Evaluation of the Relationships Between Saliency Maps and Keypoints

doi: https://doi.org/10.5954/ICAROB.2020.OS24-3

Evaluation of the Relationships Between Saliency Maps and Keypoints

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

Kazuo Ishii
Center for Socio-Robotic Synthesis, Kyushu Institute of Technology,
2-4, Hibikino Wakamatsuku, Kitakyushu, 808-0196, Japan
E-mail: rmochizuki@lsse.kyutech.ac.jp, ishii@brain.kyutech.ac.jp
www.lsse.kyutech.ac.jp/~sociorobo/ja/

Abstract

Saliency is a property of images that triggers bottom-up attention. For example, if a location in an image is sufficiently different from its surrounding and worthy of paying attention, such characteristic of image is saliency. From the point of view, the location of larger saliency is outstanding visually. On the other hand, As Image Feature extraction method, such as SIFT or SURF, robust feature matching has been realized under the existence of changing size or rotation of observed target. For the consequence, its advantage has been introduced into image stitching and Visual SRAM. However, the amount of image features is susceptive to changing photographing condition, such as luminance variety, defocus-ing etc. We assumed that feature extraction stability is large in salient region because of steep bright-ness gradient. We evaluated the relationship between saliency and feature extraction stability.

Keywords: Saliency Map, Spatial-frequency, Invariant Image Feature, Filter Tuning

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, 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.

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 (Fig. 1) We evaluated the relationship between saliency and keypoints.

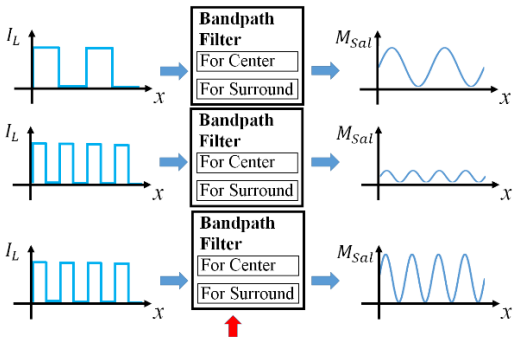


Fig. 1 Adjustment of smoothing filters against changing spatial frequency

2. Related Work

2.1. *Saliency Map*

Itti et. al. simulated human eye movement, and expressed the result as saliency maps [3]. 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 conpornents 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 [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 [3][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.

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-knvariant keypoint extraction methods have been proposed, such as SIFT[7], and BRISK[8]. 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 [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 (Fig. 2), the input image is decomposed into luminance, 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

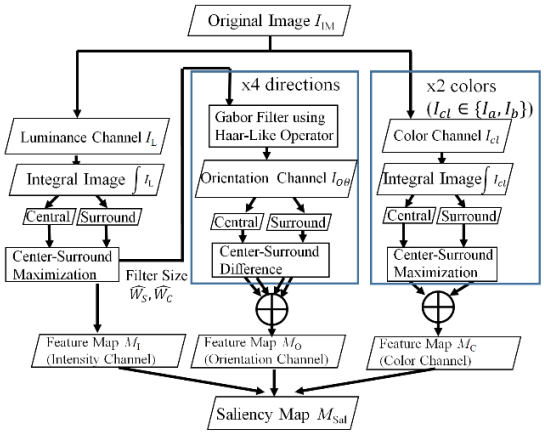


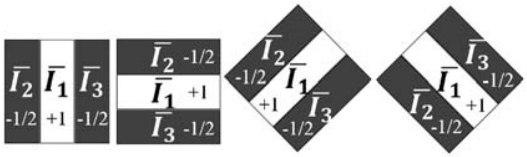
Fig. 2 Overview of proposed saliency map method

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(Fig. 4[9][10]) are convoluted on I_L. The operations are expressed as Eq. (1)

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



(a)0[deg] (b)90[deg] (c)90[deg] (d)135[deg]
Fig. 3 Haar-like Filters

3.3. *The Center and Surround*

We align two box filters F_{Bs}, F_{Bc} centered with point $\boldsymbol{p_p}$ as Fig. 4 shows. 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 [11]

