Regular Paper

Salient Object Detection Based on Direct Density-Ratio Estimation

Masao Yamanaka[1,2,a)] Masakazu Matsugu[2,b)] Masashi Sugiyama[3,c)]

Abstract: Detection of salient objects in images has been an active area of research in the computer vision community. However, existing approaches tend to perform poorly in noisy environments because probability density estimation involved in the evaluation of visual saliency is not reliable. Recently, a novel machine learning approach that directly estimates the ratio of probability densities was demonstrated to be a promising alternative to density estimation. In this paper, we propose a salient object detection method based on direct density-ratio estimation, and demonstrate its usefulness in experiments.

Keywords: salient object detection, direct density-ratio estimation, relative density-ratio estimation, Shannon entropy, density estimation

1. Introduction

Detecting salient objects in images has been extensively investigated in many computer vision applications, ranging from general object detection in web images [2], over image thumbnails [10], to computing a joint focus of attention in human robot interaction [11]. Here, a salient object indicates a region in an image that visually stands out from its surroundings and is likely to attract human attention, as illustrated in Fig. 1. A key property that makes an object salient is the visual difference from the background.

Methods of salient object detection can be divided into the top-down approach based on supervised learning [2], [9] and the bottom-up approach based on unsupervised learning [3], [6]. So far, various top-down methods have been proposed, for example, Alexe et al. combined multi-scale saliency, color contrast, edge density, and super-pixels in a Bayesian framework [2], and Liu et al. combined multi-scale contrast, center-surround histograms, and color spatial-distributions with conditional random fields [9]. However, the performance of the top-down approach depends heavily on the quality and quantity of ground truth data used for supervised learning, and gathering a large number of high-quality training data is costly. Furthermore, adding a new object category is not straightforward and human subjectivity often causes ambiguity.

On the other hand, the bottom-up approach can be easily applied in an on-line fashion with no labeling cost. A seminal work by Itti et al. [6] is based on a feature integration theory in cognitive science [13]. This method identifies salient objects based on conspicuity maps that are generated from the spatial contrast of features such as the luminance value, edge intensity, and gradient orientation. While many computational models have been developed and their applications have been explored based on this structure [3], [4], [15], fusion of feature channels remains somewhat arbitrary. Furthermore, the performance of the bottom-up approach depends on the characteristics of image features—if features are sensitive to environmental and observational noise, lighting conditions and system-specific noise can cause severe performance degradation.

The above approaches are mostly motivated biologically. On the other hand, several recent approaches attempted to model saliency computationally and mathematically. For example, Kadir et al. have introduced entropy-based saliency [7], and Hou and Zhang have computed the incremental coding length to measure the perspective entropy gain [5]. However, entropy-based methods tend to identify objects with various structures as salient, which is not always appropriate in practice.

In this paper, we propose a new method to detect salient objects in images that is robust against noise incurred by various environmental and observational factors. Our approach is based on the standard structure of cognitive visual attention models [13], where several feature channels are investigated in parallel and the conspicuity maps are fused to a single saliency map. We choose the lightness, edge intensity, and color as feature channels because they are basic features of the human attention system [16].

Our saliency computation consists of two steps: First, we sample low-dimensional features such as intensities and colors in different scales. Then, in the second step, the center-surround contrast is evaluated with a machine learning technique. More

[1] Graduate School of Science and Engineering, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8502, Japan.
[2] CANON Inc., 3-30-2 Shimomaruko, Ohta-ku, Tokyo 146-8501, Japan.
[3] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan.
a) yamanaka@sp.dis.titech.ac.jp
b) matsugu.masakazu@canon.co.jp
c) sugi@cs.titech.ac.jp

Fig. 1 Examples of salient objects in images [9].

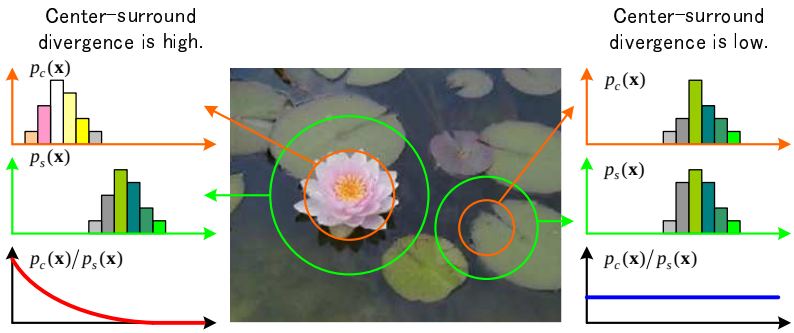


Fig. 2 Relation between visual saliency in an image and probability density ratios for low-dimensional features (e.g., color).

