Food Region Extraction Based on Saliency Detection Model*

Ayako Kitada†, Takuya Futagami† and Noboru Hayasaka†

In this paper, we propose a method that can automatically extract food regions from food images by using the saliency detection model based on a deep neural network (DNN) and the saliency thresholding method based on the average saliency value. Our experiment, using 125 food images from a food recording tool on smartphone applications, demonstrates that the proposed method significantly increased average F-measure by 4.22% or more compared with both the conventional method using local extrema and food extraction using DNN trained with 1017 food images. Our proposed method also increased average precision and recall by 0.13% or more and 11.38% or more, respectively. We also discussed the effectiveness and the future development of food extraction using the saliency detection model and saliency thresholding method on the basis of experimental results.

1. Introduction

Recently, the rate of individuals having lifestyle diseases has increased globally. To prevent lifestyle diseases, recording food intake is important[1]. Conventional food recording tools require text-based input; recording food intake using images is expected to be comparatively more efficient. Thus, smartphone applications such as FoodLog[2], DeepFoodCam[3], Im2Calories[4] have been developed that can automatically record food intake using food images.

Food recording tools recognize food using deep neural networks (DNNs)[5] or support vector machines (SVMs)[6]. The recognition requires preprocessing that extracts the image regions containing the food[7–9]. In addition, pixel-wise food extraction is required to measure the total calories of the food intake[10,11]. Therefore, this study is focused on pixel-wise food extraction from food images.

Refs. [12–15] proposed food extraction using DNN, requiring pixel-wise annotated food images as a training dataset. The food regions include various shapes, colors, and textures[12,16]. In addition, the pixel-wise annotated food images dataset is not sufficiently available on the Internet[10]. Thus, the training cost to construct a highly accurate DNN is high as mentioned by Refs. [10,12,13]. Therefore, it would be computationally economical if the food extraction did not require annotated food images.

Food extraction using saliency values[17], which tends to be higher in the regions of the image that attract human attention than in the image background, can decrease the training cost as it requires no pixel-wise annotated food images as the training dataset. Refs. [16,18] extracts the food regions under the assumption that the food, which is the main subject of the image, tends to be salient compared with the background. Ref. [16] measures the saliency by tracking the eyes of the users using wearable computer glasses. Ref. [18] measures the saliency on the basis of the local extrema of pixel values. As real-world food recording tools in smartphones[2–4] are not assumed to use measurement data from wearable computers, this study regards the method used in Ref. [18] as the conventional method.

Saliency detection performance has increased in recent years as DNN is being frequently used in saliency detection models[19]. The training cost of the saliency detection model using DNN can decrease because large annotated datasets for saliency detection (such as SALICON[20]) are available on the Internet unlike that for food extraction. However, the performance of food extraction employing saliency detection models developed in recent years has not been sufficiently discussed in the literature. Thus, we herein discuss food extraction based on a suitable saliency detection model and thoroughly evaluate its effectiveness.

To achieve the aforementioned objective, our experiment determines the saliency detection model suitable for food extraction. In addition, the effectiveness of our proposed method is evaluated applying
125 food images that are commonly used by a food recording tool.

The image-based food calorie measurement requires food recognition at pixel level to increase its accuracy as discussed in Ref. [4]. Although the food extraction using DNN [12–15] can recognize food type at pixel level, it requires the pixel-wise annotated training dataset, which is extremely laborious and costly to be prepared [21].

However, the combination of the proposed method, which can determine only whether food or background at pixel level, and bounding box food recognition such as Ref. [22] can also recognize food type at pixel level. The combination can achieve pixel level food recognition by providing the food type recognized by the bounding box food recognition to the only pixels which are identified as the food by the proposed method. Because the bounding box food recognition requires the bounding box annotation for training and the proposed method requires no annotation for food images, the combination can avoid the extremely laborious pixel-wise annotation for food images. Thus, the proposed method is expected to contribute to decrease the annotation cost to achieve the image-based food calorie measurement compared with the food extraction using DNN.