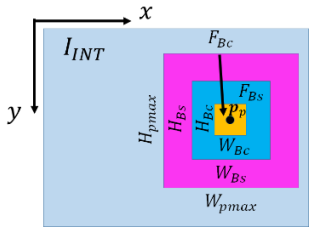


Fig. 4 Alignments of box filters

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 [11] by changing $W_{Bs}(\boldsymbol{p_p})$, $W_{Bc}(\boldsymbol{p_p})$ according to Eq. (2) and Fig. 4.

$$\widehat{I_{cs}(\boldsymbol{p_p})} = \max_{W_{Bc}(\boldsymbol{p_p}),W_{Bs}(\boldsymbol{p_p})} I_{cs}(\boldsymbol{p_p}) = \max_{W_{Bc}(\boldsymbol{p_p}),W_{Bs}(\boldsymbol{p_p})} \left|I_c(\boldsymbol{p_p}) - I_s(\boldsymbol{p_p})\right| \quad (2)$$

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 Eq. (4)(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.

$$M_C = f(M_{Ca}, M_{Cb}) \quad (4)$$
$$M_O = g(M_{O\_0}, M_{O\_45}, M_{O\_90}, M_{O\_135}) \quad (5)$$
$$M_{Sal} = h_{mix}(M_I, M_C, M_O) \quad (6)$$

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 (Fig. 5). 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.

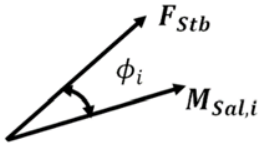


Fig. 5 Relationship between keypoint stability $\boldsymbol{F_{Stb}}$ and saliency $\boldsymbol{M_{Sal,i}}$

4.2. *Evaluation Function*

We consider two conditions of keypoints which have high stability under photographing condition variety. Firstly, the keypoints must extracted at the same location. We define the property as repeatability. Secondly, 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 Fig. 6 in different photographing condition. We define a small region of W_qxH_q [Pixels] to search identical keypoints

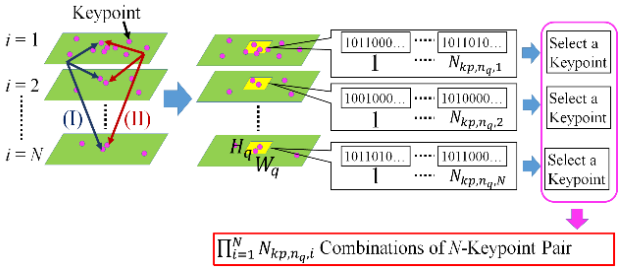


Fig. 6 Ambiguity in keypoint identification under changing photographing condition

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

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

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

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 Eq. (9) shows.

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

To obtain similarity, we select two keypoints from the same small region (as seen in Fig. 6) 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 Eq. (10). The similarity s_{Dsc,n_q} is calculated with normalization by Eq. (11) so that the range satisfies [0, 1], and r_{Derr,n_q} is smaller as the distance is larger.

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

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

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

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

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

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 $\boldsymbol{M_{Sal,i}}$, $\boldsymbol{F_{Stb}}$, respectively). ϕ_i is calculated as the angle between the two vectors (Eq. (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.

$$\phi_i = \cos^{-1}\left(\frac{\boldsymbol{M_{Sal,i}} \cdot \boldsymbol{F_{Stb}}}{\|\boldsymbol{M_{Sal,i}}\| \|\boldsymbol{F_{Stb}}\|}\right) \quad (13)$$

$$\boldsymbol{F_{Stb}} = [F_{Stb,1}, F_{Stb,2}, \dots, F_{Stb,n_q}, \dots, F_{Stb,N_q}]$$

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

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

$$\bar{\phi} = \frac{1}{N}\sum_{i=1}^{N} \phi_i \quad (14)$$

4.3. *Method*

Fig. 7 shows the experimental images (Lenna, Flower, Tree, Things). These images were selected in the database of Caltech[14] and SIDBA[15]. The spatial frequency spectrums of the images are shown in Fig.8. Lenna is a 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