specifically, two probability densities $p_c(x)$ and $p_s(x)$ of visual feature occurrences are considered for a center region and a surround region, and a divergence between these densities is approximated by the state-of-the-art machine learning method called *direct density-ratio estimation* [12]: the *ratio* of probability densities $p_c(x)/p_s(x)$ is directly estimated without separate density estimation of $p_c(x)$ and $p_s(x)$. Because density estimation is known to be a hard task [14], avoiding density estimation and directly estimating the density ratio would be more promising. Through experiments, we demonstrate that the use of direct density-ratio estimation allows robust computation of visual saliency in various scales and disengages us from the need of preparing massive ground truth data.

2. Problem Formulation

In this section, we formulate our salient object detection problem based on density ratios.

Let $\boldsymbol{x}$ be a low-dimensional feature (e.g., lightness, color, and edge) extracted from an image. Our task is to detect whether there exists a salient object based on center-surround contrast based on the low-dimensional feature. A naive approach to this problem would be to first estimate the center and surround probability density functions for low-dimensional features separately, and then evaluating the difference between center and surround regions by comparing the estimated probability density functions.

However, since non-parametric density estimation is known to be a hard problem [14], this naive approach to salient object detection may not be effective. Instead, directly estimating the *ratio* of probability densities without going through density estimation would be more promising [12]. Following this discussion, we decided to base our algorithm on the density ratio for low-dimensional feature $\boldsymbol{x}$ defined by

$$w(\boldsymbol{x}) = \frac{p_c(\boldsymbol{x})}{p_s(\boldsymbol{x})},$$

where $p_c(\boldsymbol{x})$ and $p_s(\boldsymbol{x})$ are the probability density functions for center and surround low-dimensional features, respectively. We define our saliency score S based on the density ratio w by

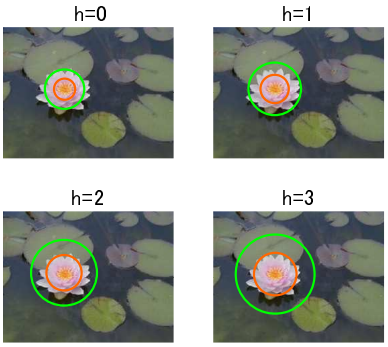


Fig. 3 Center-surround region and down sampling.

$$S = \sum_{n=1}^{N} w(\boldsymbol{x}_n),$$

where $\boldsymbol{x}_n$ and N are respectively the n-th extracted low-dimensional feature and the number of extracted low-dimensional features. Based on the score S, we detect saliency by thresholding:

$$\begin{cases} S \leq \mu & \rightarrow \text{no salient object exists,} \\ \text{otherwise} & \rightarrow \text{a salient object exists,} \end{cases}$$

where μ (≥ 0) is a predetermined threshold.

The above formulation is summarized in Fig. 2, which illustrates the relation between visual saliency in an image and the density ratio for low-dimensional features (e.g., color). The visual saliency of the center-surround region in the left-hand side of Fig. 2 is high, yielding the sum of density-ratio values to be large. On the other hand, the visual saliency of the center-surround region in the right-hand side of Fig. 2 is low, resulting in a small density-ratio values.

In practice, it may be difficult to determine the size of center-surround regions to properly detect salient objects without any preconditions. We decided to consider a hierarchy of center-surround regions with the region size decreased by factor $1/\sqrt{2}$ (see Fig. 3). This hierarchical structure makes it possible to detect salient objects in different scales from micro to macro levels. The score S_h of the h-th hierarchy is calculated as

$$S_h = \sum_{n=1}^{N} w_h(\boldsymbol{x}_n),$$

where w_h is the density ratio estimated in the h-th hierarchy. Finally, the total score S is defined by combining score S_h calculated in each hierarchy as

$$S = \sum_{h=1}^{H} S_h,$$

where H is the number of hierarchies.

If there are multiple types of low-dimensional features (e.g., lightness, color, and edge), we consider their linear combination:

$$S = \sum_{h=1}^{H} \left(S_h^{\mathrm{l}} + S_h^{\mathrm{c}} + S_h^{\mathrm{e}} \right),$$

where S_h^{l}, S_h^{c}, and S_h^{e} are respectively the scores of lightness, color, and edge in the h-th hierarchy. Finally, we build a saliency map by calculating the total score S exhaustively at all position in the image. The above formulation is summarized in Fig. 4, which denotes the schematic overview of our saliency detection system.