Furthermore, the proposed method can be applied to aid constructing the pixel-wise annotated food images dataset. Thus, the proposed method is expected to decrease the annotation cost to construct a highly accurate food extraction using DNN.

Ref. [23], which corresponds to our previously reported paper, did not compare the proposed method with food extraction using DNN in terms of food extraction accuracy. Although the food extraction accuracy of the DNN can be improved by increasing the training cost, we consider that it is relatively more important to clarify the effectiveness of the proposed method compared with effectiveness of the low-training-cost DNN trained by the annotated food images dataset available on the Internet. Thus, this study compares the food extraction accuracy between the proposed method and food extraction using low-training-cost DNN.

We note food images are assumed to be prepared for recording food intake. Thus, the food images used in this paper are regarded to include only food, tableware and table.

This paper is structured as follows. In 2. and 3., overviews of the conventional and proposed methods, respectively, are provided. In 4., we discuss the experiments conducted to clarify the effectiveness of our proposed method. In 5., we provide concluding statements.

2. Conventional Method

This section provides an overview of the conventional method [18] that extracts food on the basis of the salient region. Fig. 1 shows the flowchart of the conventional method. The details of each step of the flowchart are described in the following subsections.

2.1 Convex Hull Based on Local Extrema

To obtain the salient regions, the local extrema of the pixel value are extracted by applying the star detector [24] derived from the center surround extremas (CenSurE) [25]. The star detector is a rotation and scale-invariant extrema detector. The red points in Fig. 2 (b) depict the local extrema extracted from the input image (Fig. 2 (a)) by applying the star detector. All food images provided herein are simulated ones captured by us.

The convex hull that includes all local extrema is generated using the algorithm proposed by Ref. [26]. The green line in Fig. 2 (b) depicts the convex hull that contains the local extrema corresponding to the red points.

2.2 GrabCut

GrabCut [27], which is based on graph theory, repeatedly applies the maximum flow algorithm [28] to a graph based on both the Gaussian mixture model of the pixel values and similarity of the neighborhood pixels. Because GrabCut requires the initial regions
where the food and background are roughly divided, we use the conventional method to generate the initial regions by classifying the inner and outer areas of the convex hull as the food and the background, respectively. Fig. 2 (c) shows the extracted food by applying GrabCut to the graph initialized by the convex hull, as shown in Fig. 2 (b).

3. Proposed Method

This section provides an overview of our proposed method that extracts food by applying a suitable saliency detection model. Fig. 3 shows the flowchart of the proposed method. The details of each step of the flowchart are described in the following subsections.

3.1 Saliency Detection Model

The saliency map \( S(x,y) \), which has a saliency value for each pixel position \((x,y)\), is generated by applying the saliency detection model to the input image. As the performance of saliency detection models have been improved in recent years, food extraction accuracy is expected to increase using these models.

However, the saliency detection model most suitable for food extraction has not been discussed in the literature. Thus, this study experimentally determines the suitable saliency detection model to improve food extraction accuracy.

Generally, saliency regions are detected on the basis of handcrafted features, designed by researchers, and DNN, which obtains the efficient features via training[17]. Herein, we chose a notable saliency detection model using handcrafted features and a model using DNN on the basis of Ref.[23].

As for the saliency detection models using handcrafted feature, GMR[29], which generates a graph by using the SLIC algorithm[30] and uses superpixels located at the four sides of an image as background queries, is chosen.

As for the saliency detection models using DNN, MSI[31], which adapts a trained VGG-16 encoder[32] used as popular feature extractor in the saliency detection, is chosen. MSI uses multiple convolutional layers to obtain multiple spatial-scale features.

Fig. 4 (a) depicts the saliency map generated by applying MSI to the input image (Fig. 1 (a)).

\[ T_{\text{ave}} = \frac{\alpha}{W \cdot H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} S(x,y), \]

where \( W \) and \( H \) represent the width and height of the image. In addition, \( \alpha \) depicts a scale parameter. Ref.[35] also uses saliency threshold \( T_{\text{dis}} \) based on discriminant analysis[36].

