Saliency detection via textural contrast

Wonjun Kim* and Changick Kim

Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Yuseong-Gu, Deajeon 305-701, South Korea

**Corresponding author: jazznova@kaist.ac.kr*

Received December 5, 2011; accepted March 5, 2012;
posted March 7, 2012 (Doc. ID 159455); published May 1, 2012

We present a new approach for visual saliency detection from various natural images. It is inspired by our careful observation that the human visual system (HVS) responds sensitively and quickly to high textural contrast, derived from the discriminative directional pattern from its surroundings as well as the noticeable luminance difference, for understanding a given scene. By formulating such textural contrast within a multiscale framework, we construct a more reliable saliency map even without color information when compared to most previous approaches still suffering from the complex and cluttered background. The proposed method has been extensively tested on a wide range of natural images, and experimental results show that the proposed scheme is effective in detecting visual saliency compared to various state-of-the-art methods. © 2012 Optical Society of America

OCIS codes: 100.2960, 100.5010, 100.2000.

The human visual system (HVS) has an outstanding ability to quickly sample the most relevant regions in a given scene without any prior knowledge. Computational modeling of this system enables various applications, e.g., image retargeting, object detection, and recognition, requiring only limited processing resources. For this reason, considerable effort has been devoted to detecting salient regions, which attract the visual attention indeed, over the last few years. The basic idea of earlier work for this task is to employ high-level information (e.g., faces and pedestrians) as a useful indicator (i.e., top-down approach); however, it is hardly generalized, as the use of high-level information is not available in every image. To cope with this problem, various bottom-up approaches have been introduced, mostly based on simple low level features such as luminance, color, and orientation, followed by some center-surround operations [1–7]. This is because the local image features become stimuli of interest when they are distinguishable from their surroundings (discriminant center-surround hypothesis) [8]. On the other hand, some researchers attempt to detect irregularities as visual saliency by exploiting the frequency domain in a global view [9,10]. Even though remarkable improvements have been achieved, traditional bottom-up approaches still often fail to suppress irrelevant regions (e.g., cluttered and highly textured background) in complex scenes.

In this Letter, we propose a novel biologically inspired model for detecting salient regions from various natural images. In particular, we introduce a new stimulus of interest, highly correlated with human visual perception, i.e., textural contrast obtained from combining the difference of luminance and directional consistency between center and surrounding regions. By incorporating the proposed textural contrast into a multiscale framework, we can build more reliable saliency maps compared to traditional bottom-up models. One important advantage of the proposed method is that it greatly removes unwanted fine details while highlighting salient regions quite uniformly due to its ability to provide the contextual information regarding underlying image structures.

Specifically, motivated by the simple formulation of color contrast introduced in [4], we define our luminance contrast by allowing for the improved dynamic ranges with the nth order statistics, given as follows:

$$C^{(n)}(i) = \left| I_\mu - \frac{1}{N} \sum_{j \in B_i} I(j) \right|^n, \tag{1}$$

where I_μ denotes the mean of luminance values over the whole image (i.e., the largest surrounding region). B_i and N represent the neighbor region (5×5 pixels in our implementation) centered at the ith pixel position and its size, respectively. An example of the luminance contrast map generated by Eq. (1) is shown in Fig. 1. It is carefully observed that the second-order moment (i.e., $C^{(2)}(i)$) yields the best results by suppressing irrelevant regions while sufficiently emphasizing the salient region, i.e., a child [see Fig. 1(d)]. Furthermore, our luminance contrast map provides more reasonable response for visual saliency compared to the frequency-tuned saliency model [4], as shown in Fig. 1(b) even without color information.

With the luminance contrast, we also aim to depict the local structure of a given image based on the difference of directional consistency between center and surrounding regions. This center-surround directional pattern is invariant to simple optical variations, and it can thus provide a good approximation to the underlying image structure, which is indeed correlated with visual saliency. To do this, we allow for the structure tensor, which summarizes the dominant orientation and the energy along this direction based on the local gradient field, defined as follows:

$$\mathbf{T}(i) = \begin{bmatrix} \sum_{j \in B_i} I_x^2(j) & \sum_{j \in B_i} I_x(j) I_y(j) \\ \sum_{j \in B_i} I_x(j) I_y(j) & \sum_{j \in B_i} I_y^2(j) \end{bmatrix}, \tag{2}$$

Fig. 1. (Color online) (a) Original image, (b) color contrast by frequency-tuned model [4], (c) first-order model, (d) second-order model, and (e) fourth-order model.

