Extracting Salient Visual Attention Regions by Color Contrast and Wavelet Transformation

Wittawin Susutti, Chidchanok Lursinsap, and Peraphon Sophatsathit
Advanced Virtual and Intelligent Computing (AVIC) Research Center
Department of Mathematics, Faculty of Science
Chulalongkorn University, Bangkok 10330, Thailand
E-mail: dsgas064@gmail.com, lchidcha@chula.ac.th, peraphon.s@chula.ac.th

Abstract—**Visual attention detection is an important technique in many computer vision applications. In this paper, we propose an algorithm to extract a salient object from an image using bottom-up and top-down computations. In bottom-up computation, segment-based color contrast and attention values are employed to compose a bottom-up saliency map. In top-down computation, in-focus areas of the image are extracted to derive attention values using wavelet transforms for constructing a segment-based top-down saliency map. Attention values from both maps are combined by linear combination. The foreground/background-based salient object extraction is applied to form an output object. Experiments on 1,200 color images show that the proposed algorithm yields high level of satisfaction.**

I. Introduction

When a person sees an image, there are some parts or objects of the image that stimulate his visual system and brain. We call such a segment a visual attention region (VAR). Thus, for a given image, VAR refers to a region or regions that are distinguishable from other regions perceived by the viewer. Detecting the regions of interest in an image is the essential part in a wide range of computer vision researches, e.g., image recognition, scene understanding, and content-based image retrieval.

In order to arrive at a correct and efficient method for VAR identification, most researches in image visual attention detection utilize VAR through saliency map where computations are performed with the help of bottom-up attention model. Important features of VAR to be focused on are color, texture, intensity, and orientation. Itti [6] presented a visual attention model based on the properties of primate vision, employing the aforementioned bottom-up features. Contrast-based model was introduced by Ma [9], forming a saliency map by color contrast of image pixels and using face detection as a top-down feature. The attended areas were located with the help of a fuzzy growing algorithm. In still image, a pixel-level saliency map used color contrast as a feature represented in hierarchical attention regions [14]. Y.Hu [5] presented a robust subspace analysis-based VAR detection method. Simple features like hue and intensity were used, along with proposed subspace estimation algorithms based on generalized principal component analysis. In all these works, VAR was represented in circle and rectangle to delineate important features under investigation. To fulfill the gap between the semantic of image and low-level features, the rectangular attention region is not enough as it is limited by its boundary and dimensions. As such, an effective VAR representation takes a form of a salient object.

Several research endeavors have attempted to represent VAR as an object. Z. Yu [13] presented a rule-based VAR extraction based on real time clustering algorithms. They represented VAR in object based and arranged in hierarchical fashion. Z. C. Zhao [15] presented a segment based approach using shifts of focus of attention under Gestalt principle. H. Fu [3] proposed a segment based attention model that applied attention-driven image interpretation for image retrieval. They used color and texture as features to form a VAR object. Han [4] presented saliency based object extraction using a seed growing method. Nonetheless, object-based VAR detection approach uses only bottom-up computation scheme instead of top-down scheme. Oliva [10] used top-down information to control the salient object detection in an image. Ouerhani [11] proposed scene depth based VAR detection. T. Liu [8] proposed a supervised VAR detection by means of a set of multi-scale contrast, center-surround histogram, and color spatial distribution to locate the salient object. A conditional random field was learned and evaluated the results using labeled image by multiple users as top-down information for the detection process.

In this paper, we present a salient object extraction algorithm for image processing using both bottom-up and top-down computations. The former utilizes color contrast to locate the salient region of the image, while the later extracts information from the image focus to delimit the area of attention in the image. We translate the focus of the image to be the theme of a photograph for constructing a saliency map from the linear combination of VAR. A salient region extraction algorithm performs foreground/background object extraction to extract salient regions that subsequently form the salient object as the output.