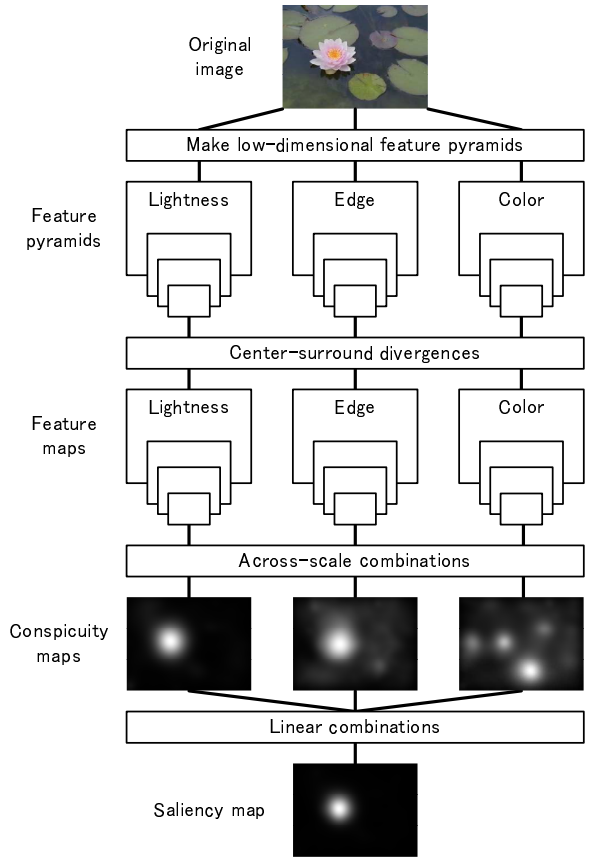


Fig. 4 Schematic overview of our saliency detection system.

The remaining question in the proposed procedure is how to estimate the density ratio function for low-dimensional features in each hierarchy:

$$w(\boldsymbol{x}) = \frac{p_{\mathrm{c}}(\boldsymbol{x})}{p_{\mathrm{s}}(\boldsymbol{x})}.$$

This is discussed in the next section.

3. Direct Density-Ratio Estimation

As described in Section 2, we use density ratio estimates for salient object detection. In this section, we show how the density ratio is directly estimated without going through density estimation.

3.1 Formulation of Density Ratio Estimation

Suppose we are given a set of N_{c} samples extracted from a center region that are drawn independently from a probability distribution P_{c} with density p_{c}:

$$\chi_{\mathrm{c}} := \left\{ \boldsymbol{x}_i^{\mathrm{c}} | \boldsymbol{x}_i^{\mathrm{c}} \in \mathfrak{R}^d \right\}_{i=1}^{N_{\mathrm{c}}} \overset{\text{i.i.d}}{\sim} P_{\mathrm{c}}.$$

We also suppose that another set of N_{s} samples extracted from a surround region that are drawn independently from (possibly) another probability distribution P_{s} with density p_{s}:

$$\chi_{\mathrm{s}} := \left\{ \boldsymbol{x}_j^{\mathrm{s}} | \boldsymbol{x}_j^{\mathrm{s}} \in \mathfrak{R}^d \right\}_{j=1}^{N_{\mathrm{s}}} \overset{\text{i.i.d}}{\sim} P_{\mathrm{s}}.$$

The goal of density ratio estimation is to estimate the density ratio function

$$w(\boldsymbol{x}) := \frac{p_{\mathrm{c}}(\boldsymbol{x})}{p_{\mathrm{s}}(\boldsymbol{x})}$$

from the samples χ_{c} and χ_{s}, where we assume $p_{\mathrm{s}}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$.

3.2 Unconstrained Least-Squares Approach to Density Ratio Estimation

Here, we review a density-ratio estimation method called *unconstrained Least-Squares Importance Fitting* (uLSIF) [8].

Let us model the density ratio function $w(\boldsymbol{x})$ by the following kernel model:

$$\widetilde{w}(\boldsymbol{x}) = \sum_{i=1}^{N_{\mathrm{c}}} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i^{\mathrm{c}}) = \boldsymbol{\alpha}' \boldsymbol{k}(\boldsymbol{x})$$

where

$$\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \ldots, \alpha_{N_{\mathrm{c}}})'$$

are parameters to be learned from data samples, $\bullet'$ denotes the transpose of a matrix or a vector,

$$\boldsymbol{k}(\boldsymbol{x}) := \left(K(\boldsymbol{x}, \boldsymbol{x}_1^{\mathrm{c}}), K(\boldsymbol{x}, \boldsymbol{x}_2^{\mathrm{c}}), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_{N_{\mathrm{c}}}^{\mathrm{c}})\right)'$$

are kernel basis functions. A popular choice of the kernel is the Gaussian function:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}\right), \tag{1}$$

where $\sigma > 0$ is the Gaussian width.