 © 2012 Optical Society of America

where I_x and I_y denote the gradient in horizontal and vertical directions, respectively. The usefulness of the structure tensor defined in Eq. (2) for our task stems from the fact that the relative discrepancy between two eigenvalues (i.e., $\lambda_1 \geq \lambda_2 \geq 0$) of $\mathbf{T}(i)$ indicates how intensively gradients in the local region are distributed along the dominant direction (i.e., the degree to which those directions are consistent). In order to help clear understanding, we illustrate the distributions of gradients obtained from selected image patches as shown in Fig. 2. As can be seen, the gradients belonging to the strong edge region (①) are intensively distributed along the dominant direction compared to those of the highly textured region (②) or the flat region (③). Thus, we define our directional consistency at each pixel position as follows:

$$\phi = (\lambda_1 - \lambda_2)^2. \tag{3}$$

Here, the larger the value ϕ is, the higher the directional consistency is. Note that the average of gradients does not guarantee the reliable measure, because aligned but oppositely oriented gradients would cancel out in this average. In what follows, the center-surround directional pattern can be formulated by using the difference of directional consistency defined in Eq. (3) between center and surrounding regions as follows:

$$D(i) = \sum_{j \in W_i} |\phi(j) - \phi(i)|, \tag{4}$$

where W_i is a set of neighboring pixels centered at the ith pixel position. Note that the size of W_i is set to 7×7 pixels in our implementation. An example of the center-surround directional pattern map is shown in Fig. 2.

Since salient regions are assumed to contain both high luminance contrast and discriminative directional patterns as mentioned, the proposed saliency map is thus computed by combining such two outputs (i.e., textural contrast) as follows:

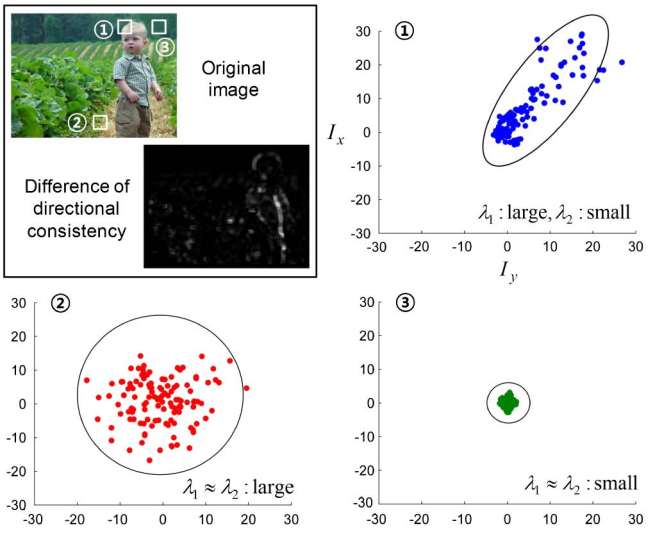

Fig. 2. (Color online) Gradients obtained from selected image patches are illustrated. Note that λ_1 and λ_2 represent the energy along the dominant orientation of the gradient field and its perpendicular direction, respectively.

$$S(i) = C^{(2)}(i) \times D(i), \tag{5}$$

where $C^{(2)}(i)$ and $D(i)$ denote the response from contrast of the luminance (generated by the second-order model) and the directional consistency at the ith pixel position, defined in Eqs. (1) and (4), respectively. Note that each response is smoothed by Gaussian filtering as in [9] and $S(i)$ is normalized to [0,255] for gray-scale representation. It is worth noting that the combination strategy defined in Eq. (5) provides the robust saliency map while effectively suppressing false positives in the background. This is because only one of two responses may be high in the background.

Since the size of the salient object is not given, visual saliency is usually computed at multiple scales. To do this, let $R = \{r_1, r_2, \cdots, r_M\}$ denote the set of scales to be considered for saliency map generation at different scales. Note that the saliency map obtained from each scale is rescaled to the size of input image (i.e., finest scale). Then, the scale-invariant saliency map is finally computed by the linear combination of outputs obtained from each scale with the same weight as follows:

$$\tilde{S}(i) = \frac{1}{M} \sum_{r \in R} S_r(i), \tag{6}$$

where S_r denotes the saliency map computed by using the scale factor r, which is subsequently rescaled to the size of the original image by using the nearest neighbor interpolation. In our implementation, we use four scale factors: $R = \{1.0, 0.7, 0.4, 0.2\}$. The scale-invariant saliency map is shown in Fig. 3. By combining outputs from each scale, we can highlight the whole region of salient objects accurately regardless of their sizes.

