Journal of Robotics, Networking and Artificial Life
Vol. 7(1); June (2020), pp. 16–21
DOI: https://doi.org/10.2991/jrnal.k.200512.004; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

Evaluation of the Relationships between Saliency Maps and Keypoints

Ryuugo Mochizuki*, Kazuo Ishii

Center for Socio-Robotic Synthesis, Kyushu Institute of Technology, 2-4, Hibikino Wakamatsuku, Kitakyushu 808-0196, Japan

ARTICLE INFO

Article History

Received 26 November 2019
Accepted 19 February 2020

Keywords

Saliency map
binary robust invariant scalable keypoint
keypoint stability

ABSTRACT

The saliency map is proposed by Itti et al., to represent the conspicuity or saliency in the visual field and to guide the selection of attended locations based on the spatial distribution of saliency, which works as the trigger of bottom-up attention. If a certain location in the visual field is sufficiently different from its surrounding, we naturally pay attention to the characteristic of visual scene. In the research of computer vision, image feature extraction methods such as Scale-Invariant Feature Transform (SIFT), Speed-Up Robust Features (SURF), Binary Robust Invariant Scalable Keypoint (BRISK) etc., have been proposed to extract keypoints robust to size change or rotation of target objects. These feature extraction methods are inevitable techniques for image mosaicking and Visual SLAM (Simultaneous Localization and Mapping), on the other hand, have big influence to photographing condition change of luminance, defocusing and so on. However, the relation between human attention model, Saliency map, and feature extraction methods in computer vision is not well discussed. In this paper, we propose a new saliency map and discuss the stability of keypoints extraction and their locations using BRISK by comparing other saliency maps.

1. INTRODUCTION

In recent years, many attempts have been done such as the selection of desired information in input information [1,2]. If attention models can be constructed to select information, the intelligence and awareness of humans can be implemented in computers.

According to Itti et al., saliency is defined as the property of images, which triggers bottom-up attentions. Saliency occurs by the local conspicuity over the entire visual scene [3]. In this model (Figure 1), input image is decomposed into luminance, color, and orientation components, then, each component is processed individually with Gaussian filter.

Considering that the saliency map is applied to environment recognition by mobile robots, various changes in photographing condition are expected to affect the input image. The change affects spatial frequency components of the image. If the spatial frequency changes, the response of Gaussian filter also changes, then, the effect reflects saliency map. Considering that the saliency map is used to select the keypoints of the image, changes in the saliency map affect the results of feature selection, then, input data of detectors vary. Thus, recognition results are influenced according to the change in photographing conditions.

For keypoint extraction, small influence is desirable in spite of the variety of object size, angle and luminance. In case of the keypoint application for object detection, repetitively extracted keypoints are ideal to select.

*Corresponding author. Email: rmochizuki@lsse.kyutech.ac.jp

In our research, we propose a method for generating saliency maps, which can absorb the effect of spatial frequency changes. If the parameters of the filters can be determined automatically, the effect of the spatial frequency change can be diminished in saliency maps (Figure 2 Bottom). We evaluated the relationship between saliency and keypoints.

2. RELATED WORK

2.1. Saliency Map

Itti et al. [3] simulated human eye movement, and expressed the result as saliency maps (Figure 1). In the process of saliency map creation input image is reduced by $1/2^n$ and nine resolutions of the images are obtained. The Center and the Surround can be obtained through the smoothing operation by a common Gaussian filter. This signal process is similar to the different responses from fovea and its neighbor in retina for the common stimuli. All the reduced images are enlarged to the same size, and the across scale difference image of the two components is normalized and added to obtain a map for each component (i.e. Luminance, Color, Orientation). Saliency map is obtained through the addition of all the maps of the three components.

According to Frintrop et al. [4], saliency map changes if the parameter of the Gaussian filters are changed. The ratio of filter parameter σ_c/σ_s is crucial for the determination of saliency. Arbitral selection of σ_c/σ_s enabled high granularity in saliency map. However, in Itti et al. [3] and Frintrop et al. [4], the parameters cannot be adjusted depending on the variety of spatial frequency. As the result, saliency map can be affected in the event of spatial frequency change (Figure 2 Top).