The rest of the paper is organized as follows. Section II elucidates a four-step algorithm encompassing saliency map bottom-up computation, saliency map top-down computation, bottom-up and top-down convolution, and saliency object extraction. The experiments and sample results are shown in Section III. Our proposed approach is discussed, accompanied by conclusion and future work in Section IV.

978-1-4244-4522-6/09/$25.00 ©2009 IEEE

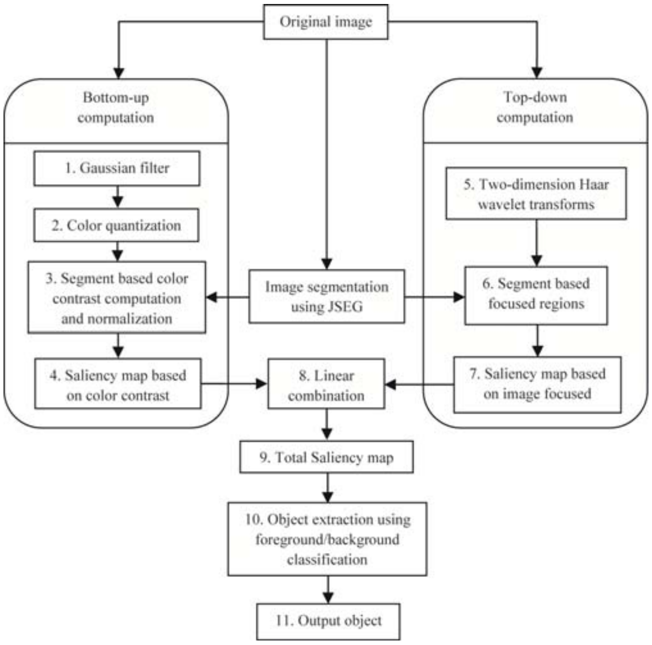

Original image
Bottom-up computation
1. Gaussian filter
2. Color quantization
3. Segment based color contrast computation and normalization
4. Saliency map based on color contrast
Image segmentation using JSEG
8. Linear combination
Top-down computation
5. Two-dimension Haar wavelet transforms
6. Segment based focused regions
7. Saliency map based on image focused
9. Total Saliency map
10. Object extraction using foreground/background classification
11. Output object

Fig. 1. The proposed visual attention framework.

II. Proposed Framework

Our proposed saliency object extraction overall work flow is shown in Fig. 1. The framework considers the saliency regions of an image on segmentation levels. The framework operates in two stages, i.e., bottom-up and top-down computations, where the total saliency value is obtained through the linear combination. In object extraction process based on foreground/background classification, there are three types of region states or seeds: foreground, background, and unlabeled seeds. The regions having high saliency value are marked as foreground seeds. The regions having low saliency value are marked as background seeds. The rest is defined as unlabeled seeds. The object extraction process forms an output saliency object from these seeds. The resulting image will be further processed by image segmentation.

Image segmentation process is an initial preprocessing work using JSEG segmentation technique [2] which is an unsupervised color-texture based image segmentation. We assume that the result of segmentation is complete. This means that we can form individual objects from segments of the segmented regions in the image. Details are described in the sections that follow.

A. Saliency map based on bottom-up computation

We use color contrast to represent the bottom-up saliency map. Color contrast is a very well-known feature that represents image saliency regions. In previous color contrast ground work [8],[9],[14], a pixel level color contrast can be determined by the color difference between the current pixel and its neighbors. Unfortunately, such a process is a computation intensive operation. To reduce computation complexity, we process color contrast computations in hierarchy. First, the original image is smoothed by Gaussian filter. Thereby, all colors in the image are quantized to 11-20 colors that do not affect image quality in human vision [2]. For each quantized color, a dominated color is selected as the representative color of the entire L*a*b* space.