In this Letter, our experiments were conducted on a total of 725 images randomly collected from Microsoft Research Asia [11] and Pattern Analysis, Statistical Modeling and Computational Learning Visual Object Classes [12] databases. To show the superiority of the proposed method, we compared ours (we refer to it as TC) with various state-of-the-art methods, which are saliency tool box (STB) [1], local contrast (LC) [2], spectral residual (SR) [9], graph-based visual saliency (GB) [3],

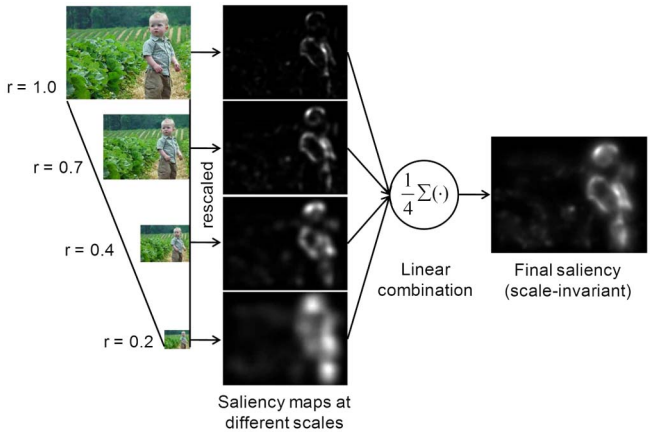

Fig. 3. (Color online) Scale-invariant saliency map. Note that the saliency map computed at each scale is resized to the size of the original image.