We determine the parameter $\boldsymbol{\alpha}$ in the model $\widetilde{w}(\boldsymbol{x})$ so that the following squared-error J_0 is minimized:

$$\begin{aligned} J_0 &:= \frac{1}{2}\int \left(\widetilde{w}(\boldsymbol{x}) - w(\boldsymbol{x})\right)^2 p_{\mathrm{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\int \widetilde{w}(\boldsymbol{x})^2 p_{\mathrm{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int \widetilde{w}(\boldsymbol{x}) p_{\mathrm{c}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\quad + \frac{1}{2}\int w(\boldsymbol{x})^2 p_{\mathrm{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \end{aligned} \tag{2}$$

where the last term is a constant and therefore can be safely ignored. Let us denote the first two terms by J:

$$\begin{aligned} J(\boldsymbol{\alpha}) &:= \frac{1}{2}\int \widetilde{w}(\boldsymbol{x})^2 p_{\mathrm{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int \widetilde{w}(\boldsymbol{x}) p_{\mathrm{c}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \frac{1}{2}\boldsymbol{\alpha}' \boldsymbol{H} \boldsymbol{\alpha} - \boldsymbol{h}'\boldsymbol{\alpha}, \end{aligned}$$

where $\boldsymbol{H}$ is the $N_{\mathrm{c}} \times N_{\mathrm{c}}$ matrix defined by

$$\boldsymbol{H} := \int \boldsymbol{k}(\boldsymbol{x})\boldsymbol{k}(\boldsymbol{x})' p_{\mathrm{s}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

and $\boldsymbol{h}$ is the N_{c}-dimensional vector defined by

$$\boldsymbol{h} := \int \boldsymbol{k}(\boldsymbol{x}) p_{\mathrm{c}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

3.3 Empirical Approximation

Since J contains the expectation over unknown densities $p_{\mathrm{s}}(\boldsymbol{x})$ and $p_{\mathrm{c}}(\boldsymbol{x})$, we approximate the expectations by empirical averages. Then we obtain

$$\begin{aligned} \widehat{J}(\boldsymbol{\alpha}) &:= \frac{1}{2N_{\mathrm{s}}}\sum_{j=1}^{N_{\mathrm{s}}} \widetilde{w}(\boldsymbol{x}_j^{\mathrm{s}})^2 - \frac{1}{N_{\mathrm{c}}}\sum_{i=1}^{N_{\mathrm{c}}} \widetilde{w}(\boldsymbol{x}_i^{\mathrm{c}}) \\ &= \frac{1}{2}\boldsymbol{\alpha}'\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \boldsymbol{\alpha}'\widehat{\boldsymbol{h}}, \end{aligned}$$

where $\widehat{\boldsymbol{H}}$ is the $N_{\mathrm{c}} \times N_{\mathrm{c}}$ matrix defined by

$$\widehat{\boldsymbol{H}} := \frac{1}{N_{\mathrm{s}}}\sum_{j=1}^{N_{\mathrm{s}}} \boldsymbol{k}(\boldsymbol{x}_j^{\mathrm{s}})\boldsymbol{k}(\boldsymbol{x}_j^{\mathrm{s}})',$$

and $\widehat{\boldsymbol{h}}$ is the N_{c}-dimensional vector defined by

$$\widehat{\boldsymbol{h}} := \frac{1}{N_{\mathrm{c}}}\sum_{i=1}^{N_{\mathrm{c}}} \boldsymbol{k}(\boldsymbol{x}_i^{\mathrm{c}}).$$

By including a regularization term, the uLSIF optimization problem is formulated as

$$\widehat{\boldsymbol{\alpha}} := \operatorname*{argmin}_{\boldsymbol{\alpha}}\left[\frac{1}{2}\boldsymbol{\alpha}'\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \boldsymbol{\alpha}'\widehat{\boldsymbol{h}} + \frac{\lambda}{2}\boldsymbol{\alpha}'\boldsymbol{\alpha}\right],$$

where $\boldsymbol{\alpha}'\boldsymbol{\alpha}/2$ is a regularizer and $\lambda(\geq 0)$ is the regularization parameter that controls the strength of regularization. By taking the derivative of the above objective function with respect to the parameter $\boldsymbol{\alpha}$ and equating it to zero, we can analytically obtain the solution $\widehat{\boldsymbol{\alpha}}$ as

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_{N_{\mathrm{c}}})^{-1}\widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}_N$ is the N-dimensional identity matrix. Finally, a density ratio estimator $\widehat{w}(\boldsymbol{x})$ is given by

$$\widehat{w}(\boldsymbol{x}) := \widehat{\boldsymbol{\alpha}}'\boldsymbol{k}(\boldsymbol{x}).$$

Thanks to the simple analytic-form expression, uLSIF is computationally more efficient than alternative density-ratio estimators which involve non-linear optimization [12].