However, the saliency thresholding method suitable for food extraction has not been discussed in the literature. Thus, this study experimentally employs the saliency thresholding method suitable for food extraction from the aforementioned three methods.

Fig. 4 (b) depicts the extracted food regions obtained by applying the saliency thresholding method based on the discriminant analysis of the saliency map (Fig. 4 (a)).

4. Experiment

This section confirms the effectiveness of our proposed method by comparing its food extraction accuracy with other methods. The saliency detection model and saliency thresholding method are also experimentally determined to construct our proposed method.

4.1 Experimental Conditions

The effectiveness of our proposed method is evaluated using 125 food images actually used in a food recording tool (hereafter called the evaluation dataset). A ground truth indicating food regions in the images was manually prepared by following the same policy.
as in Ref. [18].

First, the preliminary experiment, which clarifies the suitable saliency detection model for food extraction, is conducted to construct our proposed method. The performance of the two saliency detection models listed in 3.1 are evaluated by comparing the generated saliency map and the ground truth. The MSI was trained by the 10,000 training images provided by SALICON[20], available on the Internet. The area under a receiver operating characteristic (ROC) curve (hereinafter called AUC)[37], which is frequently used in evaluating classifiers, was used for an evaluation metric. The ROC curves plotted the FP (which shows the rate of incorrectly determined pixels in the background) on the X-axis and the TP (which shows the rate of correctly determined pixels in the food) on the Y-axis by changing the saliency threshold. The saliency detection model having the highest average AUC is employed for the proposed method.

Second, we conduct the main experiment evaluating the food extraction accuracy of the proposed method using saliency detection model suitable for food extraction. Our evaluation uses precision, recall, and F-measure as the metrics of the food extraction accuracy as shown as follows.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{2}
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \tag{3}
\]
\[
F - \text{measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}
\]

where TP shows the rate of correctly determined pixels in food. In addition, FP and FN show the rate of incorrectly determined pixels actually in the food and the background, respectively. Generally, as there is a tradeoff between the precision and recall, the F-measure, which shows the harmonic mean of the precision and recall, is used as the comprehensive evaluation metric.

The food extraction accuracy of our proposed method is compared with that of the conventional method[18] and the food extraction using DNN (hereafter called the comparison methods). Food extraction using DNN increases the training cost as it requires annotated food images as the training dataset. However, the training cost can decrease using the dataset available on the Internet. Therefore, this experiment adds food extraction using the low-training-cost DNN to the comparison method. We employed SegNet[38], which was applied to the food extraction in Refs. [14,15], as the food extraction using low-training-cost DNN. SegNet is trained by UNIMIB2016[39], which includes 1017 annotated food images available on the Internet.

In addition, 1017 annotated food images included in UNIMIB2016 were used to adjust the parameters of the proposed and conventional methods by measuring average F-measure. The saliency thresholding method employed for the proposed method was also selected by using these food images.

### 4.2 Experimental Results

#### 4.2.1 Preliminary Experiment

This section provides the results of the preliminary experiment that aims to clarify the suitable saliency detection model for food extraction. The average AUC of the GMR and MSI, listed in 3.1, are 83.71% and 90.00%, respectively. These results show that the MSI increased the average AUC by 6.29% compared with the GMR. Thus, our proposed method employs MSI to generate the saliency map.

Shapiro-Wilk test demonstrated nonnormality of the AUC provided by both GMR and MSI at a 5% significance level (the observed p-value was less than 0.01%). Thus, the Wilcoxon signed-rank test, which is categorized as the non-parametric comparison test, was applied to detect the significant difference at the same significance level, thereby detecting significant differences in AUC between the MSI and the GMR (the observed p-value was less than 0.01%).

Fig. 5 provides examples of the saliency map generated from each saliency detection model. It is evident that the MSI could increase the AUC compared with the GMR at each example because the MSI tended to provide higher and lower saliency values to the food and the background, respectively.

