine classifier as the detection model.

There are still other algorithms calculating saliency with a new definition of it. Achanta *et al.* [15] introduced a method that produces well-defined boundaries of salient objects. This is achieved by retaining substantially more frequency content of the input image than existing techniques. Goferman *et al.* [16] proposed a new type of saliency called context-aware saliency that aims at extracting the image regions representing the scene.

III. SALIENCY DETECTION ALGORITHM

Video and image are the most common visual media in everyday life. The calculation of video saliency is different from that of image. For the former, saliency is usually extracted from individual images and the spatial constraint is then added to get more consistent results. In this letter, we mainly focus on the first step, which is considered one most important
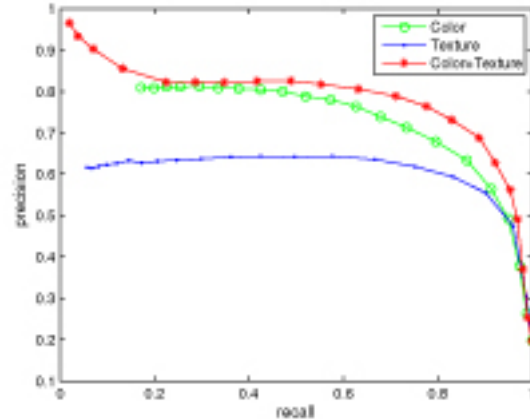
step. Thus, the discussion is constrained on images extracted from videos. Clues for contrast calculations are extracted from color, texture, and location [17].

A. Selective Color

Biological experiments indicate that opponent colors, instead of the RGB bands, are present in the early color processing of human vision system [18]. Therefore, a transformation might be applied to extract the intrinsic component of color space. For this task, several techniques including linear and nonlinear are available. Here, principle component analysis (PCA) is taken as an example to illustrate the proposed algorithm. The transformed expression of color vector by PCA is named selective color because the representation is more compact and distinctive.

Assume that the pixels in an input image are denoted as $\{X_n\}_{n=1,...,N}$, where N is the number of pixels in the image and X_n is a 3-D color vector $[r, g, b]$. The covariance matrix is then defined as

$$M = \frac{1}{N} \sum_{n=1}^{N} (X_n - \overline{X})(X_n - \overline{X})^T \qquad (1)$$

where $\overline{X} = \sum_{n=1}^{N} X_n / N$. By decomposing the covariance matrix, eigenvalues and eigenvectors can be obtained [19], [20]. The eigenvectors corresponding to larger eigenvalues are the more representative basis vectors. The original color vector is then projected to the transformed space with three orthogonal axes to get another expression. After this transformation, each pixel's saliency can be described as

$$S_c(i, R_i) = \sum_{j \neq i, j \in R_i} d_c(i, j) \qquad (2)$$

where i is the examined pixel, R_i is the supporting region for defining the saliency of pixel i, and $d_c(i, j)$ is the distance of color descriptors between i and j. We employ the ℓ_2 norm as the distance measure of color descriptors. Specifically

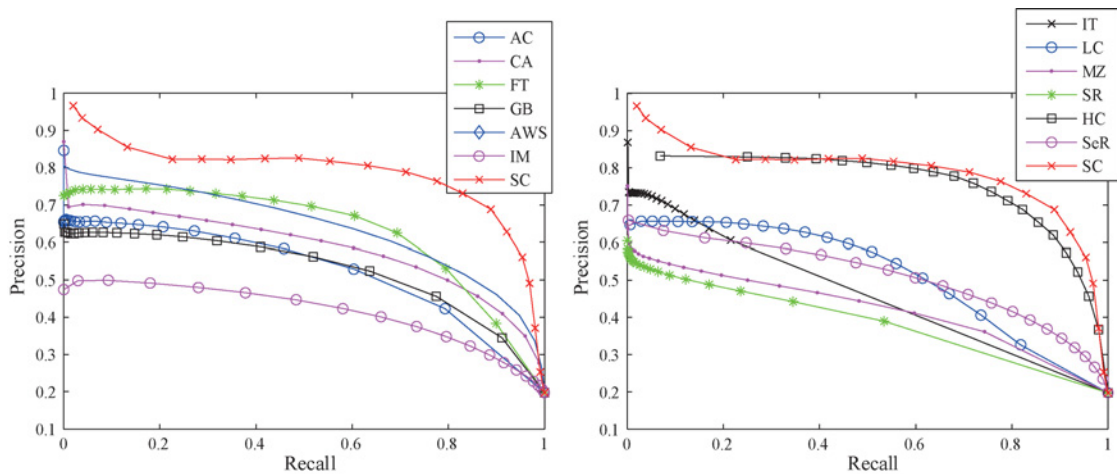$$d_c(i, j) = \|y_i - y_j\|_2 \qquad (3)$$

Fig. 3.   Precision–recall curves of saliency detection results obtained by different algorithms. This result is obtained based on 1000 images.

where y_i and y_j are the transformed color expressions of pixels i and j by PCA.