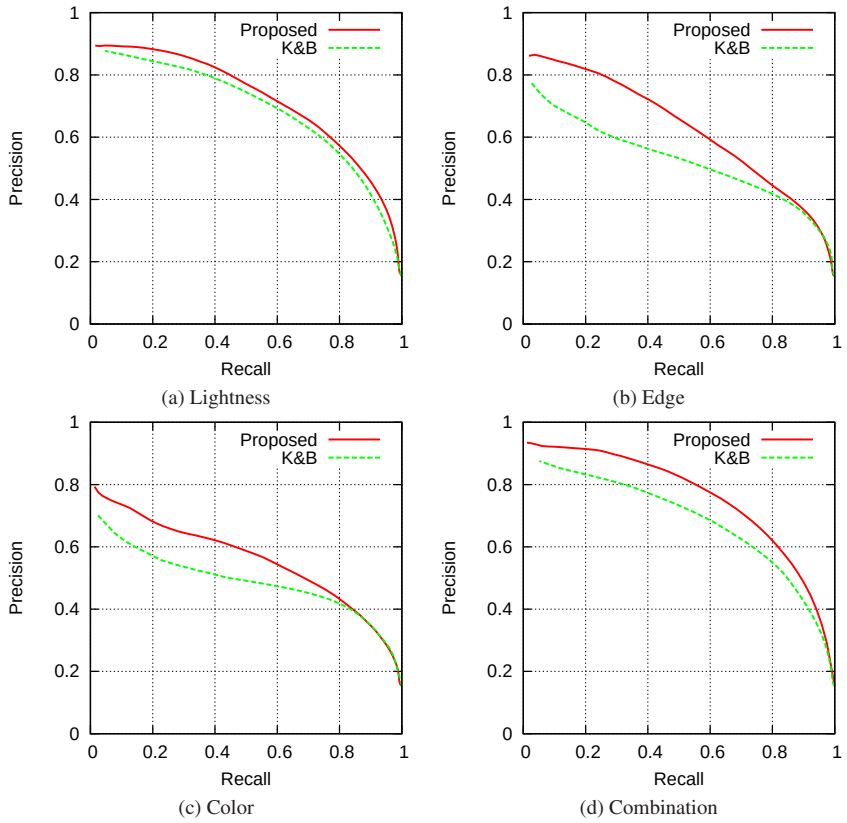
3.4 Model Selection by Cross-Validation

The practical performance of uLSIF depends on the choice of the kernel function (e.g., the kernel width σ in the case of Gaussian kernel Eq.(1)) and the regularization parameter λ. Model selection of uLSIF is possible based on *cross-validation* with respect to the error criterion J defined by Eq.(2).

More specifically, each of the sample sets $\chi_{\mathrm{c}} = \{\boldsymbol{x}_i^{\mathrm{c}}\}_{i=1}^{N_{\mathrm{c}}}$ and $\chi_{\mathrm{s}} = \{\boldsymbol{x}_j^{\mathrm{s}}\}_{j=1}^{N_{\mathrm{s}}}$ is divided into L disjoint sets $\{\chi_{\mathrm{c}}^l\}_{l=1}^L$ and $\{\chi_{\mathrm{s}}^l\}_{l=1}^L$. Then an uLSIF solution $\widetilde{w}_l(\boldsymbol{x})$ is obtained using $\chi_{\mathrm{c}}\backslash\chi_{\mathrm{c}}^l$ and $\chi_{\mathrm{s}}\backslash\chi_{\mathrm{s}}^l$ (i.e., all samples without χ_{c}^l and χ_{s}^l), and its J-value for the hold-out samples χ_{c}^l and χ_{s}^l is computed as

$$\widehat{J}_{\mathrm{CV}}^l := \frac{1}{2|\chi_{\mathrm{s}}^l|}\sum_{\boldsymbol{x}_{\mathrm{s}}\in\chi_{\mathrm{s}}^l} \widetilde{w}_l(\boldsymbol{x}_{\mathrm{s}})^2 - \frac{1}{|\chi_{\mathrm{c}}^l|}\sum_{\boldsymbol{x}_{\mathrm{c}}\in\chi_{\mathrm{c}}^l} \widetilde{w}_l(\boldsymbol{x}_{\mathrm{c}}),$$

where $|\chi|$ denotes the number of elements in the set χ. This procedure is repeated for $l = 1, \ldots, L$, and the average of $\widehat{J}_{\mathrm{CV}}^l$ over all l is computed as



(a) Lightness (b) Edge

(c) Color (d) Combination

Fig. 5 Precision-recall curves.

$$\widehat{J}_{\mathrm{CV}} := \frac{1}{L}\sum_{l=1}^{L}\widehat{J}^{l}_{\mathrm{CV}}.$$

Finally, the model (the kernel width σ and the regularization parameter λ in the current setup) that minimizes $\widehat{J}^{l}_{\mathrm{CV}}$ is chosen as the most suitable one.