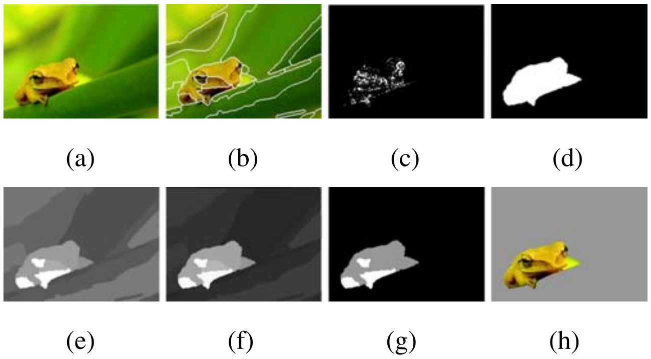


Fig. 2. An example of how the salient region is extracted by our algorithm. (a) original image (b) segmented image (c) focal points in an image using wavelet transforms (d) saliency map based on top-down computation (e) saliency map based on bottom-up computation (f) total saliency map (g) white segments represent foreground seeds, black segments represent background seeds, and gray segments represent unlabeled seeds (h) The experimental result.

A segment-based color contrast of each image segment ($SalCC$) is defined as

$$SalCC_i = \sum_{j=1}^{S} (\|A_i - A_j\| \times n_j) \qquad (1)$$

where $SalCC_i$ denotes color contrast value of segment i, A_i is the mode of the dominating color in segment i, S is the number of segments in the image, and n_j is the total pixels in segment j.

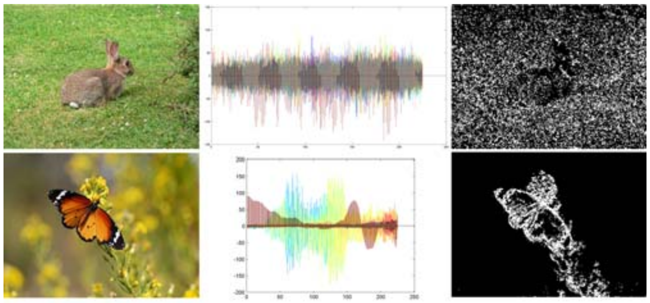
Hence, normalization is performed on all $SalCC$ values in the range [0, 1]. A bottom-up saliency map can be constructed from the $SalCC$ as shown in Fig. 2(e).

B. Saliency map based on top-down computation

The top-down computation poses difficulties in selecting appropriate features to use. If the given image is not a human picture, face and human features will be useless to incorporate in the detection process.

To overcome feature selection problem, we resort to features that can be applied in a wide range of images not being restricted to only human or specific objects. Therefore, we propose the in-focus region of an image as the top-down feature for the saliency map computation. The in-focus region is the information provided by the photographer for the viewers to indicate which regions or objects are in focus. In order to determine these in-focus regions, we use two-dimension Haar wavelet transform technique to classify in-focus regions and out-of-focus regions of the image [7].

Generally, most images are divided into two types: a low depth and a high depth of field image. Fig. 3 shows the characteristics of the low and high depth of field images as the results of applying two-dimension Haar wavelet transforms.



(a) (b) (c)

Fig. 3. Two examples of different depths. First row shows the high depth of field image and the second row shows low depth of field images. (a) Given images, rabbit and butterfly. (b) The wavelet transforms graph of both images. (c) The focal points of both images after applying wavelet transforms.

To check whether the input image is a low or high depth of field image, we compute the ratio of high frequency regions over all image regions (R). In this work, we set the maximum of R (mR) to be 0.7. From our observation on the wavelet transform results in the low depth of field images, there were many levels of the low depth of field images. If R was very close to 1, it indicated that the high frequency pixels dispersed over the entire image. Thus, the image is a high depth of field type. On the other hand, if hight frequency pixels clutter around some regions of the image, the value of R will drop farther from 1. In which case, the image is considered a low depth of field type.

From the results of wavelet transform, we extract high frequency pixels as the focal points of the image. The focal point filter of each pixel is defined as

$$focal\_point_{i,j} = \begin{cases} 1 & \text{if } |p_{i,j}| \geq std_w \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $p_{i,j}$ is the wavelet transform value of pixel at row i and column j of the image, std_w is the standard deviation of the image pixel values from wavelet transform. Some residuals may exist which appear as scattered noise in the image. These noisy spots are eliminated immediately before determining an in-focus region. The in-focus region is the segment of the image that has the focal point in the segment.