The supporting region can be as small as the 8-neighbors or as large as the whole image. But, a larger region may produce a more reliable result because it can exclude the influence of noise. In this case, the computational cost is expensive for calculating $O(N^2)$ distances. In this letter, it is chosen as the other regions different from where the examined pixel belongs.

In order to further improve the efficiency by reducing the computational complexity, the transformed color space is first clustered into k_c groups $\{c_m\}_{m=1:k_c}$ by k-means. Therefore, each pixel's color is represented by its nearest group center. By this means, the computation is reduced to looking up a distance dictionary D_c of $k_c \times k_c$ dimensions, with its element D_{ij} being the distance between the ith and j^{th} color prototypes. Accordingly, 2 should be changed into

$$S_c(i, R_i) = \sum_{\varphi(j), j \neq i} \hbar_{\varphi(j)} D_c(\varphi(i), \varphi(j)) \tag{4}$$

where $\varphi(i)$ is the function mapping pixel i to its corresponding prototype color and $\hbar_{\varphi(j)}$ is the frequency of $\varphi(j)$ in region R_i.

B. Selective Texture

An image is not just a random collection of colorful pixels, but a meaningful arrangement of them. Different organizations of these pixels form different textures, which would provide us with descriptive information. The obtained textures can be described by different ways, but are generally characterized by the outputs of a set of filters. Since these outputs span a wide range of high dimensional spaces, we hope to express them in a more compact manner. Therefore, these texture expressions are first clustered by k-means. Then, the cluster centers are treated as the representative textures. Each texture in the future is denoted as its nearest texture prototype. We call this expression selective texture.

As an example, the filter bank used in this letter is rotated copies of a Gaussian derivative and its Hilbert transform along the y-axis, which model the symmetric receptive fields of simple cells in a visual cortex [21]. To be more specific, they are

$$f_1(x, y) = \frac{d^2}{dy^2}\left(\frac{1}{C}\exp(\frac{y^2}{\sigma^2})\exp(\frac{x^2}{\ell^2\sigma^2})\right)$$
$$f_2(x, y) = Hilbert(f_1(x, y)) \tag{5}$$

where σ is the scale, ℓ is the ratio of the filter, and C is a normalization constant. Similarly, the texture saliency of pixel i is defined as

$$S_t(i, R_i) = \sum_{j \neq i, j \in R_i} d_t(i, j) \tag{6}$$

where R_i and $d_t(i, j)$ have an analogous meaning, as in Section III-A. Since the texture descriptor is a continuous-valued vector, the problem of computational complexity also exists. In order to be more efficient, a limited number of textons are trained from a set of images. This equals to saying that the textures are quantized to k_t textons. Then, a distance dictionary D_t can also be constructed to be used as a lookup table. This leads to the following form of (6):

$$S_t(i, R_i) = \sum_{\phi(j), j \neq i} \hbar_{\phi(j)} D_t(\phi(i), \phi(j)) \tag{7}$$

where these notations have a similar meaning to those in Section III-A, but they reflect the texture information.