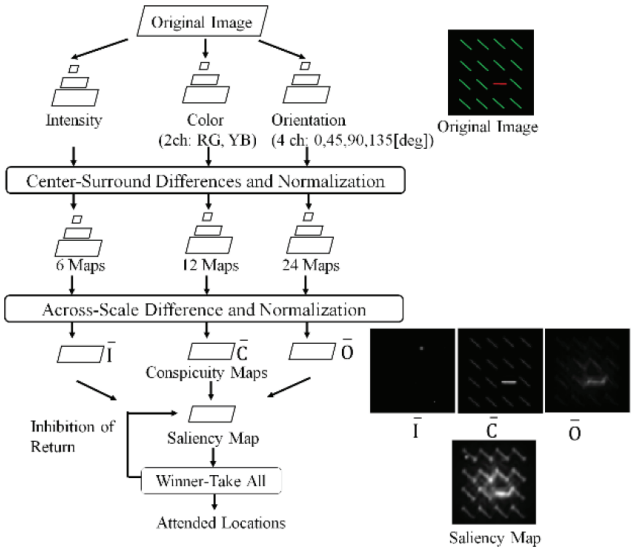

Figure 1 | Itti's saliency map.

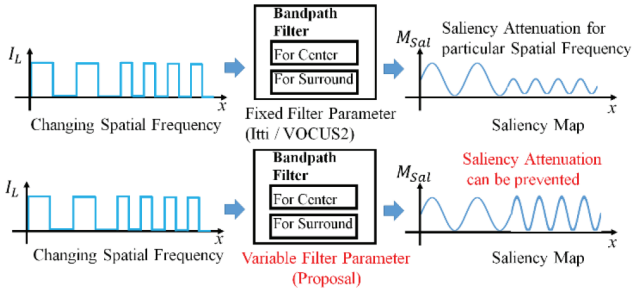

Figure 2 | Adjustment of smoothing filters under saliency map generation against changing spatial frequency.

2.2. Keypoint Extraction

Keypoint extraction is often applied for object recognition tasks [5], image stitching tasks [6], etc. by robot vision. A keypoint has a co-ordinate, a descriptor which explains brightness gradient in the neighborhood. In the object recognition task, the database image and the newly observed image are searched. Recently, scale-invariant keypoint extraction methods have been proposed, such as SIFT [7], and Binary Robust Invariant Scalable Keypoints (BRISK) [8] (Figure 3). As the result, the stability of object detection has been improved. However, if photographing conditions (brightness of the environment, size of the observed object, focusing conditions, camera internal parameters, etc.) change, the number of extracted keypoints changes significantly. Stably extracted keypoints are desirable for the use of object detection tasks by robot vision.

3. PROPOSAL OF SALIENCY MAP

3.1. Outline

In this research, we developed the theory of Frintrop et al. [4] to mitigate the effect of spatial frequency variation. The strategy is automatic adjustments of σ_c and σ_s. In the saliency map generation process (Figure 4), the input image is decomposed into luminance,

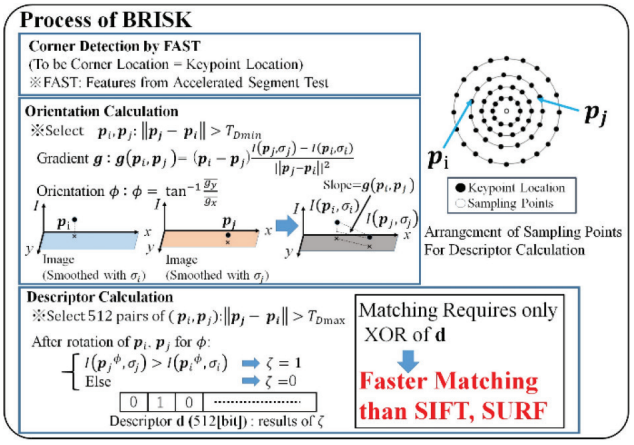

Figure 3 | BRISK Keypoint Extraction Process.