#### 4.2.2 Main Experiment

Table 1 shows the food extraction accuracy obtained from each method. Our proposed method experimentally employed the saliency threshold \( T_{\text{ave}} \) based on the average saliency value. Our proposed method increased the average F-measure, which shows the comprehensive metric, by 4.22% or more compared with the comparison methods (the conventional method and SegNet). The Wilcoxon signed-rank test, which was selected as the comparison test on the basis of the result of the Shapiro-Wilk test (the observed p-value was less than 0.01%), demonstrated that the F-measure of the proposed method was significantly larger than that of the comparison methods at the 5% significant level (observed p-values were less than 0.34%).

Furthermore, our proposed method increased both precision and recall by 0.13% or more and 11.38% or more, respectively, compared with the comparison methods. Although the recall of our proposed method was not significantly different from that of the comparison methods (observed p-values were more than 22.59%), the precision of our proposed method was significantly different (observed p-values were less than 0.01%).

The experiment confirmed the effectiveness of our proposed method as all metrics of the food extraction accuracy increased. The effectiveness of the proposed method was also confirmed as it significantly increased the recall and F-measure.

Fig. 6 depicts the examples of the extracted food regions by each method. Although our proposed
method tends to determine the dishes as the food regions, all food is observed to be extracted. Compared with SegNet, the proposed method, which tended to erroneously misdetermine the dishes as the food regions, correctly determined the rice and fish as the food regions. Since this tendency increased the recall, thereby increasing the F-measure of the proposed method compared with the SegNet as shown in Table 1. Compared with the conventional method, our proposed method was observed to determine the background as the background correctly.

4.3 Discussion
This section discusses the effectiveness and future development of the proposed method on the basis of the experimental results.

4.3.1 Saliency Detection Model
There are two types of food images: those that include single or multiple food items in the image (hereafter called single or multiple food images). Our analysis showed that the evaluation dataset contains 51 single food images and 74 multiple food images.

The single food image, corresponding to the image of the first row in Fig. 6, tends to contain food regions at the center of the image compared with the multiple food image, which corresponds to the image of second row in Fig. 6. This tendency can also be confirmed from the images in Fig. 7, which compares the probability of finding food in each pixel between single and multiple food images. These figures are obtained from the ground truth that indicate food regions and were resized to $256 \times 256$ pixels.

Table 2 shows the average AUC for single and multiple food images by each saliency detection model. The effectiveness of the MSI can also be confirmed from this result because MSI can increase the average AUC compared with the GMR at both single and multiple food images.

Although the MSI does not have significant differences in AUC between single and multiple food images (observed p-value was more than 10.96%), the GMR significantly decreased the AUC of the multiple food images (observed p-value was less than 0.01%). The reason is that the GMR makes the saliency value around the center of the image high (hereafter called center prior) unlike the MSI, as the subject of the image tends to appear at the center. The center prior was unsuitable for multiple food images that do not tend to contain the food regions at the center of the image compared with single food images. We also confirmed that other models in Ref. [23] with the center prior demonstrated the same tendency. For future development, we should note that the saliency detection model that has the center prior was not effective in multiple food images.

4.3.2 Saliency Thresholding Method
Table 3 compares the average F-measure provided by combinations of the saliency detection model and

---

Table 1 Average extraction accuracy of each method (CM: conventional method, PM: proposed method)

| Metric     | Method | CM | SegNet | PM |
|------------|--------|----|--------|----|
| Precision  | %      | 61.33 | 59.59  | 61.46 |
| Recall     | %      | 77.00 | 68.95  | 88.38 |
| F-measure  | %      | 68.28 | 63.93  | 72.50 |

---

Fig. 5 Saliency map provided by saliency detection models [29,31]
the saliency thresholding method. In the table, our proposed method corresponds to the combination of MSI and the saliency thresholding method based on the average $T_{\text{ave}}$. The column of $T_i$ shows the average F-measure obtained from the ideal saliency thresholding method, adjusting the threshold to each saliency map in terms of F-measure. Although our proposed method increased the average F-measure by 4.22% or more compared with the comparison methods as shown in 4.2.2, the proposed method is observed to have room for increasing the average F-measure by 3.04% by improving the saliency thresholding method.