C. Center Prone Prior

Salient objects are prone to be at the center of an image, which has been pointed out by [11] and [16]. This selective principle is mostly incorporated by putting more weights on the area in the center and less on the area near the boundary. But, in this letter, the reweighing is done on regions instead of pixels because region-based calculations can resist a certain level of noise. Therefore, the input image is first overseg-mented by a mean-shift algorithm [22] using the default parameter setting. Then, the segmented regions are treated as the basic unit for calculating saliency. We denote the obtained regions as $\{R_k\}_{k=1,...,r}$. After that, the center prone prior is added. However, it is usually found that a salient object lies at the boundary of the image and the attention is not absolutely centered. So, we adopt the principle of [6] that emphasizes
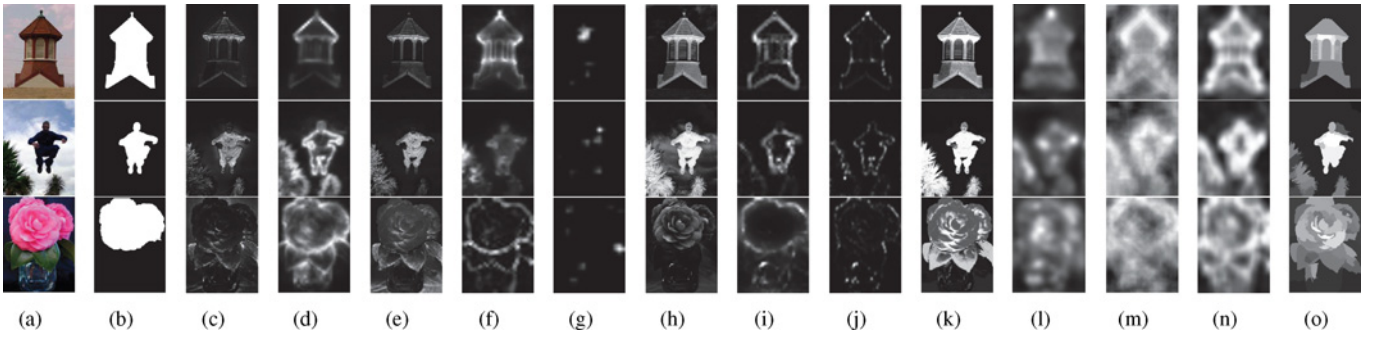
Fig. 4.   (a) Original image. (b) Ground truth. Saliency maps produced by (c) AC [23], (d) CA [16], (e) FT [15], (f) GB [12], (g) IT [1], (h) LC [24], (i) MZ [25], (j) SR [9], (k) HC [6], (l) AWS [26], (m) IM [27], (n) SeR [28], and the (o) SC proposed in this letter.
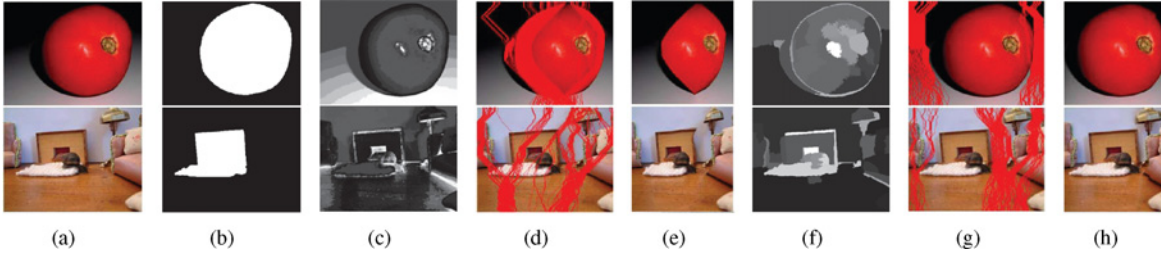

Fig. 5.   (a) Original image. (b) Ground truth. (c) Saliency map obtained by HC [6]. (d) Its corresponding carved seams. (e) Resized image. (f) Saliency map obtained by the proposed algorithm SC. (g) Its corresponding carved seams. (h) Resized image.

more on the close region and less on the far region for the examined one. Therefore, we have

$$
\begin{aligned}
S(R_k) &= \sum_{l \neq k} e^{\{-\rho_s(k,l)\}} [\lambda_c S_c(R_k, R_l) + \lambda_t S_t(R_k, R_l)] \\
&= \sum_{l \neq k} \sum_{p \in R_k} e^{\{-\rho_s(k,l)\}} [\lambda_c S_c(p, R_l) + \lambda_t S_t(p, R_l)]
\end{aligned}
\tag{8}
$$

where $S(R_k, R_l)$ is the contrast between regions R_k and R_l, λ_c and λ_t are the weighting factors between color and texture with the constraint $\lambda_c + \lambda_t = 1$, and $\rho_s(k, l)$ is the spatial distance between the centers of regions k and l. In addition, regions with μ $(0-1$, in percentage) edge pixels lying on the boundary of the image should be punished with their saliency reduced to 50%.