For a low depth of field image, a segment-based in-focus saliency of each segment ($SalF$) is defined as

$$SalF_i = \begin{cases} 1 & \text{if } i \text{ is an in-focus region} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

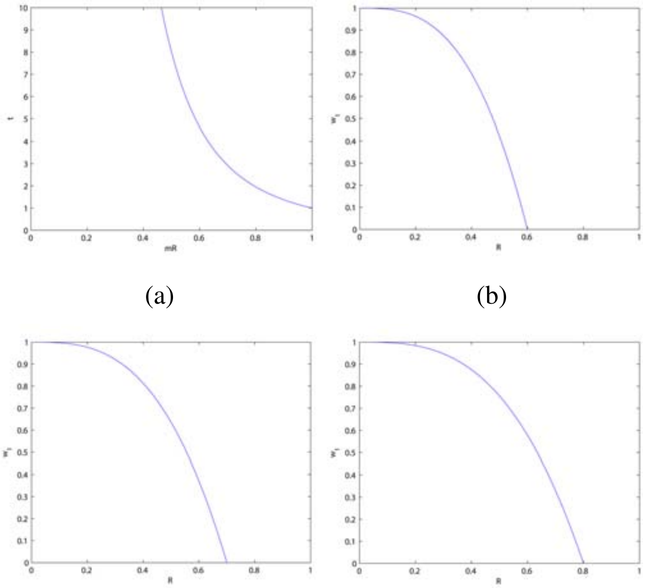
In a high depth of field image, the $SalF$ value of every segment equals to zero. Thus, we can construct a top-down saliency map from $SalF$ as shown in Fig. 2(d).

C. Bottom-up and top-down interaction

To study the interaction between both models, we apply a linear combination of top-down and bottom-up saliency map computations based on the following assumptions:

1) Regardless of the image type, the color contrast regions are still considered as a complimentary stimulus to the viewers' attention; and
2) In-focus regions of the image will attract the viewer and increase the overall attention value of the image.

From these assumptions, the saliency values of focused regions should increase if the input image is a focused image. In a de-focused image, however, the color contrast is the major feature.



(c) (d)

Fig. 4. Graphs between mR and t. (a) The relational plot of mR and t, where X-axis represents the values of mR and Y-axis represents the values of t. The relational plot of R and w_1 based on different values of mR. (b) mR equals to 0.6 and t equals to 4.6296. (c) mR equals to 0.7 and t equals to 2.9155. (d) mR equals to 0.8 and t equals to 1.9531.

The total saliency value of each segment in the image is defined by the following formula

$$SalT_i = \frac{w_1 SalF_i + SalCC_i - w_2}{1 + w_1} \tag{4}$$

where $SalT_i$ is the total saliency value of image segment i. w_1 is in-focus weight defined by (5) and w_2 by (7)

$$w_1 = \begin{cases} 1 - (t \times R^3) & \text{if } R \leq mR \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$t = \frac{1}{mR^3} \tag{6}$$

$$w_2 = SalF_i \times \frac{w_1}{3} \times e^{-\frac{SalCC_i^2}{2}} \tag{7}$$

where t is a weight value depending on mR. The relation of t and mR is described in Fig. 4. w_2 is an adjusted value of two combined saliency values derived from the effect of range of the in-focus regions. Fig. 5 shows sample results with different values of t and mR.