3.5 Relative Density-Ratio Estimation

Depending on the condition of the denominator density $p_s(\boldsymbol{x})$, the density-ratio value $p_c(\boldsymbol{x})/p_s(\boldsymbol{x})$ can be unbounded (i.e., it can be infinity). This is actually problematic because the non-parametric convergence rate of uLSIF is governed by the sup-norm of the true density-ratio function [17]:

$$\max_{\boldsymbol{x}} \frac{p_{\mathrm{c}}(\boldsymbol{x})}{p_{\mathrm{s}}(\boldsymbol{x})}.$$

To overcome this problem, *relative density-ratio estimation* [17] was introduced. The *β-relative density ratio* is defined by

$$\frac{p_{\mathrm{c}}(\boldsymbol{x})}{\beta p_{\mathrm{c}}(\boldsymbol{x}) + (1-\beta)p_{\mathrm{s}}(\boldsymbol{x})}.$$

This is reduced to the plain density ratio if $\beta = 0$, and it tends to be smoother as β gets larger. More specifically, it can be confirmed that the β-relative density ratio is bounded above by $1/\beta$ for $\beta > 0$, even when the plain density ratio $p_{\mathrm{c}}(\boldsymbol{x})/p_{\mathrm{s}}(\boldsymbol{x})$ is unbounded. Thus, estimation of the relative density ratio is expected to be more reliable than that of the plain density ratio.

The β-relative density ratio can be learned in the same way as the plain density ratio. Indeed, the optimization problem of a relative variant of uLSIF, called RuLSIF, is given as the same form as uLSIF; the only difference is the definition of the matrix $\widehat{\boldsymbol{H}}$:

$$\widehat{\boldsymbol{H}} = \frac{\beta}{N_{\mathrm{c}}}\sum_{i=1}^{N_{\mathrm{c}}}\boldsymbol{k}(\boldsymbol{x}^{\mathrm{c}}_i)\boldsymbol{k}(\boldsymbol{x}^{\mathrm{c}}_i)' + \frac{1-\beta}{N_{\mathrm{s}}}\sum_{j=1}^{N_{\mathrm{s}}}\boldsymbol{k}(\boldsymbol{x}^{\mathrm{s}}_j)\boldsymbol{k}(\boldsymbol{x}^{\mathrm{s}}_j)'.$$

Thus, the RuLSIF solution can still be obtained analytically. Furthermore, RuLSIF was proved to possess an even better non-parametric convergence property than uLSIF [17]. For this reason, we decided to use RuLSIF to compute our saliency score S, where we use all samples $\{\boldsymbol{x}^{\mathrm{c}}_i\}_{i=1}^{N_{\mathrm{c}}}$ and $\{\boldsymbol{x}^{\mathrm{s}}_j\}_{j=1}^{N_{\mathrm{s}}}$ to compute S.

4. Experiments

In this section, we experimentally compare the proposed method with the method proposed by Kadir and Brady [7] (K&B), on the *MSRA salient object database* [9].

In the K&B method, visual saliency S is defined as

$$S(r) = H(r)\cdot W(r), \tag{3}$$

where $H(r)$ denotes the Shannon entropy in the local region with size r:

$$H(r) = -\int p(I,r)\log p(I,r)\mathrm{d}I.$$

Here, $p(I,r)$ represents the probability density for low-dimensional feature I (e.g., lightness, edge, and color). $W(r)$ in Eq.(3) is a weight function defined by

$$W(r) = \frac{r^2}{2r-1}\int\left|\frac{\partial p(I,r)}{\partial r}\right|\mathrm{d}I.$$

Fig. 1 shows some images in the database[*1] and ground-truth

[*1] As pointed out in Liu et al. [9], saliency detection for images with large objects is too easy. Here, we choose 200 images from the database that contain small objects of size less than 20% of the image size.