The table also suggests that the saliency thresholding method suitable for the food extraction was not always the method based on the average $T_{\text{ave}}$. Thus, the suitable thresholding method should be determined depending on the employed saliency detection model.

Generally, the saliency thresholding method based on the discriminant analysis is known to be effective as it can set the threshold without any parameter unlike other methods. However, the saliency thresholding method based on the discriminant analysis tended to decrease the average F-measure in Table 3. This is because the saliency value histogram does not always have a bimodal distribution, which the discriminant analysis assumes, as shown in Fig. 8; Fig. 8 shows the histogram that decreases monotonously based on the saliency map by MSI of the image at the second row in Fig. 5.

### 5. Conclusion

This study proposes automatic food region extraction from food images based on a combination of saliency detection and saliency thresholding. Our

| Food type | SDM | STM |
|-----------|-----|-----|
|           | GMR[29] | MSI[31] | $T_{\text{fix}}$ | $T_{\text{ave}}$ | $T_{\text{dis}}$ | $T_i$ |
| Single [%] | 89.18 | 90.14 | 65.29 | 64.70 | 65.05 | 72.12 |
| Multiple [%] | 79.94 | 89.90 | 72.11 | 72.50 | 63.08 | 75.54 |

Fig. 6 Results of food extraction with each method

Fig. 7 Probability of finding food in each pixel (whiter pixels show higher probability)

Table 2 Average AUC of each saliency detection model (SDM) for single and multiple food images

Table 3 Average F-measure of food extraction by combinations of the saliency detection model (SDM) and saliency thresholding method (STM)

Fig. 8 Saliency value histogram of the second row in Fig. 5 provided by MSI[31]
proposed method experimentally employs the MSI, which detects saliency on the basis of multiple convolutional layers, and the saliency thresholding method based on average saliency values.

The effectiveness of our proposed method was evaluated by using 125 food images utilized in a food recording tool on smartphones. Our proposed method significantly increased the average F-measure, which shows the comprehensive evaluation metric, by 4.22% or more compared with the comparison methods. Our proposed method was also effective as it increased all metrics of the food extraction accuracy.

Furthermore, we discussed our proposed method in terms of effectiveness and future development. First, the saliency detection models were discussed. The effectiveness of the MSI was confirmed as it increased the average AUC for both the single and multiple food images. In addition, the MSI did not have significant difference in AUC between single and multiple food images unlike the other saliency detection models. This is because MSI does not make the saliency value around the center of the image high unlike the other models.

Second, saliency thresholding methods were discussed. Our discussion suggests that the suitable saliency thresholding method depends on the employed saliency detection model. In addition, the proposed method was observed to have room for increasing the average F-measure by 3.04% by improving the saliency thresholding methods. Although saliency thresholding methods based on discriminant analysis are effective as they can set the threshold without any parameters, they tend to decrease the average F-measure. This is because the histogram of the saliency values does not have a bimodal distribution assumed by the discriminant analysis.

**References**

[1] M. Fujishiro: Dietary management of obesity; *Journal of Nihon University Medical Association*, Vol. 78, No. 4, pp. 223–229 (2019)

[2] A. Kiyoharu: Image recognition-based tool for food recording and analysis; *FoodLog: Connected Health in Smart Cities*, Springer, Cham, pp. 1–9 (2020)

[3] R. Tanno, K. Okamoto and K. Yanai: DeepFood-Cam: A CNN-based real-time mobile food recognition system; *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, p. 89 (2016)

[4] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang and K. P. Murphy: Im2Calories: Towards an automated mobile vision food diary; *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pp. 1233–1241 (2015)