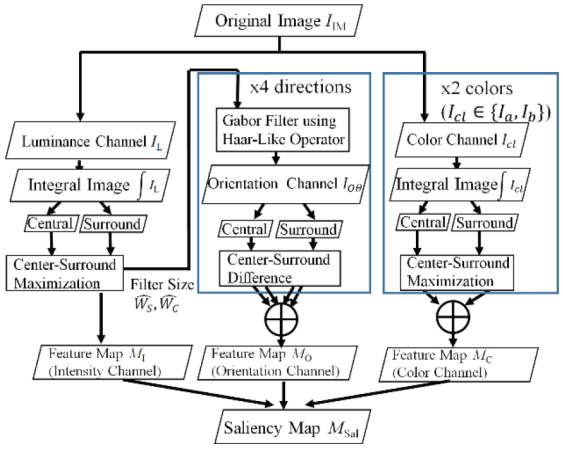

Figure 4 | Overview of proposed saliency map method.

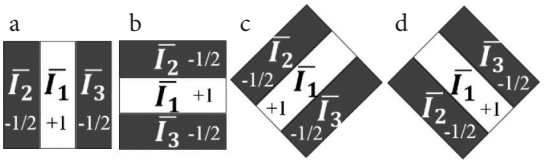

Figure 5 | Haar-like Filters. (a) 0°, (b) 90°, (c) 90°, (d) 135°.

color, and orientation components in advance. For each component, the Center and the Surround are generated by the combination of integral image and box filters. The parameter of the filters are automatically adjusted so that the pixel values of the across scale difference are maximized. The across scale differences of all three components are merged to form saliency map.

3.2. Decomposition of Input

We utilize CIE-Lab color system to simplify the difference of complimentary color channel. I_L, I_a and I_b indicates luminance, color (Red-Green), color (Blue-Yellow) component, each other. For the obtainment of orientation component, Haar-Like Filters (Figure 5 [9,10]) are convoluted on I_L. The operations are expressed as Equation (1).

$$I_\theta(\boldsymbol{p}_p) = \left| \overline{I_1(\boldsymbol{p}_p)} - \frac{1}{2}\overline{I_2(\boldsymbol{p}_p)} - \frac{1}{2}\overline{I_3(\boldsymbol{p}_p)} \right| \quad (1)$$

3.3. The Center and Surround

We align two box filters F_{Bs}, F_{Bc} centered with point $\boldsymbol{p}_p$ as shown in Figure 6. The filters are used for convolution to generate the Center and Surround. The filter widths W_{Bs}, W_{Bc} can be variable up to W_{pmax} and fulfills $W_{Bs} > W_{Bc}$. This arrangement is same as Mochizuki et al. [11].

3.4. Filter Adjustment

To obtain across scale difference of luminance, color components, we maximize the pixel value of the difference $I_{cs}(\boldsymbol{p}_p)$ as in Mochizuki et al. [11] by changing $W_{Bs}(\boldsymbol{p}_p)$, $W_{Bc}(\boldsymbol{p}_p)$ according to Equation (2) and Figure 6.

$$\begin{aligned} \widehat{\boldsymbol{I}_{cs}(\boldsymbol{p}_p)} &= \max_{W_{Bc}(\boldsymbol{p}_p), W_{Bs}(\boldsymbol{p}_p)} \boldsymbol{I}_{cs}(\boldsymbol{p}_p) \\ &= \max_{W_{Bc}(\boldsymbol{p}_p), W_{Bs}(\boldsymbol{p}_p)} \left| \boldsymbol{I}_c(\boldsymbol{p}_p) - \boldsymbol{I}_s(\boldsymbol{p}_p) \right| \end{aligned} \quad (2)$$

Here, $W_{Bs}(\boldsymbol{p}_p)$, $W_{Bc}(\boldsymbol{p}_p)$ satisfies $\widehat{W_{Bs}(\boldsymbol{p}_p)}, \widehat{W_{Bc}(\boldsymbol{p}_p)}$.