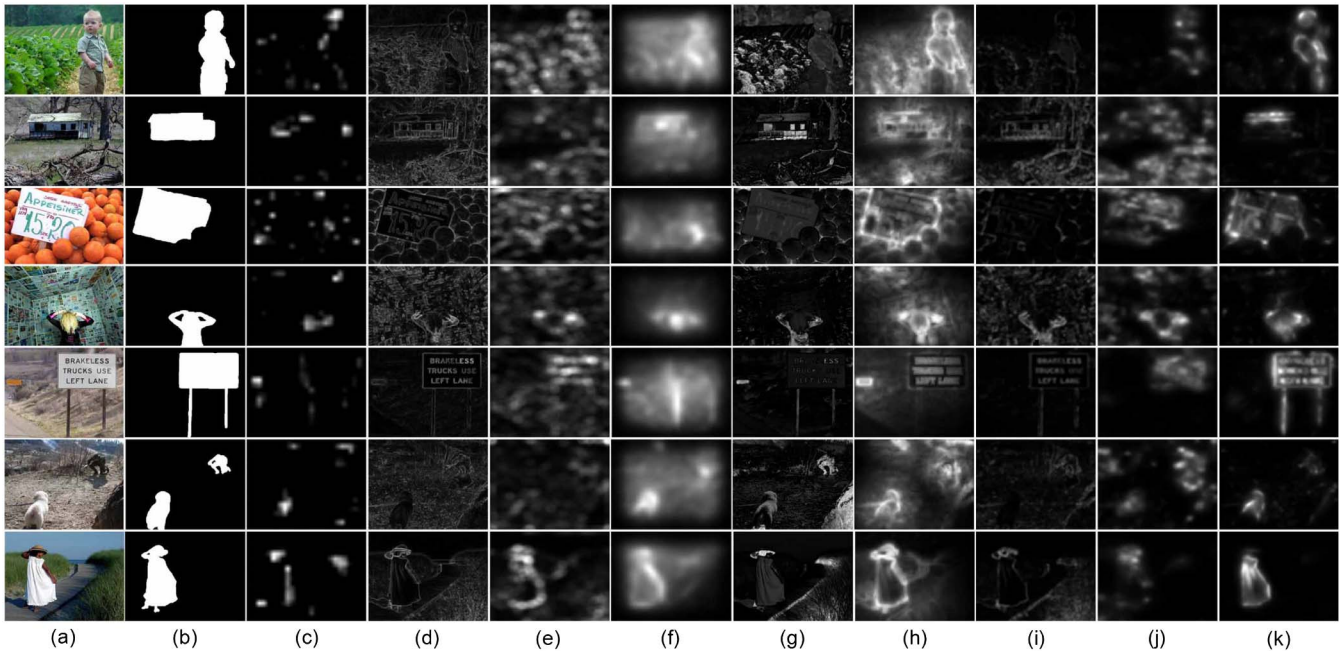

Fig. 4. (Color online) Some examples of visual saliency detection. (a) Original image, (b) ground truth, (c) saliency tool box (STB) [1], (d) local contrast method (LC) [2], (e) spectral residual (SR) [9], (f) graph-based visual saliency (GB) [3], (g) frequency-tuned method (FT) [4], (h) context-aware visual saliency (CA) [5], (i) spatial-frequency distribution (SFD), (j) difference of ordinal signatures (DOS) [7], (k) Proposed method (TC).

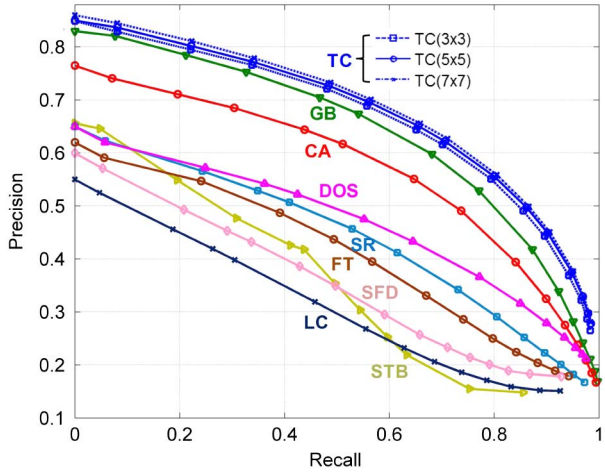

Fig. 5. (Color online) ROC curve for the quantitative analysis.

frequency-tuned (FT) [4], context-aware visual saliency (CA) [5], spatial-frequency distribution (SFD) [6], and difference of ordinal signatures (DOS) [7]. Some results of saliency detection are shown in Fig. 4. For the quantitative evaluation, we also plot the receiver operating characteristic (ROC) curve based on recall and precision as shown in Fig. 5. Note that the ground truth images are manually generated. As shown in these figures, the best performance is achieved with the proposed method. Note that the proposed method is evaluated with various block sizes for computing the structure tensor. We also demonstrate the average processing time taken by some competitive methods (i.e., SR, GB, FT, CA, DOS, and TC) in Table 1. Note that algorithms were tested using a Core2Duo 3.0 GHz machine with C implementation.

In conclusion, we have presented a novel method for saliency detection from various natural images.

Table 1. Performance Comparison of the Processing Time (Most Images Have Resolution 400 × 300 on Our Database)

Method	SR	GB	FT	CA	DOS	TC
Speed (sec)	0.02	0.93	0.03	26.88	0.06	0.58

References

1. L. Itti, C. Koch, and E. Niebur, IEEE Trans. Pattern Anal. Machine Intell. **20**, 1254 (1998).
2. Y. F. Ma and H. J. Zhang, in *Proceedings of ACM International Conference on Multimedia*, 374 (2003).
3. J. Harel, C. Koch, and P. Perona, in *Proceedings of Advances in Neural Informational Processing Systems*, 545 (2007).
4. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, Proc. IEEE, 1597 (2009).
5. S. Goferman, L. Z. Manor, and A. Tal, Proc. IEEE, 2376 (2010).
6. Y. Xu, Y. Zhao, C. Jin, J. Qu, L. Liu, and X. Sun, Opt. Lett. **35**, 475 (2010).
7. W. Kim, C. Jung, and C. Kim, IEEE Trans. Circuits Syst. Video Technol. **21**, 446 (2011).
8. D. Gao, V. Mahadevan, and N. Vasconcelos, J. Vision **8**, 13 (2008).
9. X. Hou and L. Zhang, Proc. IEEE, 1 (2007).
10. C. Guo and L. Zhang, IEEE Trans. Image Process. **15**, 185 (2010).
11. T. Liu, J. Sun, N. -N. Zheng, X. Tang, and H. -Y. Shum, Proc. IEEE, 1 (2007).
12. M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, Int. J. Comput. Vis., **88**, 303 (2010).