[5] P. Dhruv and S. Naskar: Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review; *Proceedings of the International Conference on Machine Learning and Information Processing*, pp. 367–381 (2020)

[6] U. Maulik and D. Chakraborty: Remote sensing image classification: A survey of support-vector-machine-based advanced techniques; *IEEE Geoscience and Remote Sensing Magazine*, Vol. 5, No. 1, pp. 33–52 (2017)

[7] V. H. Reddy, S. Kumari, V. Muralidharan, K. Gigoo and B. S. Thakare: Literature survey food recognition and calorie measurement using image processing and machine learning techniques; *Proceedings of the 2nd International Conference on Communications and Cyber Physical Engineering*, pp. 23–37 (2020)

[8] L. Jiang, B. Qiu, X. Liu, C. Huang and K. Lin: DeepFood: Food image analysis and dietary assessment via deep model; *IEEE Access*, Vol. 8, pp. 47477–47489 (2020)

[9] Y. Kawano and K. Yanai: Real-time mobile food recognition system; *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1–7 (2013)

[10] T. Ege, W. Shimoda and K. Yanai: A new large-scale food image segmentation dataset and its application to food calorie estimation based on greyscale of rice; *Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management*, pp. 82–89 (2019)

[11] L. Zhou, C. Zhang, F. Liu, Z. Qiu and Y. He: Application of deep learning in food: A review; *Comprehensive Reviews in Food Science and Food Safety*, Vol. 18, No. 6, pp. 1793–1811 (2019)

[12] Y. Wang, F. Zhu, C. J. Boushey and E. J. Delp: Weakly supervised food image segmentation using class activation maps; *Proceedings of the 2017 IEEE International Conference on Image Processing*, pp. 1277–1281 (2017)

[13] G. Ciocca, D. Mazzini and R. Schettini: Evaluating CNN-based semantic food segmentation across illuminants; *Proceedings of the 7th International Workshop on Computational Color Imaging*, pp. 247–259 (2019)

[14] J. O. Pinzón-Arenas, R. Jiménez-Moreno and C. G. Pachón-Suescún: ResSeg: Residual encoder-decoder convolutional neural network for food segmentation; *International Journal of Electrical and Computer Engineering*, Vol. 10, No. 2, pp. 1017–1026 (2020)

[15] N. Jamil, N. A. N. Redzuan, M. F. Ismail and W. A. W. Ramli: Evaluation of VGG networks for semantic image segmentation of Malaysian meals; *Proceedings of the 1st International Conference on Informatics, Engineering, Science and Technology* (2019)

[16] H. C. Chen, W. Jia, X. Sun, Z. Li, Y. Li, J. D. Fernstrom and M. Sun: Saliency-aware food image segmentation for personal dietary assessment using a wearable computer; * Measurement Science and Technology*, Vol. 26, No. 2, p. 025702 (2015)

[17] A. Borji, M. M. Cheng, Q. Hou, H. Jiang and J. Li: Salient object detection: A survey; *Computational Visual Media*, Vol. 5, No. 2, pp. 117–150 (2019)

[18] H. Sugiyama, C. Morikawa and K. Aizawa: Segmentation of food images by local extrema and GrabCut; *The Journal of the Institute of Image Information and Television Engineers*, Vol. 66, No. 5, pp. J179–J181 (2012) (in Japanese)

[19] N. Kumar: Thresholding in salient object detection:
a survey; *Multimedia Tools and Applications*, Vol. 77, No. 15, pp. 19139–19170 (2018)

[20] M. Jiang, S. Huang, J. Duan and Q. Zhao: Salicon: Saliency in context; *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1072–1080 (2015)

[21] S. Belharbi, I. B. Ayed, L. McCaffrey and E. Granger: Deep active learning for joint classification and segmentation with weak annotator; *arXiv preprint*, arXiv:2010.04899 (2020)

[22] H. Ye and Q. Zou: Food recognition and dietary assessment for healthcare system at mobile device end using mask R-CNN; *Proceedings of the International Conference on Testbeds and Research Infrastructures*, pp. 18–35 (2019)