On the other hand, for orientation component, to obtain across scale differences the sizes of the Haar-like Filters are set to $\widehat{W_{Bs}(\boldsymbol{p}_p)}$, $\widehat{W_{Bc}(\boldsymbol{p}_p)}$, then, the filters are convoluted with I_L. The responses of the Center and the Surround are denoted as $I_{\theta\_c}(\boldsymbol{p}_p)$, $I_{\theta\_s}(\boldsymbol{p}_p)$. The across scale differences of all directions are obtained and merged to map $M_{O\_\theta}(\boldsymbol{p}_p)$.

3.5. Saliency Map Generation

Map M_C (for Color), and M_O (for Orientation) are obtained by Equations (4) and (5). Saliency map M_{Sal} is formed through the merge of M_I, M_C, M_O with Eq. (6). The functions f_{mix}, g_{mix}, h_{mix} for merging maps can be selected arbitrarily.

$$M_C = f(M_{Ca}, M_{Cb}) \quad (4)$$

$$M_O = g(M_{O\_0}, M_{O\_45}, M_{O\_90}, M_{O\_135}) \quad (5)$$

$$M_{Sal} = h_{mix}(M_I, M_C, M_O) \quad (6)$$

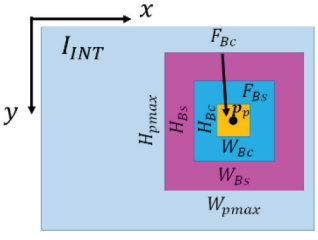

Figure 6 | Alignments of box filters.

4. EVALUATION OF THE RELATIONSHIP BETWEEN SALIENCY AND KEYPOINT EXTRACTION

4.1. Outline of the Experiment

In this experiment, we assume that the selected image keypoints are used for object detection. Thus, we evaluate the relationship between saliency $\boldsymbol{M}_{Sal}$ and feature stability $\boldsymbol{F}_{Stb}$. Suppose the number of small regions is N_q, $\boldsymbol{F}_{Stb}$ and $\boldsymbol{M}_{Sal}$ are expressed in line vector of N_q dimensions. However, we treat $\boldsymbol{M}_{Sal}$ and $\boldsymbol{F}_{Stb}$ as two dimensions (Figure 7). Then, we calculate the relationship ϕ_i by obtaining inner product $\boldsymbol{F}_{Stb} \cdot \boldsymbol{M}_{Sal}$. The saliency maps were generated by conventional methods (i.e. Itti method, VOCUS2) and our proposal to compare ϕ_i. The source codes for the experiment are Simpsal [12] by Caltech for Itti method, and [13] for VOCUS2. We chose BRISK [8] as keypoint extraction method because descriptor is expressed in binary system. Such system is reported to require shorter time for matching than SIFT [7]. Furthermore, the descriptor has properties of rotation and scale invariance.

4.2. Evaluation Function

We consider two conditions of keypoints which have high stability under photographing condition variety. First, the keypoints must extracted at the same location. We define the property as repeatability. Second, the descriptors must remain the same, that is, the similarity.

To evaluate keypoint stability, keypoint displacement have to be considered because of image flicker, resize of observed object size. For example, the combination of the same keypoints is considered as (I) or (II) in Figure 8 in different photographing condition. We define a small region of $W_q \times H_q$ [Pixels] to search identical keypoints.

Suppose $N_{kp,n_q,i}$ keypoints are extracted at n_q-th small region under i-th photographing condition, the variance σ_{kp,n_q} of

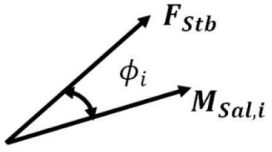

Figure 7 | Relationship between keypoint stability $\boldsymbol{F}_{Stb}$ and saliency $\boldsymbol{M}_{Sal,i}$.