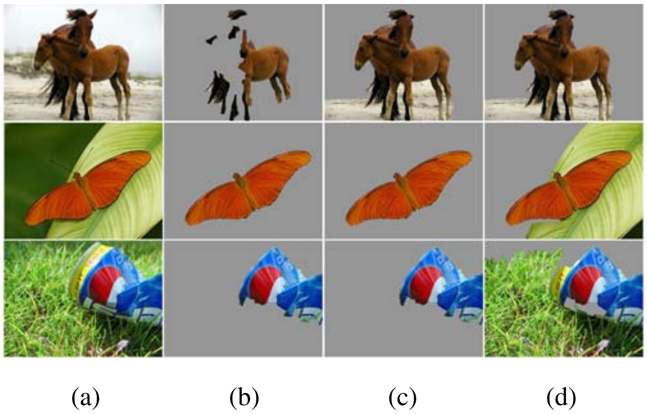


Fig. 5. Outcomes from different values of mR and t. (a) Given images. (b) mR equals to 0.6 and t equals to 4.6296. (c) mR equals to 0.7 and t equals to 2.9155. (d) mR equals to 0.8 and t equals to 1.9531.

D. Saliency object extraction

J.F. Wang [12] presented a foreground/background classification based method for object extraction. An input image can be divided into many regions by watershed technique. The regions so obtained are labeled according to foreground or background information that is provided by the user. The unlabeled neighboring regions are enqueued to form a hierarchical queues based either on color similarity or index number. The regions in the hierarchical queue are then dequeued from the lowest index number and assigned the same label as the neighboring labeled region having the smallest color distance. For unlabeled neighboring regions being dequeued that have not been labeled or are not in the hierarchical queue are enqueued with the index number. The process terminates when every region is dequeued.

We apply the above algorithm to suit the saliency object extraction with slight modifications. Instead of taking user-defined foreground seeds and background seeds, we replace the foreground seeds by the high saliency value regions and background seeds by the low saliency value regions. These seeds are subsequently classified into three types as follows

$$rst\left(i\right)=\begin{cases} fseed & \text{if } SalT_i \geq max_j\left(SalT_j\right)-std \\ bseed & \text{if } SalT_i \leq min_j\left(SalT_j\right)+\frac{std}{2} \\ nseed & \text{otherwise.} \end{cases} \quad (8)$$

where $rst(i)$ is the status of segment i, $fseed$ is the segment status to indicate foreground, $bseed$ is the segment status to indicate background. We call them $labeled$ segments. The segments that are neither foreground nor background are classified as $nseed$ or $unlabeled$ segments. std is the standard deviation of $SalT$.

We construct a region adjacency graph to represent the relation between neighboring segments of the image. In segment labeling step, segment type assignment as foreground or background is determined based on the index number of the hierarchical queue. For each unlabeled segment B adjacent to any labeled segments, we define saliency index value function as

$$q_B = floor\left(min_n\left|SalT_B - SalT_i\right| \times 100\right) \quad (9)$$

where q_B is the index number of segment B and n is the total number of labeled segments adjacent to segment B.

For the hierarchical queue, the segment C with the lowest index number is dequeued and classified as foreground or background by the formula

$$St^{*} = argmin_p\left|SalT_C - SalT_k\right| \quad (10)$$

where St^{*} is the segment status of the minimum distance of $SalT$ value between segment C and the labeled segment k. p is the total number of labeled segments adjacent to segment C.

After every segment is labeled, the results or salient objects represent the foreground segments of the image. Fig. 6 shows sample results with three types of saliency map.

III. EXPERIMENTAL AND RESULTS

We employed 1,200 test images which were divided into two sets, namely, test set A and test set B. Test set A contained 1,000 well-segmented of both low and high depth of field images that were randomly selected from standard image database [8]. Test set B contained 200 low depth of field images which were taken from various sources on the Internet. Therefore, the attention value was very subjective depending on knowledge and experience of the viewers. We considered only the images that possessed only one attention region or object. The results were rated in three levels, namely, good, accept, and failed by nine users. The descriptions of the three level ratings are given in Table I.

Table II displays the results of the experiments. In test set A, 71% of images are marked as "Good", while "Failed" marked images are 13%. Fig. 7 shows sample results on test set A. From our observation on failed case, there were two main causes involved. First, the attention object in the image was not a stand out contrast. Second, the attention object was over segmented by the segmentation algorithm, due to blurred color contrast in some regions. Thus, some parts of the attention object were missing after extraction as shown in Fig. 9.