[23] T. Futagami, A. Kitada and N. Hayasaka: Food region extraction by applying saliency detection model; *Proceedings of the 64th Annual Conference of the Institute of Systems, Control and Information Engineers (ISCIE)*, pp. 57–62 (2020) (in Japanese)

[24] M. Weinmann: Visual features—From early concepts to modern computer vision; *Advanced Topics in Computer Vision*, Springer, pp. 1–34 (2013)

[25] M. Agrawal, K. Konolige and M. R. Blas: Censure: Center surround extremas for realtime feature detection and matching; *Proceedings of the 10th European Conference on Computer Vision*, pp. 102–115 (2008)

[26] J. Sklansky: Measuring concavity on a rectangular mosaic; *IEEE Transactions on Computers*, Vol. 21, No. 12, pp. 1355–1364 (1972)

[27] C. Rother, V. Kolmogorov and A. Blake: GrabCut: interactive foreground extraction using iterated graph cuts; *ACM Transactions on Graphics*, Vol. 23, No. 3, pp. 309–314 (2004)

[28] Y. Boykov and V. Kolmogorov: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision; *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 9, pp. 1124–1137 (2004)

[29] C. Yang, L. Zhang, H. Lu, X. Ruan and M. H. Yang: Saliency detection via graph-based manifold ranking; *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3166–3173 (2013)

[30] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. S{"u}sstrunk: SLIC superpixels compared to state-of-the-art superpixel methods; *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 11, pp. 2274–2282 (2012)

[31] A. Kroener, M. Senden, K. Driessen and R. Goebel: Contextual encoder-decoder network for visual saliency prediction; *Neural Networks*, Vol. 129, pp. 261–270 (2020)

[32] K. Simonyan and A. Zisserman: Very deep convolutional networks for large-scale image recognition; *arXiv preprint*, arXiv:1409.1556 (2014)

[33] A. Singh, C. H. H. Chu and M. A. Pratt: Multiresolution superpixels for visual saliency detection; *Proceedings of the IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vision Pro-
cessing*, pp. 1–8 (2014)

[34] X. Xu, N. Mu, H. Zhang and X. Fu: Salient object detection from distinctive features in low contrast images; *2015 IEEE International Conference on Image Processing*, pp. 3126–3130 (2015)

[35] P. Khuwuthayakorn, A. R. Kelly and J. Zhou: Object of interest detection by saliency learning; *Proceedings of the 11th European Conference on Computer Vision*, pp. 636–649 (2010)

[36] N. Otsu: A threshold selection method from gray-level histograms; *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, No. 1, pp. 62–66 (1979)

[37] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba and F. Durand: What do different evaluation metrics tell us about saliency models?; *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 3, pp. 740–757 (2018)

[38] V. Badrinarayanan, A. Kendall and R. Cipolla: SegNet: A deep convolutional encoder-decoder architecture for image segmentation; *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 12, pp. 2481–2495 (2017)

[39] G. Ciocca, P. Napoletano and R. Schettini: Food recognition: a new dataset, experiments, and results; *IEEE Journal of Biomedical and Health Informatics*, Vol. 21, No. 3, pp. 588–598 (2016)

Authors

Ayako Kitada

She received her B.S degree from Osaka Electro-Communication University in 2021. Her current research interests include image processing and image recognition.

Takuya Futagami (Member)

He received his B.S., M.S., and Ph.D. degrees from Osaka University in 2013, 2015, and 2021 respectively. He has been a visiting researcher at Osaka Electro-Communication University since 2018. His current research interests include image processing and image recognition. He is a member of IEICE.

Noboru Hayasaka

He received his B.S., M.S., and Ph.D. degrees from Hokkaido University in 2002, 2004, and 2007, respectively. He is currently an associate professor at Osaka Electro-Communication University. His current research interests include speech processing and speech recognition. He is a member of ASJ, IEICE, and IEEE.