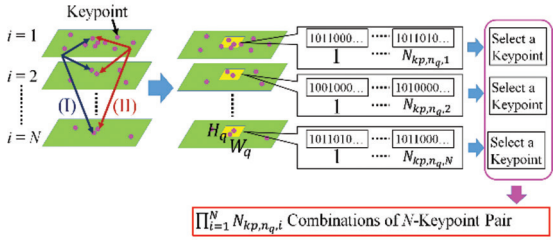

Figure 8 | Ambiguity in keypoint identification under changing photographing condition.

keypoint number is obtained by Equation (7). The average $\overline{N_{kp,n_q}}$ (for N variations of a parameter) of extracted keypoints is obtained by Equation (8).

$$\sigma_{kp,n_q} = \frac{1}{N}\sum_{i=1}^{N}(N_{kp,n_q,i} - \overline{N_{kp,n_q}})^2 \tag{7}$$

$$\overline{N_{kp,n_q}} = \frac{1}{N}\sum_{i=1}^{N} N_{kp,n_q,i} \tag{8}$$

r_{ftr,n_q} is obtained by the normalization of σ_{kp,n_q} r_{ftr,n_q} should be larger if σ_{kp,n_q} is smaller as shown in Equation (9).

$$r_{ftr,n_q} = 1 - \frac{\sigma_{kp,n_q}}{\max_{n_q}\sigma_{kp,n_q}} \tag{9}$$

To obtain similarity, we select two keypoints from the same small region (as seen in Figure 8) and different photographing conditions, then calculate Hamming distance between the two descriptors. To obtain average Hamming distance of all combinations of the keypoint pairs, we use Equation (10). The similarity s_{Dsc,n_q} is calculated with normalization by Equation (11) so that the range satisfies [0,1], and r_{Derr,n_q} is smaller as the distance is larger.

$$r_{Derr,n_q} = \frac{\min_K \sum_{l=1}^{N-1}\sum_{m=l+1}^{N} d_H(d_{n_q,l,k_l}, d_{n_q,m,k_m})}{c\binom{N}{2}} * \frac{1}{L_D} \tag{10}$$

$$\text{s.t. } \boldsymbol{K} = [k_1, k_2, k_3, \ldots, k_{N_i}], l, m = 1, 2, \ldots, N_i, l \neq m$$

$$s_{Dsc,n_q} = 1 - \frac{r_{Derr,n_q}}{\max_{n_q} r_{Derr,n_q}} \tag{11}$$

Keypoint stability of F_{Stb,n_q} is calculated by the weighting of r_{Derr,n_q} and s_{Dsc,n_q} as shown in Equation (12).

$$F_{Stb,n_q} = w r_{ftr,n_q} + (1-w) s_{Dsc,n_q} \tag{12}$$

For the saliency M_{Sal,i,n_q}, The maximum response of M_{Sal} is searched within each small region. Maximum saliency and feature stability are expressed as N_q dimensions of line vectors (denoted as $\boldsymbol{M}_{\boldsymbol{Sal,i}}$, $\boldsymbol{F}_{\boldsymbol{Stb}}$, respectively). ϕ_i is calculated as the angle between the two vectors [Equation (13)]. To be noted that r_{ftr}, s_{Dsc} are calculated only for the regions where keypoints are extracted more than twice during N variations of photographing conditions.

$$\phi_i = \cos^{-1}\left(\frac{\boldsymbol{M}_{Sal,i}\cdot\boldsymbol{F}_{Stb}}{\|\boldsymbol{M}_{Sal,i}\|\|\boldsymbol{F}_{Stb}\|}\right) \tag{13}$$

$$\boldsymbol{F}_{\boldsymbol{Stb}} = [F_{Stb,1}, F_{Stb,2}, \ldots, F_{Stb,n_q}, \ldots, F_{Stb,N_q}]$$

$$\boldsymbol{M}_{\boldsymbol{Sal,i}} = [M_{Sal,i,1}, M_{Sal,i,2}, \ldots, M_{Sal,i,n_q}, \ldots, M_{Sal,i,N_q}]$$

The average $\bar{\phi}$ is obtained according to Equation (14).

$$\bar{\phi} = \frac{1}{N}\sum_{i=1}^{N}\phi_i \tag{14}$$

4.3. Method

Figure 9 shows the experimental images (Lenna, Flower, Tree, Things). These images were selected in the database of Caltech [14] and Standard Image Data BAse (SIDBA) [15]. The spatial frequency spectrums of the images are shown in Figure 10. Lenna is well known for test image to be used image analysis. Flower has wider spectrum than Lenna with higher frequency component. As well as the comparison of Things and Tree, Things has higher frequency component than Tree.