In test set B, the results show that our approach is very effective in the low depth of field images. The "Good" rating was 87% of the images, whereas "Failed" rate occurred only 3%. Fig. 8 shows sample results on test set B.

TABLE I
DESCRIPTION OF RATING LEVELS

Rating of result	Description
Good	The saliency object is extracted with few other segments.
Accept	The saliency object is in the image result but contain some unwanted segments.
Failed	The saliency object is not in the image result or comes with many unwanted segments.

TABLE II
EXPERIMENTAL RESULTS

Test set	Level of satisfactory		
	Good	Accept	Failed
A	71%	16%	13%
B	87%	10%	3%

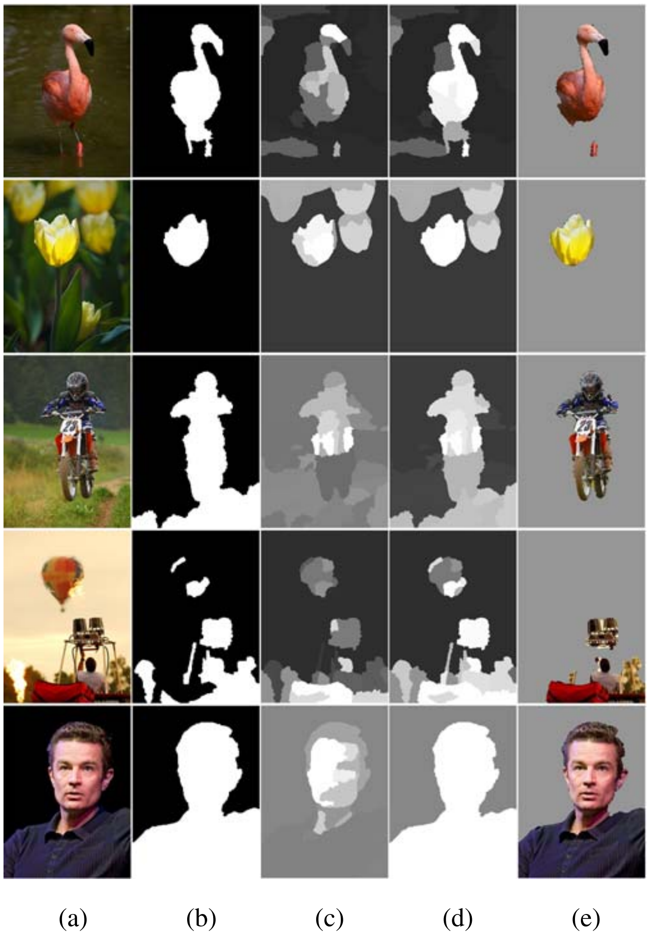


Fig. 6. Saliency object extraction outcomes. (a) Given images. (b) Saliency map based on top-down computation. (c) Saliency map based on bottom-up computation. (d) Total saliency map. (e) The experimental results.

IV. CONCLUSION

The proposed algorithm proves to be satisfactory in locating the salient region of images in various categories and compositions. The combined "Good" and "Accept" of 87% on standard test set A demonstrates the viability of the algorithm effectiveness. Such a claim is reaffirmed by additional arbitrary images from test set B that yielded 97% accuracy. Despite some essential complexities [1] that are innate to the problem under investigation as depicted in Fig. 9, our vision is set on improving related computer vision applications such as scene analysis, content-based image retrieval, image recognition, and in particular, motion picture object extraction.