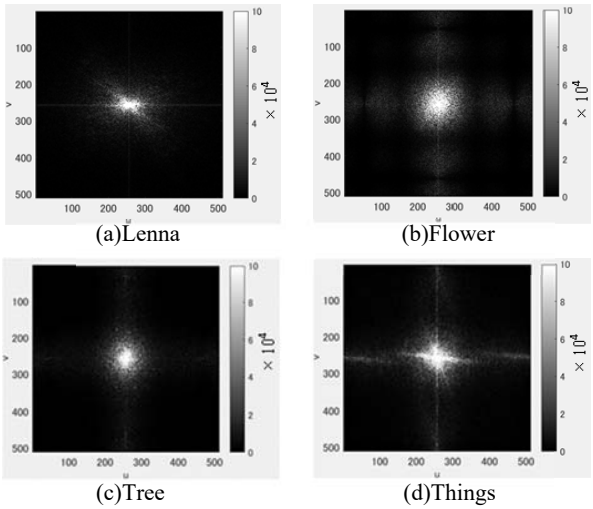
keypoint number is $I_{Max,i}/I_{Max,1}$ for luminance, $W_{Obj,i}/W_{Obj,1}$ for object size, each other, whose range is from 0.5 to 1.0 with the step 0.1 of increase.

For changing $W_{Obj,i}$, we selected images of no white backbround, (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 $W_{Obj,i}$. 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 $W_{IM} \times H_{IM}$ = 512x512 [Pixel].

(a)Lenna (b)Flower

(c)Tree (d)Things

Fig. 7 Experimental Images (※Cited from [14][15])



(a)Lenna (b)Flower

(c)Tree (d)Things

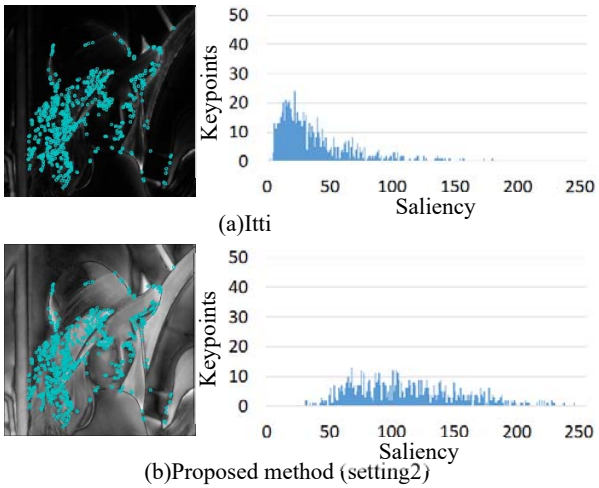
Fig. 9 Spectrum of spatial frequency

4.4. *Result and Discussion*

Table 1, 2 shows the relationship of $\boldsymbol{M_{Sal}, r_{ftr}}$ and $\boldsymbol{M_{Sal}, s_{dsc}}$ 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 Table 1, 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 to conventional method.

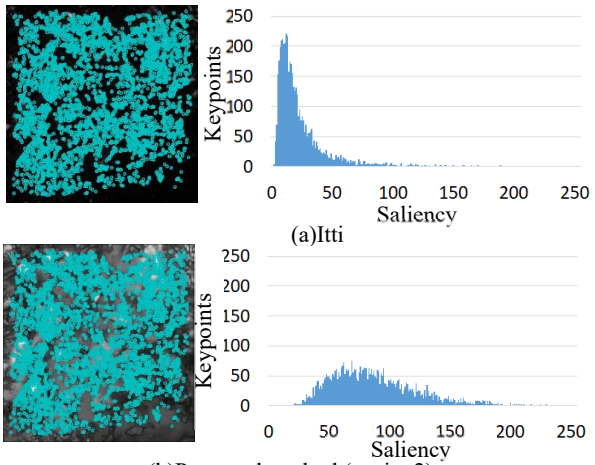
Fig.10 (for Lenna), 11 (for Flower) shows the location of keypoints on saliency map (Left), the histogram which indicates the response of saliency at the locations respectively. There is difference in frequency component, however, for the case of the proposal, the location of the peak in the histogram is higher saliency than other saliency map. Thus the inner product in Eq. (13) becomes larger. The change of $W_{Obj,i}$ means the



(a)Itti

(b)Proposed method (setting2)

Fig. 10 Keypoint location (Lenna)



(a)Itti

(b)Proposed method (setting2)

Fig. 11 Keypoint location (Flower)