The photographing condition to adjust to vary extracted keypoint number is $I_{Max,i}/I_{Max,1}$ for luminance, $W_{Obj,i}/W_{Obj,1}$ for object size, each other, whose range is from 0.5 to 1.0 with the step 0.1 of increase.

For changing $W_{Obj,i}$, we selected images of no white background, (i.e. Tree and Flower). We selected $T_{FAST} = 20$ (T_{FAST}: Threshold of FAST Score [8]) and $I_{Max,1} = 255$ during the adjustment of $I_{Max,i}$ and

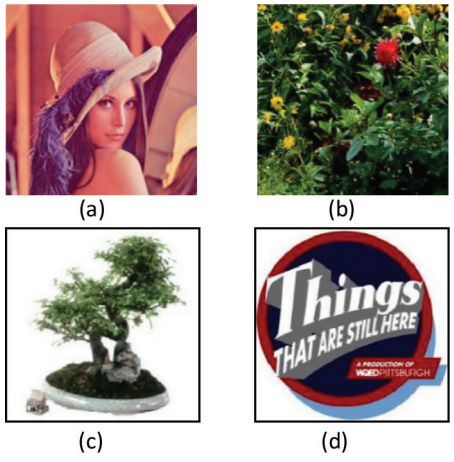

Figure 9 | Experimental images (*Cited from [14,15]). (a) Lenna. (b) Flower. (c) Tree. (d) Things.

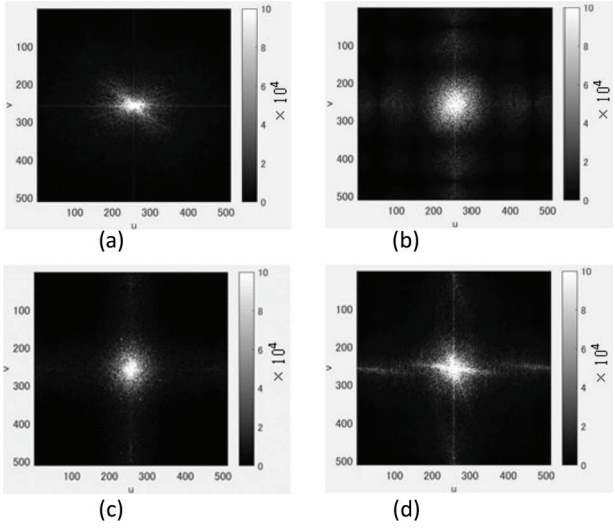

Figure 10 | Spectrum of spatial frequency which FFT algorithm returns as spatial frequency components. $u = v = 256$ for DC component. The brightness means the intensity of the component.

$W_{Obj,i}$. As the setting of the proposal, for Setting 1, $W_{pmax} = W_{IM}/4$ and for Setting 2, $W_{pmax} = W_{IM} / 2$. W_{IM} indicates the image width. The resolution of the image is $W_{IM} \times H_{IM} = 512 \times 512$ [Pixel].

4.4. Results and Discussion

Tables 1 and 2 shows the relationship of $\boldsymbol{M}_{Sal}$, $\boldsymbol{r}_{ftr}$ and $\boldsymbol{M}_{Sal}$, $\boldsymbol{s}_{Dsc}$ under variable $I_{Max,i}$, and $W_{Obj,i}$, each other. Flower has higher frequency component than Lenna, and Things has higher frequency component than Tree.

We discuss the comparison of $\bar{\phi}$ under variable $I_{Max,i}$. Referring to Tables 1 and 2 for VOCUS2 and Itti, $\bar{\phi}$ was large for high spatial frequency. While, for proposal, $\bar{\phi}$ is less influenced by spatial frequency change compared with conventional method.