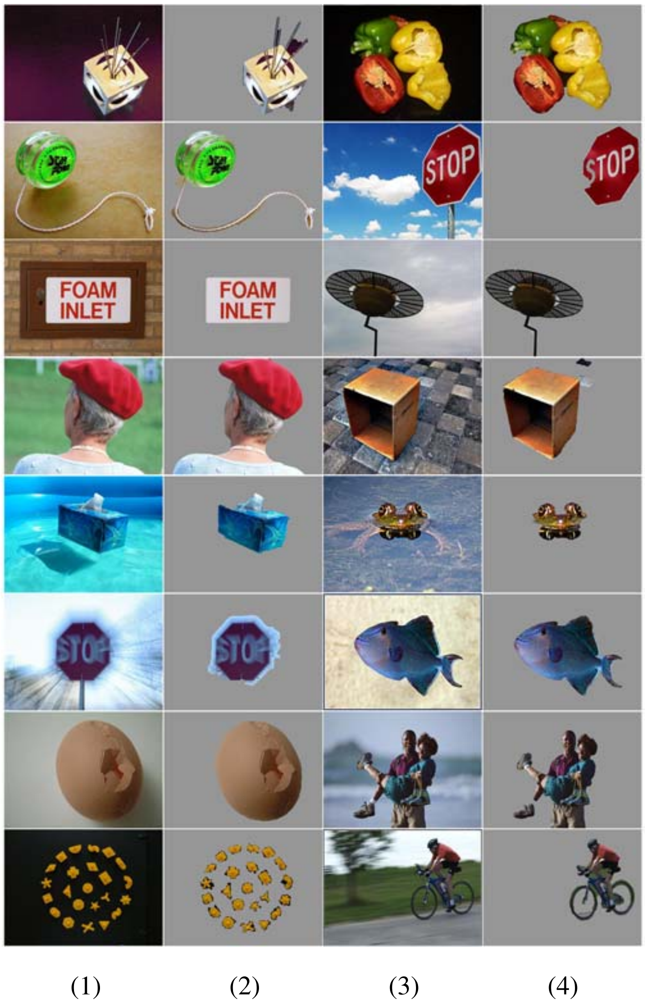


Fig. 7. Object detection outcomes on test set A. The odd columns are input images and the even columns are the experimental results.

ACKNOWLEDGMENTS

This research was supported by a grant from the Strategic Scholarships for Frontier Research Network of Thai Doctoral Program, Office of the Commission on Higher Education, Ministry of Education, Thailand.

REFERENCES

[1] F.P. Brooks, Jr., "No silver bullet: essence and accidents of software engineering", *Computer,* vol. 20, no. 4, pp. 10-19, 1987.

[2] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video", *IEEE Trans. on PAMI,* vol. 23, no. 8, pp. 800-810, 2001.

[3] H. Fu, Z. Chi, and D. Feng, "Attention-driven image interpretation with application to image retrieval", *Pattern Recognition,* vol. 39, no. 9, pp. 1604-1621, September 2006.

[4] J. Han, M. Li, H. Zhang, and L. Guo, "Automatic attention object extraction from images", *IEEE Conf. on ICIP,* vol. 2, pp. 403-406, September 2003.

[5] Y. Hu, D. Rajan, and L. T. Chia, "Detection of visual attention regions in images using robust subspace analysis", *Journal of Vision Communication and Image Representation,* vol. 19 no. 3, pp. 199-216, April 2008.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", *IEEE Trans. on PAMI,* vol. 20, no. 11, pp. 1254-1259, 1998.