IV. EXPERIMENTS

In this section, experiments are done on the dataset [15] of 1000 images with ground truth labeled salient objects, among which 200 images for training while the rest 800 as testing samples. This dataset contains enough varieties and has received recognition in [6] and [15], and other works. The obtained results are finally compared with 12 popular algorithms to justify the effectiveness of the proposed method.

A. Parameter Selection

The first step in the experiments is to select proper parameters for the proposed algorithm. There are mainly three types of parameters to be decided. The first one is the number of prototype colors and textons, k_c and k_t, in addition to the parameter μ. Generally, k_c and k_t are set as 32, 64, or 128. But, experiments show that the selection of 128 is computationally infeasible, while 32 is too small to be distinguishable. Therefore, they are set to 64. As for μ, it is empirically set to 0.2.

The second parameter is the dimension of color component after PCA. The reduced color representation on the orthogonal axes may be 1-D, 2-D, or 3-D. In order to get a quantitative evaluation, three indexes based on 200 training images are calculated with respect to different dimensions. They are precision, recall, and F-measure. The statistical results are shown in Fig. 1. From the bars representing averaged precision, recall, and F-measure, it is obvious that the dimension of two is the best choice. A smaller dimension will lead to a decrease of performance while a larger one will increase no performance but the computational cost.

The third parameter is λ_c and λ_t for weighting color saliency and texture saliency. Experiments of different parameter combinations are also done on the 200 training images; the corresponding parameters $\lambda_c = 0.8$ and $\lambda_t = 0.2$ are selected as the best weighting factors.

B. Saliency Detection Results

With the selected parameters, experiments are conducted on the other 800 images. In order to justify that the combination of color and texture is more effective than the single clue, saliency maps are calculated independently using color and texture information. Then, the results are compared with the one that integrates them with 8. Fig. 2 illustrates the results. From the precision-recall curves, it is apparent that the combination of color and texture is superior to the basic clue alone. This result justifies the proposition that increased image clues support better for saliency detection. In fact, each

information channel, color or texture, can only reflect the perceptual knowledge from one aspect. But, their results might be complementary for each other. For example, when the saliency from color is not good enough, their corresponding saliency from textures may be satisfying. This phenomenon appears frequently in the experiments. In this case, the integrated saliency map ensures a consistent result. The above procedure can be explained with the early representation of the human visual system [2].

Then, the results of the proposed algorithm are compared with 12 state-of-the-art algorithms. They are, respectively, AC [23], CA [16], FT [15], GB [12], IT [1], LC [24], MZ [25], SR [9], HC [6], ASW [26], IM [27], and SeR [28]. The principle for selecting these algorithms is their recency, variety, and popularity. To be more specific, CA, HC, AWS, and IM are proposed in the last two years. GB, IT, and SR have been cited over 200 times. GB is based on graph theory, SR and FT frequency motivated, CA context related and LC spatially and temporally combined. AC, FT, and SeR are local based, and HC, MZ are global based. Our aim is to compare the proposed method with different kinds of saliency detection techniques to see whether the presented selective mechanism is effective. The codes for implementing these algorithms are from the authors' homepages or [6]. Every algorithm processes all of the 800 images and the results are compared with the ground truth labeling to get a quantitative evaluation. Fig. 3 illustrates the results. The precision–recall curves are obvious to show that the proposed algorithm outperforms the other ones. Achieving at the same precision value, the proposed algorithm can detect more salient regions; with the same recall value, the proposed algorithm is more accurate.

Several example results are also presented in Fig. 4 for qualitative evaluation. These saliency maps are representative for the algorithms' performance. A careful investigation of the details of the produced maps would find out that the proposed algorithm generates more consistent results than the other ones. The saliency of the target object is profoundly distinguished with the background. This is generally because the selective clues can be more distinguishable than other representations and the spatial constraint is more appropriate. Therefore, the proposed model can ensure that the target object stands out in a greater probability in each image.

V. APPLICATION IN RETARGETING

In order to comprehensively evaluate the proposed algorithm, the detected saliency map is further employed in an application of seam carving. Seam carving is an important technique for content-based visual retargeting [29]–[31]. In this letter, the saliency map is evaluated in the context of retargeting application by defining the energy function with it. Then, the seams with the minimum energy paths are removed according to a graph-cut framework [29]. In the end, the media is resized to 75% width of the original one and the results are judged subjectively. By this means, the effectiveness of saliency detection can be evaluated.