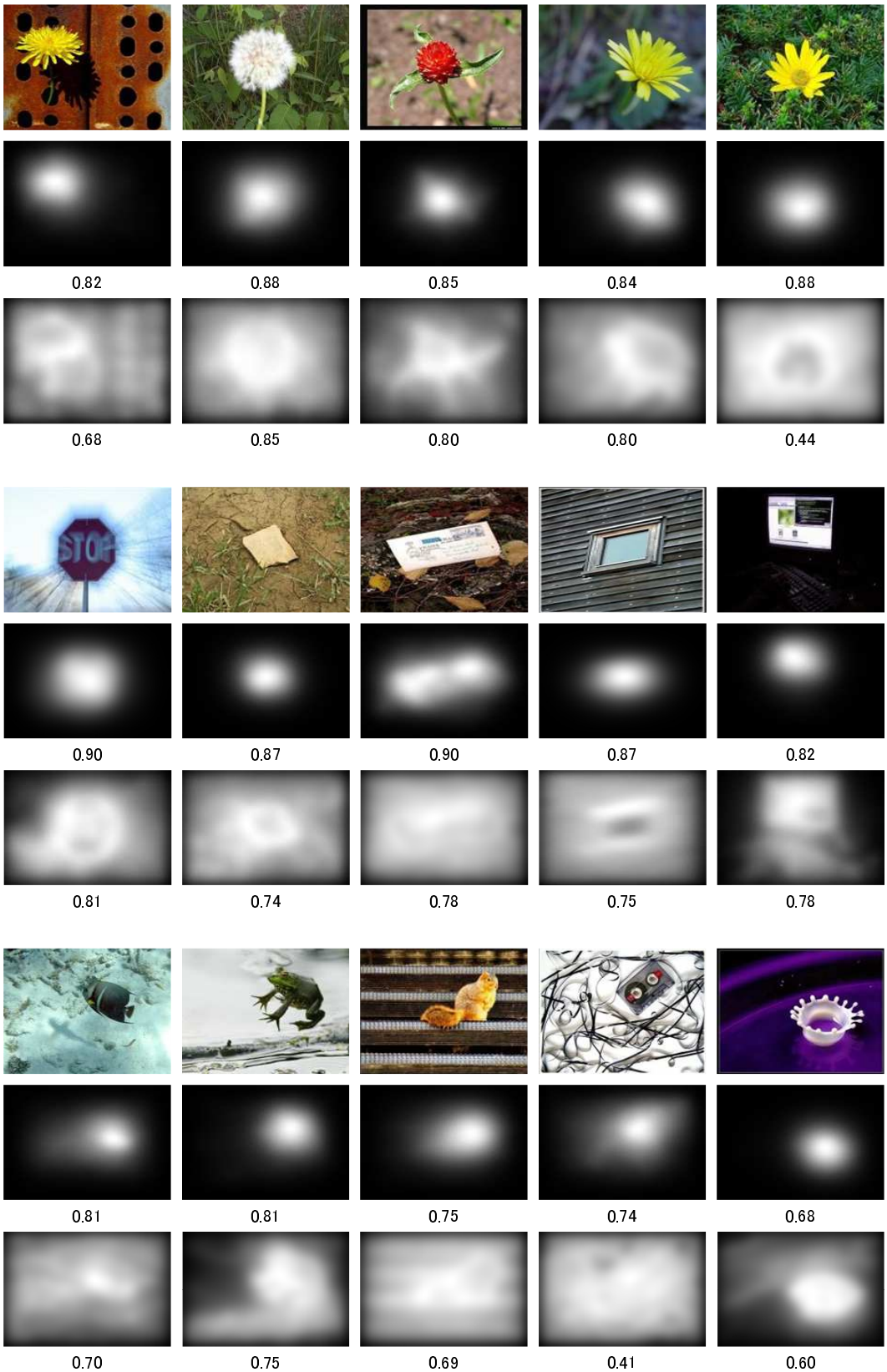


Fig. 6 Experimental results on the MSRA dataset. Top rows: Original images. Middle rows: Saliency maps obtained by the proposed method. Bottom rows: Saliency maps obtained by the K&B method. The number below each image is the maximum F-score.