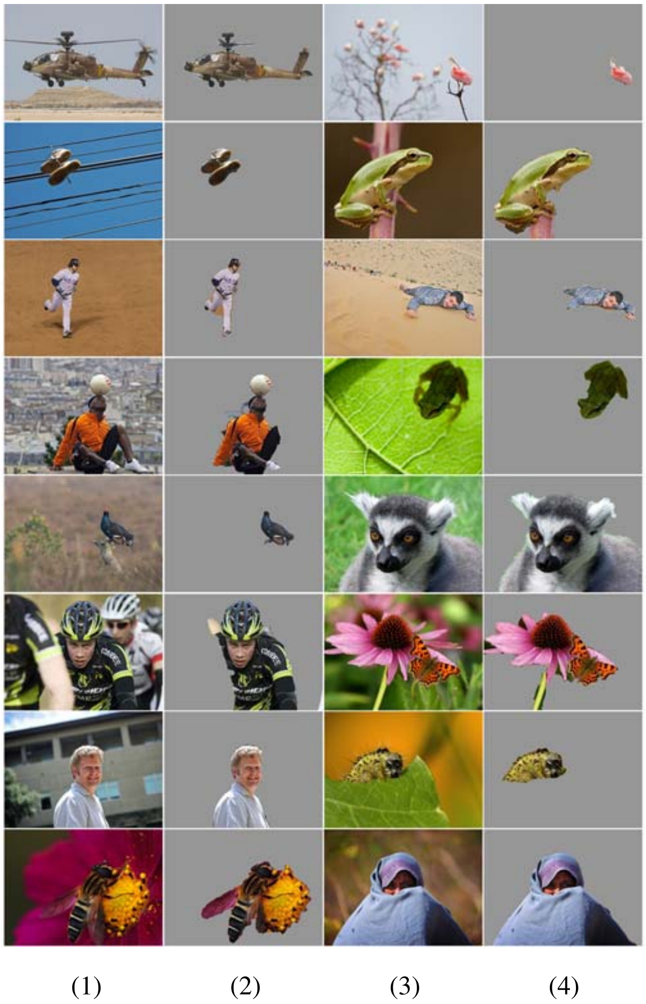


Fig. 8. Object detection outcomes on test set B. The odd columns are input images and the even columns are the experimental results.

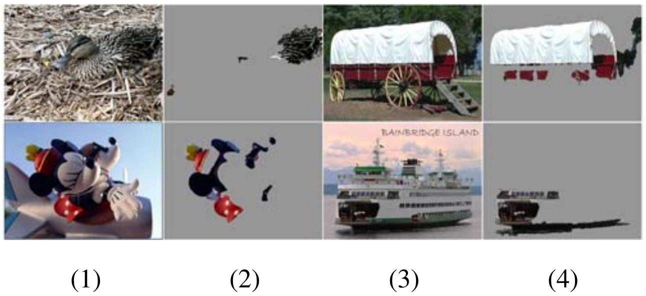


Fig. 9. Failed case of object detection outcomes. The odd columns are input images and the even columns are the experimental results.

[7] Y. Lila, C. Lursinsap, R. Lipikorn, and S. Satoh, “3D shape recovery from single image by using texture information”, *IEEE Conf. on ICCAS,* pp. 2801-2806, October 2008.

[8] T. Liu, J. Sun, N. N. Zheng, X. Tang, and H. Y. Shum, “Learning to detect a salient object”, *IEEE Conf. on CVPR,* June 2007.

[9] Y. F. Ma and H. J. Zhang, “Contrast-based image attention analysis by using fuzzy growing”, In *Proceedings of ICMM,* pp. 374-381, 2003.

[10] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, “Top-down control of visual attention in object detection”, *IIEEE Conf. on ICIP,* vol. 1, pp. 253-256, September 2003.

[11] N. Ouerhani and H. Hugli, “Computing visual attention from scene depth”, *IEEE Conf. on ICPR,* vol. 1, pp. 375-378, 2000.

[12] J. F. Wang, H. J. Hsu, and J. S. Li, “Intelligent object extraction algorithm based on foreground/background classification”, *EUC Workshops 2005,* LNCS 3823, pp. 101-110, 2005.

[13] Z. Yu and H. S. Wong, “A rule based technique for extraction of visual attention regions based on real-time clustering”, *IEEE Trans. on multimedia,* vol. 9 no. 4, pp. 766-784, June 2007.

[14] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues”, *In Proceedings of the* 14^{th} *annual ACM international conference on multimedia,* 2006.

[15] Z. C. Zhao and A. N. Cai, “Selective extraction of visual saliency objects in images and videos”, *IEEE Conf. on IIHMSP,* vol. 1, pp. 198-201, November 2007.