The algorithm of HC [6], which has a comparable performance according to the precision–recall curves in Fig. 3,

is chosen to be compared with the proposed SC algorithm. Fig. 5 shows the experimental results of seam carving. It is manifest that the saliency maps by the proposed algorithm are better than the HC maps, with a higher and consistent saliency degree in the target area. Therefore, the removed red seams in Fig. 5(d) and (g) are mainly from the background regions instead of disturbing the target ones.

VI. CONCLUSION AND FUTURE WORK

In this letter, an algorithm of saliency detection based on SC was first proposed. Motivated by the success of opponent analysis evidenced in human visual experiments, the most distinguishable and selective components of color, texture, and location were chosen and incorporated into the saliency detection framework. Experimental results showed that the presented algorithm outperformed the other 12 competitive ones in terms of the metric of precision–recall.

In the future, high level features, such as boundary, shape, and specific prior knowledge in applications, will be further explored. Other techniques for component data representation, including LLE, LPP, etc., will also be considered, especially when the feature vector is of higher dimensionality. In addition, the application in videos will also be explored, which we think is a straightforward extension.

REFERENCES

[1] 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, Nov. 1998.

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

[3] H. Li and K. N. Ngan, "Unsupervized video segmentation with low depth of field," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 12, pp. 1742–1751, Dec. 2007.

[4] H. Ha, J. Park, S. Lee, and A. C. Bovik, "Perceptually unequal packet loss protection by weighting saliency and error propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 9, pp. 1187–1199, Sep. 2010.

[5] W. Ke, C. Chen, and C. Chiu, "Bita/swce: Image enhancement with bilateral tone adjustment and saliency weighted contrast enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 3, pp. 360–364, Mar. 2011.

[6] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu, "Global contrast based salient region detection," in *Proc. CVPR*, 2011, pp. 409–416.

[7] R. Achanta and S. Susstrunk, "Global contrast based salient region detection," in *Proc. ICIP*, 2010, pp. 2653–2656.

[8] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition: A gentle way," in *Proc. 2nd Workshop Biologically Motivated Comput. Vision*, vol. 2525. Nov. 2002, pp. 251–267.

[9] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. CVPR*, 2007, pp. 1–8.

[10] S. Frintrop, M. Klodt, and E. Rome, "A real-time visual attention system using integral images," in *Proc. 5th Int. Conf. Comput. Vision Syst.*, 2007.

[11] J. Han, K. Ngan, M. Li, and H. 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.

[12] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS*, 2006, pp. 545–552.

[13] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," in *Proc. CVPR*, 2007, pp. 1–8.

[14] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. ICCV*, 2009, pp. 2106–2113.

[15] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. CVPR*, 2009, pp. 1597–1604.

[16] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. CVPR*, 2010, pp. 2376–2383.

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

[18] G. Buchsbaum and A. Gottschalk, "Trichromacy, opponent colours coding and optimum colour information transmission in the retina," *Proc. Roy. Soc. London*, vol. 220, pp. 89–113, Nov. 1983.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[20] M. Brown and S. Susstrunk, "Multi-spectral sift for scene category recognition," in *Proc. CVPR*, 2011, pp. 177–184.

[21] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *Int. J. Comput. Vision*, vol. 43, no. 1, pp. 7–27, 2001.

[22] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[23] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk, "Salient region detection and segmentation," *Comput. Vision Syst.*, vol. 5008, pp. 66–75, May 2008.

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

[25] Y. F. Ma and H. J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Multimedia*, 2003, pp. 374–381.

[26] A. García-Díaz, X. R. Fernández-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image Vision Comput.*, vol. 30, no. 1, pp. 51–64, 2012.

[27] N. Murray, M. Vanrell, X. Otazu, and C. A. Párraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. CVPR*, 2011, pp. 433–440.

[28] H. J. Seo and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *Proc. CVPR*, 2009, pp. 45–52.

[29] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–9, 2008.

[30] A. Mansfield, P. V. Gehler, L. J. V. Gool, and C. Rother, "Scene carving: Scene consistent image retargeting," in *Proc. ECCV*, 2010, pp. 143–156.

[31] D. Han, X. Wu, and M. Sonka, "Optimal multiple surfaces searching for video/image resizing: A graph-theoretic approach," in *Proc. ICCV*, 2009, pp. 1026–1033.