Figures 11 (for Lenna) and 12 (for Flower) show the location of keypoints on saliency map (Left), the histogram which indicates the response of saliency at the locations respectively. There is difference in frequency component, however, for the case of the proposal, the location of the peak in the histogram is higher saliency than other saliency map. Thus the inner product in Equation (13) becomes larger. The change of $W_{Obj,i}$ means the image reduction, then high frequency component increases. The proposal recorded smaller $\bar{\phi}$ than others. As the results, the M_{Sal} of our proposal turned out to have larger correlation in feature stability and saliency, which means our proposal is more suitable for keypoint selection.

Table 1 | Comparison of $\bar{\phi}$ (variable $I_{Max,i}$)

Image	VOCUS2		Itti	Proposal	
	1/10	5/10		Set 1	Set 2
(a) $w = 1$ ($F_{Stb} = \boldsymbol{r}_{ftr}$)					
Lenna	18.13	20.63	31.61	19.94	**18.04**
Flower	29.00	29.64	41.39	19.72	**18.98**
Tree	21.85	24.34	33.97	20.76	**20.10**
Things	31.24	29.77	28.36	19.60	**18.37**
(b) $w = 0$ ($F_{Stb} = \boldsymbol{s}_{Dsc}$)					
Lenna	**17.20**	18.45	33.15	20.05	17.65
Flower	28.85	29.32	42.18	19.66	**18.82**
Tree	18.97	21.53	33.90	19.24	**17.27**
Things	28.58	27.40	26.67	15.21	**13.11**

Bold type indicates the best value of $\bar{\phi}$ among the all settings mentioned in the table.

Table 2 | Comparison of $\bar{\phi}$ (variable $W_{Obj,i}$)

Image	VOCUS2		Itti	Proposal	
	1/10	5/10		Set 1	Set 2
(a) $w = 1$ ($F_{Stb} = \boldsymbol{r}_{ftr}$)					
Tree	22.62	23.22	34.06	**21.87**	22.16
Things	28.74	25.83	29.56	18.96	**17.68**
(b) $w = 0$ ($F_{Stb} = \boldsymbol{s}_{dsc}$)					
Tree	24.57	25.37	35.30	**23.24**	23.88
Things	28.30	27.06	29.74	21.52	**20.09**

Bold type indicates the best value of $\bar{\phi}$ among the all settings mentioned in the table.

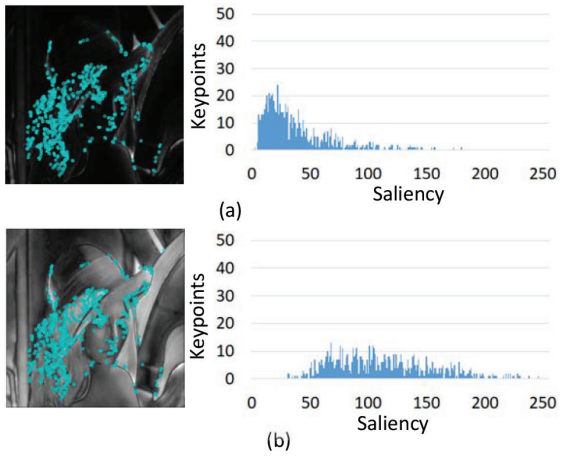

Figure 11 | Keypoint location (Lenna). (a) Itti. (b) Proposed method (setting 2).

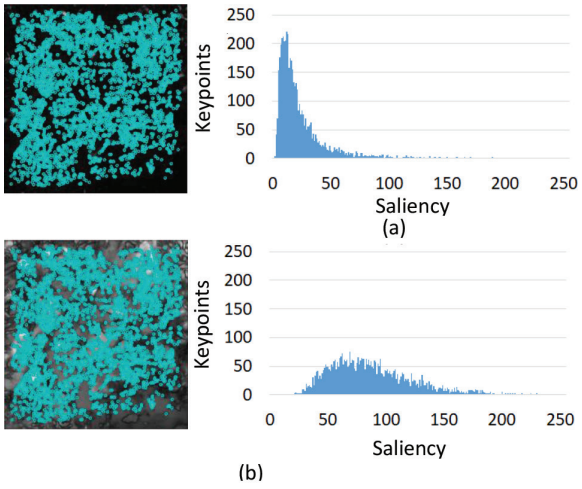

Figure 12 | Keypoint location (Flower). (a) Itti. (b) Proposed method (setting 2).