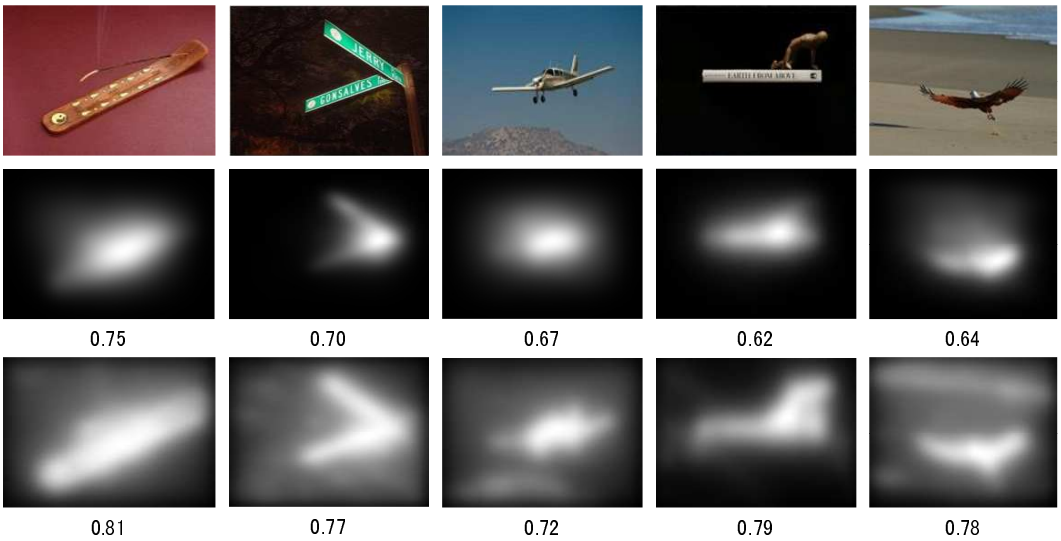


Fig. 7 Examples where the proposed methods do not perform well. The proposed method is not suitable for finding elongated salient regions.

saliency maps. In the proposed RuLSIF-based method, we fix the parameters at $N_c = 50$, $N_s = 50$, $H = 8$, $\beta = 0.1$, and the sizes of center and surround regions are 0.2 and 0.3 of the entire image.

The quality of an obtained saliency map is evaluated according to Achanta et al. [1]: A binary map is constructed from an obtained saliency map by varying a threshold on the intensity values in $[0, 255]$. Then each of these 256 maps is compared with the ground-truth binary map and the precision and recall scores are computed.

We compare precision-recall curves for each low-dimensional feature (lightness, edge, and color) and for the combined feature in Fig. 5. The graphs show that our method tends to outperform the K&B method for each low-dimensional feature channel and it more clearly outperforms the K&B method for the combined feature.

Fig. 6 shows examples of saliency maps obtained by the proposed approach and the K&B method; below each image, the maximum F-score (i.e., the maximum of the harmonic mean of precision and recall) is described. Overall, the proposed method gives much better results both in visual quality and the F-score.

A potential weakness of the proposed method is that if a salient object is highly elongated, its shape cannot be extracted sharply (see Fig. 7). This weakness is caused by our search strategy that density ratios between *spherical* regions are estimated. Indeed, this weakness can be overcome by considering elongated regions in the saliency search. However, this in turn increases the computational cost significantly.

5. Conclusion

We presented a new approach to computing visual saliency based on direct density-ratio estimation. Direct density-ratio estimation is an emerging machine learning technique that allows us to systematically avoid density estimation, which is known to be a hard task. Based on an estimated density ratio, we determined the contrast of the center and the surround feature distributions for lightness, edge, and color channels. Through experiments, we demonstrated that our proposed approach outperforms the K&B method which is based on probability density estimation.

We experimentally found that the proposed method cannot sharply identify a salient object if its shape is elongated, which is due to our search strategy that density ratios between *spherical* regions are estimated. If elongated regions are used for saliency search, this problem can be overcome in principle. Thus, this weakness of the proposed method is not an essential limitation. However, naively employing various elongated regions in saliency search increases the computation cost significantly. Thus, we will develop a computationally efficient way to handle this problem in our future work.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[2] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, 2010.

[3] S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*, volume 3899 of *Lecture Notes in Artificial Intelligence*. Springer, 2006.

[4] S. Frintrop, E. Rome, and H. I. Christensen. Computational visual attention systems and their cognitive foundation: A survey. *ACM Transactions on Applied Perception*, 7(1):6:1–6:39, 2010.

[5] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *Advances in Neural Information Processing Systems 21*, pages 681–688, 2009.

[6] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[7] T. Kadir, A. Zisserman, and M. Brady. *An Affine Invariant Salient Region Detector*, volume 3021 of *Lecture Notes in Computer Science*. Springer, 2004.

[8] T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

[9] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Y. Shum. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):353–367, 2011.

[10] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2232–2239, 2009.

[11] B. Schauerte and G. A. Fink. Focusing computational visual attention

in multi-modal human-robot interaction. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 6:1–6:8, 2010.

[12] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge, UK, 2012.

[13] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[14] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

[15] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[16] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501, 2004.

[17] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems 24*, pages 594–602, 2011.