5. CONCLUSION

In this research, we proposed method of saliency map generation which consists of adaptive filter adjustment to spatial frequency to aim at preventing fluctuation of saliency caused by different input image and photographing condition change. Our saliency map method turned to be suitable for selecting keypoints less affectable by photographing condition change compared with other conventional methods.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

REFERENCES

[1] A.A. Dalve, S. Shiravale, Real time traffic signboard detection and recognition from street level imagery for smart vehicle, Int. J. Comput. Appl. 135 (2016), 18–22.

[2] H. Wang, X. Dong, J. Shen, X. Wu, Z. Chen, Saliency-based adaptive object extraction for color underwater images, Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering, Atlantis Press, Paris, France, 2013, pp. 2651–2655.
[3] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998), 1254–1259.
[4] S. Frintrop, T. Werner, G.M. García, Traditional saliency reloaded: a good old model in new shape, 2015 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Boston, MA, USA, 2015, pp. 82–90.
[5] M. Aly, M. Munich, P. Perona, Bag of words for large scale object recognition properties and benchmark, Proceedings of the Sixth International Conference on Computer Vision Theory and Applications (VISAPP), 2011, pp. 299–306.
[6] M. Brown, D.G. Lowe, Automatic panoramic image stitching using invariant features, Int. J. Comput. Vis. 74 (2007), 59–73.
[7] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004), 91–110.
[8] S. Leutenegger, M. Chili, R.Y. Siegwart, BRISK: binary robust invariant scalable keypoints, 2011 International Conference on Computer Vision, IEEE, Barcelona, Spain, 2011, pp. 2548–2555.
[9] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Kauai, HI, USA, 2001, pp. 511–518.
[10] R. Lienhart, A. Kuranov, V. Pisarevsky, Empirical analysis of detection cascades of boosted classifiers for rapid object detection, Joint Pattern Recognition Symposium, DAGM 2003: Pattern Recognition, Springer, Berlin, Heidelberg, 2003, pp. 297–304.
[11] R. Mochizuki, S. Yasukawa, K. Ishii, Proposition of saliency map based on the maximization of center-surround difference, Proceedings of the International Conference on Artificial Life and Robotics, vols. 24, 2019, pp. 487–492.
[12] MATLAB Saliency –Caltech Vision- HomePage. Available from: http://www.vision.caltech.edu/~harel/share/gbvs.php.
[13] Saliency System VOCUS2 Universität Bonn Home Page. Available from: http://pages.iai.uni-bonn.de/frintrop_simone/vocus2.html.
[14] Caltech Database Homepage. Available from: http://www.vision.caltech.edu/Image_Datasets.
[15] Kanagawa Institute of Technology. Available from: http://www.ess.ic.kanagawa-it.ac.jp/app_images_j.html.

AUTHORS INTRODUCTION

Dr. Ryuugo Mochizuki

He received his master of engineering degree at Kyushu Institute of Technology in 2008. Then he has been involved in spec test and design of Integrated Circuit as a worker in Shikino High-tech CO., LTD. up to 2013. His research topic during the doctor course was the relationship between scale-invariant keypoint extraction and saliency in the domain of image processing. He finished PhD degree in September 2019.

Dr. Kazuo Ishii

He is a Professor in the Kyushu Institute of Technology, where he has been since 1996. He received his PhD degree in engineering from University of Tokyo, Tokyo, Japan, in 1996. His research interests span both ship marine engineering and Intelligent Mechanics. He holds five patents derived from his research. His lab got “Robo Cup 2011 Middle Size League Technical Challenge 1st Place” in 2011. He is a member of the Institute of Electrical and Electronics Engineers, the Japan Society of Mechanical Engineers, Robotics Society of Japan, the Society of Instrument and Control Engineers and so on.